Evolution of substrate specificity and protein-protein interactions in three enzyme superfamilies

Dissertation

zur Erlangung des Doktorgrades der
Naturwissenschaften (Dr. rer. nat.) der
Fakultät für Biologie und vorklinische
Medizin der Universität Regensburg

vorgelegt von
Maximilian Plach
aus Bad Kötzting

Januar 2017

Das Promotionsgesuch wurde eingereicht am 11.01.2017.

Die Arbeit wurde angeleitet von Prof. Dr. Reinhard Sterner.

Unterschrift:

Abstract

Superfamilies are a classification system to combine proteins that are related through a common evolutionary origin, share similar sequences, structures, and core reaction mechanisms, but exert different functions. Today, for most superfamilies tens of thousands of sequences and hundreds of structures are known and most of the different functions of their members have been elucidated. Superfamilies thus provide a formal and biologically sensible framework to study evolutionary relationships between proteins. In the present work, the frameworks of three enzyme superfamilies were utilized to get insights into several important aspects of enzyme evolution.

The first part of this work addresses the question how enzymatic mono- and bi-functionality have evolved in the superfamily of ribose-binding $(\beta\alpha)_8$-barrel sugar isomerases. This superfamily contains the homologous enzymes HisA and TrpF, which catalyze similar reactions in histidine and tryptophan biosynthesis, as well as the bi-functional enzyme PriA, which catalyzes both the HisA and TrpF isomerization reactions. HisA and TrpF are ubiquitous in Archaea and Bacteria, whereas PriA is only found in certain Actinobacteria. These species have lost the dedicated TrpF enzyme and PriA is consequently part of both tryptophan and histidine biosynthesis. Much has been speculated on the evolutionary relationship of these enzymes and whether the bi-functionality of PriA is a remnant from ancient evolutionary times or a more recent development in Actinobacteria. Using ancestral sequence reconstruction it was demonstrated in this work that evolutionary ancestors of modern HisA enzymes display bi-functionality, reminiscent of PriA. A detailed enzymatic characterization of three reconstructed HisA ancestors showed that they catalyze not only the HisA but also the TrpF reaction with comparable catalytic efficiencies *in vitro*. Metabolic complementation experiments with *his*A and *trp*F deficient *Escherichia coli* strains furthermore demonstrated that the bi-functional HisA ancestors could support both histidine and tryptophan biosynthesis *in vivo*. By a combination of sequence- and network-based *in silico* methods, several modern HisA enzymes were subsequently identified that possess sequence motifs typical for bi-functional PriA enzymes. The enzymatic characterization of three such modern HisA representatives revealed that they are also bi-functional, albeit to a lesser extent, although the respective organisms possess dedicated TrpF enzymes. Thus, the ancestral bi-functionality has pertained for billions of years in HisA enzymes, without any obvious selective pressure. Consequently, a new model for the evolution of HisA, TrpF, and PriA was proposed: The bi-functionality of ancient HisA variants may have played an important role in maintaining early metabolism by supporting both histidine and tryptophan biosynthesis. After the emergence of dedicated TrpF enzymes the bi-functionality of the ancestors became expendable and diminished to the level observed in modern HisA enzymes. However, the inherent bi-functionality of HisA contributed to the robustness of microbial metabolism and made possible to compensate the loss of a dedicated *trp*F gene in some Actinobacteria. In these organisms, the available bi-functionality of HisA was exploited, selected for, and enhanced, which eventually led to the modern PriA enzymes.

The second part of this work deals with the evolution of substrate specificity and secondary metabolic enzymes in a superfamily of chorismate-utilizing enzymes, named MST-superfamily. Chorismate is a central metabolic node molecule and the starting point for the biosynthesis of various important metabolites, including aromatic amino acids, folate, or iron-chelating siderophores. The MST-enzymes catalyze the committed steps of these biosynthetic pathways and are highly similar in sequence, structure, and reaction mechanism. However, the MST-enzymes that are part of primary metabolic pathways employ exclusively ammonia as a nucleophile to aminate chorismate, whereas those that are part of secondary metabolic pathways exclusively employ water as a nucleophile to hydroxylate chorismate. Based on the notion that secondary metabolic enzymes are descendants of primary metabolic ones, it was investigated in this part of this work by which mechanism the transition from primary metabolic to secondary metabolic MST-enzymes went along with a change in nucleophile-specificity from ammonia to water. Initially, network-based, phylogenetic, and structure-based *in silico* methods were applied to identify two key amino acids in the nucleophile access channel of the active site that distinguish primary-metabolic/ammonia-utilizing and secondary-metabolic/water-utilizing MST-enzymes. The importance of these key positions was subsequently examined by rationally designing sixteen variants of the MST-enzyme anthranilate synthase, which normally employs ammonia as a nucleophile. The enzymatic characterization of these variants by HPLC-MS showed that the right combination of amino acids at the two key positions indeed resulted in a broadening of nucleophile specificity to also include water. These anthranilate synthase variants hydroxylated chorismate and formed isochorismate with efficiencies comparable to native secondary-metabolic/water-utilizing isochorismate synthases. Moreover, these variants were still able to employ ammonia as a nucleophile

and formed their native product anthranilate; hence they were bi-functional. These experiments demonstrated that nucleophile specificity in the MST-superfamily can readily switch from ammonia to water. Moreover, the observed bi-functionality of the anthranilate synthase variants argues that the evolution of secondary metabolic MST-enzymes may have proceeded through bi-functional intermediates. Such metabolic generalists may have allowed for the formation of novel metabolites (isochorismate) while maintaining the formation of important primary metabolic metabolites (anthranilate). This scenario consequently does not *a priori* require gene duplication events and thus precludes negative metabolic effects linked to retaining redundant gene copies.

The third part of this work pursues the question how protein-protein interaction specificity is assured in superfamilies of structurally related protein complexes and how the determinants of interaction specificity have evolved. Specific interactions between proteins are vital for almost all cellular functions. This specificity is usually achieved by shape and electrostatic complementarity of protein interfaces. However, the number of different protein folds and interface geometries found in Nature is limited, due to the constraints imposed by efficiently packing hydrogen-bonded secondary structure elements. It is thus a challenging question how interaction specificity is achieved despite structural limitations and how the formation of non-physiological complexes is avoided when several possible interaction partners with similar interface geometries are available. In order to address this problem, initially a comprehensive computational survey of the interface geometries of over 300 bacterial, heteromeric protein complexes and all their homologs of respective superfamilies was performed. This survey revealed that in about 10% of the superfamilies interface geometries vary significantly between related complexes that share homologous subunits. In these cases interfaces were extended by so-called interface add-ons, which typically comprise 10-20 amino acids, form well-defined secondary structure elements, and significantly contribute to complex stability. These characteristics suggested that interface add-ons differentiate between structurally related protein complexes and contribute to interaction specificity through negative design. In order to back this assumption, the case of the interface add-on found in a superfamily of glutamine amidotransferase complexes involved in tryptophan and folate biosynthesis was subsequently analyzed in detail. These complexes comprise synthase and glutaminase subunits that interact to transfer ammonia from glutamine to an acceptor substrate. A subset of synthase subunits exclusively involved in tryptophan biosynthesis contains the interface add-on, whereas it is absent in all other homologous synthase subunits, including those exclusively involved in folate biosynthesis. The comprehensive experimental characterization of 54 combinations of different synthase and glutaminase subunits by

chromatographic methods, light scattering, mass spectrometry, and enzyme kinetics demonstrated that the presence or absence of the interface add-on determines interaction specificity. An *in silico* genetic profiling of over 15 000 archaeal and bacterial genomes together with *in vivo* growth assays showed that the interface add-on found in complexes of tryptophan biosynthesis is biologically relevant for preventing cross-interactions with the homologous complexes of folate biosynthesis, which would lead to harmful metabolic cross-talk that negatively affects cellular fitness. It was finally shown by protein design that the evolution of the interface add-on in these complexes most likely proceeded via intermediary complexes with relaxed interaction specificity. In conclusion, this part of this work demonstrates that interface add-ons are evolutionary tools to facilitate interaction specificity in superfamilies of homologous proteins or in cases where a protein has to discriminate between several potential interaction partners that share similar interface geometries.

In summary, the presented work leads to an improved understanding of the mechanisms behind the evolution of enzymatic mono- and bi-functionality, emphasizes the importance of generalist, bi- or multi-functional enzymes for the evolution of secondary metabolic pathways, and finally describes a so far overlooked structural tool for the evolutionary specification of protein-protein interactions.

Kurzfassung der Arbeit

Superfamilien stellen ein Klassifikationsschema zur Gruppierung von Proteinen dar, die auf einen gemeinsamen evolutionären Ursprung zurück gehen und Ähnlichkeiten in Sequenz, Struktur und Reaktionsmechanismus besitzen, jedoch unterschiedliche Funktionen ausüben. Die meisten Superfamilien sind heute umfassend charakterisiert und häufig sind für eine Superfamilie mehr als zehntausend Sequenzen, mehrere hundert Strukturen sowie die zentralen Funktionen bekannt. Superfamilien erlauben daher als formales und gleichzeitig biologisch sinnvolles Rahmenwerk die Untersuchung der evolutionären Zusammenhänge zwischen Proteinen. Für die vorliegende Arbeit wurden drei Enzym-Superfamilien herangezogen, um Einblicke in mehrere wichtige Fragestellungen der Evolution von Enzymen zu gewinnen.

Der erste Teil dieser Arbeit befasst sich mit der Evolution enzymatischer Mono- und Bifunktionalität in der Superfamilie der Ribose-bindenden Zuckerisomerasen mit $(\beta\alpha_8)$-Fass Struktur. Diese Superfamilie enthält unter anderem die verwandten Enzyme HisA und TrpF, die ähnliche Reaktionen in der Biosynthese der Aminosäuren Histidin und Tryptophan katalysieren. Zusätzlich enthält die Familie auch das bifunktionale Enzym PriA, welches sowohl die HisA, als auch die TrpF Isomerisierungsreaktion katalysiert. Während HisA und TrpF in den meisten Archaeen und Bakterien vorkommen, ist PriA ausschließlich in bestimmten Actinobakterien zu finden. Diesen Spezies fehlt ein dediziertes TrpF Enzym und PriA ist daher essentiell für die Biosynthese sowohl von Histidin als auch von Tryptophan. Es wurden bisher mehrere Hypothesen über die genauen evolutionären Zusammenhänge dieser drei Enzyme aufgestellt; so auch beispielsweise ob die in PriA beobachtete Bifunktionalität ein Überrest von evolutionsgeschichtlich alten Enzymen ist oder ob sie sich erst in neuerer Zeit speziell in Actinobakterien entwickelt hat. Anhand der Rekonstruktion anzestraler Proteinsequenzen wurde in dieser Arbeit gezeigt, dass die evolutionären Vorfahren moderner HisA Enzyme eine zu PriA ähnliche Bifunktionalität aufweisen. Die genaue enzymatische Charakterisierung dreier HisA-Vorläufer ergab, dass diese nicht nur die HisA-Reaktion, sondern auch die TrpF-Reaktion *in vitro* mit vergleichbarer Effizienz katalyiseren. Mittels metabolischer Komplementationsexperimente wurde weiterhin gezeigt, dass die bifunktionalen HisA-Vorläufer *in vivo* die Aufrechterhaltung von Histidin- und Tryptophanbiosynthese ermöglichen. Darauf aufbauend wurden mittels einer Kombination aus sequenz- und netzwerkbasierten *in silico* Methoden mehrere moderne HisA Enzyme identifiziert, die Sequenzmotive enthalten, die für bifunktionale PriA Enzyme typisch sind. Die enzymatische Charakterisierung dreier solcher HisA Vertreter zeigte, dass diese ebenfalls einen gewissen Grad an Bifunktionalität aufweisen. Interessanterweise besitzen die zugehörigen bakteriellen Spezies jedoch eigenständige TrpF Enzyme. Die ursprüngliche Bifunktionalität von HisA hat sich folglich über mehrere Milliarden Jahre erhalten; offenbar ohne direkten Selektionsdruck oder evolutionären Vorteil. Aufbauend auf diesen Ergebnissen wurde abschließend ein neues Modell zur Erklärung der Evolution von HisA, TrpF und PriA vorgeschlagen: In der evolutionären Frühzeit waren bifunktionale HisA Varianten vermutlich wichtig für die Aufrechterhaltung einer primitiven Art von Metabolismus, da sie sowohl die entscheidende Isomerisierungsreaktion in der Histidinbiosynthese, als auch die analoge Reaktion in der Tryptophanbiosynthese katalysieren konnten. Nachdem im Laufe der Evolution allerdings für letztere Reaktion dedizierte TrpF Enzyme entstanden waren, führte ein geringerer Selektionsdruck auf HisA zu einer Abnahme der Bifunktionalität auf das Niveau moderner HisA Vertreter. Dennoch spielte die den HisA Enzymen eigene Bifunktionalität eine wichtige evolutionäre Rolle, da sie die Anpassungsfähigkeit des bakteriellen Metabolismus vergrößerte. So ermöglichten bifunktionale HisA Varianten beispielsweise den Verlust eines dedizierten TrpF Enzyms in Actinobakterien auszugleichen. In diesen Organismen wurde die HisA Bifunktionalität in Folge durch Selektion verstärkt, woraus letztendlich die modernen, bifunktionalen PriA Enzyme hervorgingen.

Im zweiten Teil dieser Arbeit wurde die Evolution von Substratspezifität und sekundärmetabolischen Enzymen am Beispiel einer Superfamilie von Chorismat-bindenden Enzymen, der MST-Superfamilie, untersucht. Chorismat ist ein zentraler metabolischer Knotenpunkt und der Ausgangspunkt für die Biosynthese zahlreicher wichtiger Metabolite wie der aromatischen Aminosäuren, Folsäure, oder eisenchelatierender Siderophore. Die initialen Reaktionen dieser Biosynthesewege werden ausgehend von Chorismat von Enzymen der MST-Superfamilie katalysiert, die hohe Ähnlichkeiten in Sequenz, Struktur und Reaktionsmechanismus aufweisen. Dennoch nutzen diejenigen MST-Enzyme, die Teil von primärmetabolischen Stoffwechselwegen sind, ausschließlich Ammoniak als Nukleophil für die Bildung von aminosubstituierten Chorismatderivativen, wogegen die MST-Enzyme, die Teil von sekundärmetabolischen Wegen sind, ausschließlich Wasser als Nukleophil für die Bildung von hydroxysubstituierten Chorismatderivativen nutzen. Gemäß der Vorstellung, dass sich sekundärmetabolische Enzyme aus den evolutionär

älteren primärmetabolischen Enzymen entwickelt haben, wurde in diesem Teil der Arbeit zunächst untersucht, inwiefern der Übergang von primär- zu sekundärmetabolischen MST-Enzymen mit dem Übergang von Ammoniak auf Wasser als Nukleophil zusammenhängt. Dazu wurden zunächst mit netzwerkbasierten, phylogenetischen und strukturbasierten *in silico* Methoden zwei entscheidende Aminosäuren in einem Zugangskanal des aktiven Zentrums identifiziert, die sich signifikant zwischen den primärmetabolischen, Ammoniak-umsetztenden und den sekundärmetabolischen, Wasser-umsetztenden MST-Enzymen unterscheiden. Der Beitrag dieser beiden Aminosäuren zur Substrat- bzw. Nukleophilspezifität wurde mittels rationalem Design von sechzehn Varianten des primärmetabolischen MST-Enzyms Anthranilatsynthase überprüft. HPLC-MS Experimente zeigten, dass verschiedene Kombinationen von zueinander passenden Aminosäuren an den beiden identifizierten Positionen zu einer Verbreiterung der Nukleophilspezifität von Ammoniak auf Wasser führen. Die aktiven Varianten waren in der Lage, Chorismat unter Verwendung von Wasser als Nukleophil zu Isochorismat umzuwandeln; deren Effizienz war dabei mit der von nativen Isochorismatsynthasen vergleichbar. Diese Varianten konnten ebenso Ammoniak als Nukleophil für die Bildung von Anthranilat zu nutzen, waren also bifunktional. Zusammenfassend wurde durch diese Experimente gezeigt, dass die Nukleophilspezifität von MST-Enzymen ohne große Hürden von Ammoniak auf Wasser ausgedehnt werden kann. Die beobachtete Bifunktionalität der Anthranilatsynthasevarianten lässt vermuten, dass die Evolution der sekundärmetabolischen MST-Enzyme über vergleichbare, bifunktionale Enzymintermediate verlaufen ist. Solche, oft als metabolische Generalisten bezeichnete, Enzyme könnten die Bildung von neuartigen Sekundärmetaboliten (Isochorismat) unter gleichzeitiger Aufrechterhaltung der Biosynthese wichtiger primärer Metaboliten (Anthranilat) erlaubt haben. Ein solches evolutionäres Szenario setzt nicht zwingenderweise eine initiale Genduplikation voraus und vermeidet folglich negative metabolische Effekte, die beispielsweise aus der Beibehaltung von duplizierten, redundanten Genkopien resultieren.

Der dritte Teil dieser Arbeit befasst sich mit der Problematik, wie spezifische Protein-Protein Interaktionen in Superfamilien von strukturell ähnlichen Proteinkomplexen sichergestellt werden und wie die für Interaktionsspezifität entscheidenden Faktoren evolviert sind. Praktisch alle zellulären Prozesse sind essentiell abhängig von spezifischen Interaktionen zwischen Proteinen. Die nötige Spezifität wird generell durch das Zusammenspiel von räumlich und elektrostatisch komplementären Proteinoberflächen, den so genannten *interfaces*, erreicht. Allerdings ist die Zahl der in der Natur vorkommenden Proteinstrukturen und folglich auch die Zahl der Protein *interfaces* begrenzt; dieser Umstand ist bedingt durch den Zwang bei der Proteinfaltung die 2D-Strukturelemente räumlich möglichst

effizient zu packen. Zentrale Fragen sind daher, wie trotz der strukturellen Limitationen ausreichende Interaktionsspezifität sichergestellt und wie die Bildug unphysiologischer Komplexe vermieden wird, wenn mehrere Interaktionspartner mit strukturell ähnlichen *interfaces* vorhanden sind. Um sich diesen Problemen zu nähern, wurden in diesem Teil der Arbeit zunächst die *interface* Geometrien von über 300 bakteriellen, heteromeren Proteinkomplexen im Vergleich zu ihren homologen Verwandten aus den entsprechenden Superfamilien bioinformatisch ausgewertet. Diese Untersuchung zeigte, dass in 10% aller Komplexe, die strukturell ähnliche Untereinheiten besitzen, deutliche Variationen in den *interface* Geometrien auftreten. In diesen Fällen sind die *interfaces* durch zusätzliche, definierte Sekundärstrukturelemente, die so genannten *interface add-ons*, erweitert, die aus ca. 10-20 Aminosäuren bestehen und signifikante Beiträge zur Stabilität der Proteinkomplexe leisten. Aufgrund dieser Eigenschaften lässt sich folgern, dass *interface add-ons* als eine Art negatives Designelement wirken, dadurch strukturell ähnliche Proteinkomplexe voneinander differenzieren und somit zur Interaktionsspezifität von Proteinkomplexen beitragen. Um diese Vermutung experimentell zu untersuchen, wurde anschließend eine Superfamilie von Glutaminamidotransferasekomplexen, die an der Biosynthese von Tryptophan und Folsäure beteiligt sind, genauer untersucht: Diese heteromeren Komplexe bestehen aus Synthase- und Glutaminaseuntereinheiten, welche miteinander interagieren, um Ammoniak von Glutamin auf ein Akzeptorsubstrat zu übertragen. Ein Teil der Glutaminamidotransferasen dieser Superfamilie – ausschließlich solche, die an der Biosynthese von Tryptophan beteiligt sind – trägt in der Synthaseuntereinheit ein *interface add-on*. Alle anderen homologen Synthaseuntereinheiten der Superfamilie – auch diejenigen, die an der Biosynthese von Folsäure beteiligt sind – besitzen dieses *interface add-on* nicht. Die umfassende Charakterisierung von insgesamt 54 Kombinationen verschiedener Synthase- und Glutaminaseuntereinheiten mittels chromatografischer, biophysikalischer, massenspektrometrischer und enzymkinetischer Methoden zeigte eindeutig, dass die An- bzw. Abwesenheit des *interface add-ons* in der Synthaseuntereinheit die Spezifität für die Interaktionen mit den Glutaminaseuntereinheiten bestimmt. *In silico* erstellte Profile von über 15 000 archaeellen und bakteriellen Spezies sowie *in vivo* Wachstumsexperimente machten weiterhin deutlich, dass das *interface add-on* der Komplexe aus der Tryptophanbiosynthese biologische Relevanz besitzt: Das *interface add-on* verhindert Kreuzinteraktionen mit den Untereinheiten des homologen Komplexes aus der Folsäurebiosynthese, welche sich ansonsten negativ auf die zelluläre Fitness auswirken und das Zellwachstum behindern. Abschließend wurde anhand von rational designten Glutaminaseuntereinheiten aufgezeigt, dass in der Evolution des *interface add-ons* dieser Superfamilie vermutlich intermediäre Komplexe mit breiterer

Interaktionsspezifität aufgetreten sind. Bei *interface add-ons* handelt es sich folglich um evolutionäre Instrumente, die es ermöglichen, spezifische Protein-Protein Interaktionen zu etablieren, wenn aufgrund von ähnlichen *interface* Geometrien zwischen mehreren potentiellen Interaktionspartnern unterschieden werden muss.

Zusammenfassend vertieft die vorliegende Arbeit das Verständnis der Mechanismen, die der Evolution von enzymatischer Mono- und Bifunktionalität zu Grunde liegen, unterstreicht die Bedeutung von multifunktionellen, generalistischen Enzymen für die Evolution des Sekundärmetabolismus und beschreibt ein bisher nicht erkanntes strukturelles Element zur evolutionären Spezifizierung von Protein-Protein Interaktionen.

List of Publications

This cumulative dissertation is composed of the following published or submitted manuscripts:

A **Plach, M.G.**, Reisinger, B., Sterner, R., and Merkl, R. (2016). Long-term persistence of bi-functionality contributes to the robustness of microbial life through exaptation. *PLoS Genetics* 12:e1005836

B **Plach, M.G.**, Löffler, P., Merkl, R., and Sterner, R. (2015). Conversion of anthranilate synthase into isochorismate synthase: Implications for the evolution of chorismate-utilizing enzymes. *Angewandte Chemie International Edition* 54:11270-11274

C **Plach, M.G.**, Semmelmann, F., Busch, F., Busch, M., Heizinger, L., Wysocki, V.H., Merkl, R., and Sterner, R. (2017). Evolutionary diversification of protein-protein interactions by interface add-ons. *Submitted for Publication*

In the course of this work, I contributed to further publications, which are not part of the dissertation:

D Veprinskiy, V., Heizinger, L., **Plach, M.G.**, and Merkl, R. (2017). Assessing *in silico* the recruitment and functional spectrum of bacterial enzymes from secondary metabolism. *BMC Evolutionary Biology*. Accepted.

E Kandlinger, F., **Plach, M.G.**, and Merkl, R. (2017). AGeNNT: annotation of enzyme families by means of refined neighborhood networks. *Submitted for publication*

Personal Contributions

Publication A
The research was designed by myself, Bernd Reisinger, Reinhard Sterner, and Rainer Merkl. The experiments were performed by myself and Bernd Reisinger in equal parts. Ancestral sequence reconstruction was done by Rainer Merkl. The work was supervised by Reinhard Sterner and Rainer Merkl, and the publication was written by myself, Reinhard Sterner, and Rainer Merkl.

Publication B
The research was designed and the experimental was work performed by myself. Patrick Löffler performed molecular dynamics simulations. The work was supervised by Rainer Merkl and Reinhard Sterner, and the publication was written by myself, Patrick Löffler, Rainer Merkl, and Reinhard Sterner.

Publication C
The research was designed by myself, Florian Semmelmann, Rainer Merkl, and Reinhard Sterner. Cloning, expression, purification, and experimental characterization of glutaminases and synthases were done by myself and Florian Semmelmann. Florian Busch performed mass spectrometry experiments and was supervised by Vicky Wysocki. All other experiments were performed by myself. Computational tools and scripts were provided by myself, Markus Busch, and Leonhard Heizinger. Rainer Merkl and Reinhard Sterner supervised the work. The publication was written by myself, Florian Semmelmann, Rainer Merkl, and Reinhard Sterner.

Table of contents

1 General Introduction

1.1 A short history of evolution

Life in its entire spectrum formally bursts with complexity. One, for instance, might think of the remarkable biosynthetic capabilities of microorganisms, the specialization and connectivity of individual cells in multi-cellular plants and animals, or the highly sophisticated cognitive abilities of the human brain. These developments did, of course, not come overnight but rather from a painstakingly slow but gradual adaptation process. This process, referred to as *evolution*, was first formulated as a scientific theory in the early 19th century (Hoyle and Wickramasinghe, 2000). At that time scholars recognized that the driving forces behind the differentiation of life into the countless species were constant adaptation to ecological niches and competition for resources (Darwin, 1859).

But what did evolution begin with? One longstanding theory suggests that life itself started in the primordial oceans. During a time-period with a reducing atmosphere (Lazcano and Miller, 1996; Oparin and Morgulis, 2003) complex organic compounds formed from carbon dioxide or methane under the influence of ultraviolet light or electricity (Bada, 2013; Miller, 1953). Despite some disagreements (Bernhardt, 2012), it is widely acknowledged that RNA was among these first compounds (de Farias et al., 2016; Gilbert, 1986). This "RNA world" concept suggests that RNA molecules could catalyze simple biochemical reactions and also store the information needed for their replication. The concept is supported by the observation that some modern biosynthetic machineries like ribosomes contain catalytic RNA molecules and by the simple enzymatic reduction of ribose nucleotides that yields deoxyribose nucleotides as they appear in the later emerged DNA (Higgs and Lehman, 2015).

Evolution really kicked off when the information stored in RNA and DNA became meaningful in early forms of genes. These served as blueprints for small primordial peptides that were able to interact with RNA and to support RNA-based catalysis (Alva et al., 2015). These small peptides are thought to have gradually fused, multiplied, and recombined to eventually give rise to the relatively small set of not more than a thousand different modern protein geometries or folds (Chothia, 1992). Recent experimental discoveries identified sets of primordial peptides (Farías-Rico et al., 2014), showed that they can self-assemble and form larger, biochemically active proteins (Smock et al., 2016), and highlighted that duplication and fusion can lead to symmetric proteins structurally similar to modern ones (Longo and Blaber, 2014; Park et al., 2015; Richter et al., 2010).

The inclusion of such primitive proteins together with RNA, DNA, and other organic molecules in cell-like bubbles made of amphipathic fatty molecules was the basis for first simple self-replicating units. These units eventually formed the last universal common ancestor (LUCA) of all living things, which existed approximately between 3.8 and 4.5 billion years ago (Nisbet and Sleep, 2001). It is assumed that this enigmatic single entity was rather simple (Di Giulio, 2011; Koonin, 2003; Woese, 1998), yet its true nature and physiology is still widely debated (Kim and Caetano-Anollés, 2011; Weiss et al., 2016). For instance, analyses of gene content, co-evolution, and proteomes favor a more sophisticated LUCA physiology (Doolittle, 2000; Ouzounis et al., 2006; Tuller et al., 2010). Originating from the LUCA, formation and diversification of species ("macro-evolution") took on, first yielding uni-cellular microorganisms like archaea and bacteria and eventually the evolutionary young multi-cellular plants and animals. Given some remarkable similarities between different animal species or between animals and plants, this common origin has already been noticed in the 17th century (Hoyle and Wickramasinghe, 2000). Indeed, modern methods of genetic analyses allow one to reconstruct the evolutionary relationship of all known species back to their common root in a so-called tree-of-life (**Figure 1**).

Finally, there is one more point to stress out: Evolution is no completed process that has begun sometime in the past and that has led to all the species and organisms we know today. On the contrary, evolution is highly dynamic and new species, new organisms, and new functions constantly emerge. This is best seen from the rapid development of microbial drug resistance against common antibiotics that have only been in use for nearly a decade (Barlow and Hall, 2002; Boyanova et al., 2015; Hall, 2004) or the appearance of microbial enzymes that are capable of degrading chemicals that have not been in the ecosystem for longer than 50 years (Copley, 2009; Hartley et al., 2006; Janssen et al., 2005; Seffernick and Wackett, 2016). The herbicide atrazine, for example, was invented in 1958 and has been applied in millions of tons since then. Already 40 years later, bacterial strains were identified that are able to degrade atrazine completely and metabolize its degradation products (Boundy-Mills et al., 1997; de Souza et al., 1998b). The atrazine-hydrolyzing enzymes produced by these bacteria are assumed to have emerged very recently from amidohydrolases like melamine deaminase in soil-living Pseudomonas species (de Souza et al., 1998a; Seffernick et al., 2001; Shapir et al., 2007).

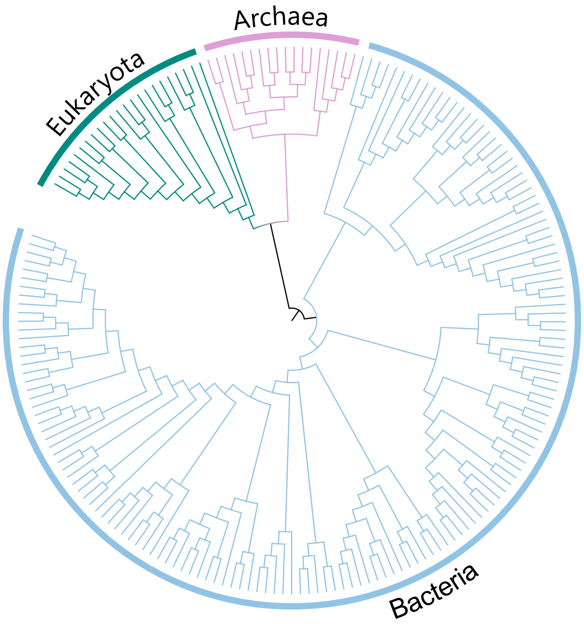


Figure 1. Tree-of-life.
Phylogenetic tree, as constructed from the comparison of 31 concatenated genes that occur in 191 species with fully sequenced genomes (Ciccarelli et al., 2006). In this illustration the root of the tree is in the middle and the branches radiate outwards, with each new species descending from a common ancestor resulting in a new branch. Note that branch lengths are not to scale. Bacterial and archaeal species are spanned by blue and magenta arcs, respectively. Eukaryotic species are marked by a green arc.

1.2 The driving forces of evolution

The common ancestry of all species and the relationship between them makes clear how new species emerge: By segregation from already existing ones. Such speciation events take place when two populations of the same species get separated by some kind of barrier, for instance an entrapment in an ecological niche, and adapt to their respective conditions. Over time the populations will genetically drift apart, up to a point where genetic exchange between the daughter and the parental population is no longer possible (Safran and Nosil, 2012). Evolution is thus driven by constant adaptation processes to changing environments through the emergence of novel functionalities, which are eventually just novel genes that lead to novel proteins with novel capacities.

But how do novel genes arise? In fact, this is simply a matter of chance. DNA is not just an information carrier that is filed somewhere inside a cell but is constantly transcribed into RNA blueprints for proteins and replicated for genetic reproduction. Although a highly complex and sophisticated catalytic machinery has developed for these essential tasks, errors occasionally occur. For instance, sometimes the wrong nucleotide is incorporated into replicated DNA thus changing the genetic information (Kunkel and Bebenek, 2000). Additionally, response systems can fail to repair DNA damage coming from the exposure to UV light or oxidizing compounds and different parts of a genome can be recombined due to viruses or erroneous DNA ligation (Friedberg, 2003; Hakem, 2008).

These changes in the genetic information are known as mutations, which can be classified into insertions, deletions, and substitutions, depending on whether additional nucleotides were inserted into the sequence, a stretch of nucleotides was lost, or a single nucleotide was changed to another one. Although the effects of mutations in protein coding genes are buffered by a redundancy in coding nucleotide triplets, many mutations are deleterious. For instance, a mutation may lead to a loss of function by substituting a hydrophobic residue in the protein core with a polar one, which could cause the protein to misfold. Likewise, substituting a catalytic residue in an enzyme might render it inactive. For uni-cellular organisms like bacteria such mutations can have lethal consequences, clearing the corresponding mutants from the population and the gene pool. However, by chance, some mutations will occur at just the right spot; they may lead to a more stable protein by filling a hole in its structure with a bulky amino acid, may allow the defense against antibiotic agents by modifying the catalytic capacity of a hydrolase enzyme, or increase the catalytic efficiency of an enzyme by introducing an amino acid that aids in binding its substrate. Such mutations lead to novel functionalities that can give the mutant a lead in competition and adaptation, eventually allowing it to prevail in the population and conserve the new functionality.

1.3 The concept of homology

A great caveat of mutational adaptation is, as mentioned above, that most mutations are deleterious and severely impair the viability of organisms. Today, evolution is thus no longer viewed as a strictly linear, gradual adaptation process but more as a drift through a landscape of neutral mutations, which are neither beneficial nor harmful, with random fixation of mutations (Kimura, 1983). Moreover, it has been realized that other processes of greater genetic variation must exist that enable the accumulation of mutations while buffering their negative effects. Ohno postulated in his seminal work *Evolution by Gene Duplication* (Ohno, 1970) that the duplication of a gene or even larger genomic segments creates redundant copies that are free from selective pressure, which allows for the accumulation of otherwise detrimental or lethal mutations.

Genes that are related through such duplication events and that have diverged in function are formally referred to as *paralogs* (**Figure 2**). The frequency of paralogs in bacterial genomes is quite high; already in a genome comprising 2000 genes about one fourth of the genes are paralogs (Hooper and Berg, 2003). Although the term paralog, in its original definition referred to genes (Fitch, 1970), it is commonly used to describe the relationship of proteins. In this sense, paralogous proteins share many properties like their overall three-dimensional structure but often differ in crucial features like enzymatic function or stability. For example, the enzyme 1-(5-phosphoribosyl)-5-[(5-phosphoribosylamino)-methylideneamino]-imidazole-4-carboxamide isomerase (HisA) and the cyclase subunit of the imidazole glycerol phosphate synthase (HisF) share the same fold and are similar with respect to their amino acid sequences, but have evolved quite different catalytic activities for the biosynthesis of the amino acid histidine.

In contrast to paralogs, *orthologs* are genes that are related through speciation events and that have retained their original function in the descendant species (Fitch, 1970), because they are still subject to the same selective pressure as in the paternal species (**Figure 2**). Following the example from above, the HisA enzymes from *Thermotoga maritima* and *Salmonella enterica* are orthologs, because they both have descended from their common HisA ancestor and they both still catalyze the fourth step of histidine biosynthesis.

In a generalization of these two concepts, genes and their products are called *homologs*, if they are related to each other by descent from a common ancestor through a series of duplication and speciation events (**Figure 2**). However, it is complicated to infer homology for genes or proteins that have diverged from a common ancestor billions of years ago, because their sequences often show hardly any similarities. Thus, a more practical definition of homology is applied today: Two genes (or two proteins) are homologs, if they are simply similar enough that it is unlikely that this similarity arose just by coincidence. The most common lower threshold for inferring homology of proteins is a minimum of 20-30% identical amino acids across the length of their aligned sequences, as this value correlates with an almost certain structure similarity (Orengo and Thornton, 2005).

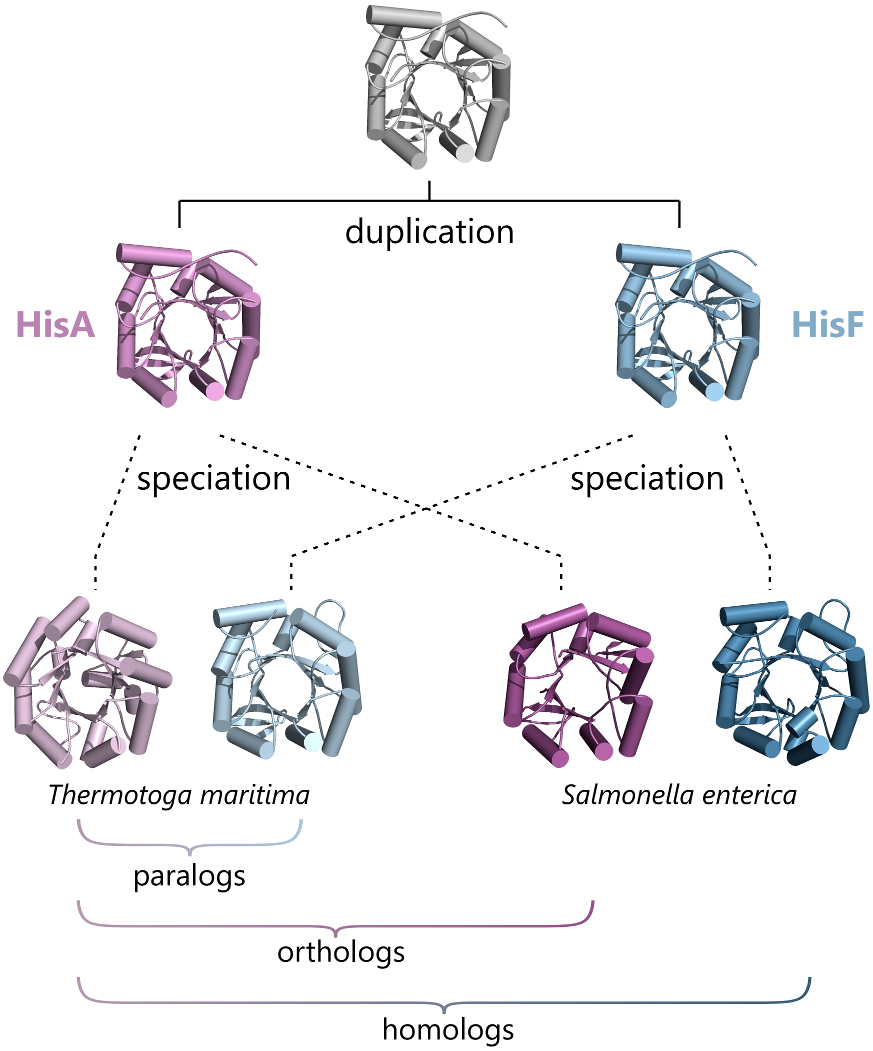


Figure 2. Illustration of the concepts of paralogs, orthologs, and homologs.
The enzymes HisA (magenta) and HisF (blue) have emerged from a common ancestor by gene duplication but have evolved different functions in the biosynthesis of the amino acid histidine. Their relation is apparent from structural similarity and the similar amino acid sequences of modern HisA and HisF enzymes (41% and 42% for the pairs from *S. enterica* and *T. maritima*, respectively, determined with EMBOSS Needle (Rice et al., 2000)). HisA and HisF are paralogs, because they are related through a duplication event and have diverged in their respective functions. In contrast, orthologs are genes or proteins that are related through a speciation event and that have retained the same function. In general, homologs are genes or proteins that are related through descent from a common ancestor by a series of duplication and speciation events. The homology between HisA from *T. maritima* and HisF from *S. enterica* is evident from their structural similarity and the high degree of sequence similarity (43%, determined with EMBOSS Needle).

1.4 Protein superfamilies

Homology is no completely precise concept: For instance, closely related homologs may share almost identical structures and perform very similar functions, like the catalysis of the same reaction with just a slightly different substrate. On the other hand, distantly related proteins that may have little in common – their structures may be rather strong variations of the same core fold and they may catalyze completely different chemistries – are counted as homologs as long as their sequences are similar enough. Thus a more fine-grained classification of homology in the context of proteins is necessary. Such is provided by combining homologs that have highly similar sequences (e.g. >40% identity) and that also share common functional properties into *families*. In turn, several families can be further grouped together to larger entities whose members have less similar sequences (generally <30% identity) and have fewer traits in common, for instance only the overall three-dimensional structure or a core reaction chemistry. Margaret Dayhoff first coined the term *superfamily* for these groups to reflect their super-ordinate character and broader evolutionary relationship compared to families (Dayhoff, 1965). Superfamilies are frequently functionally diverse (Almonacid and Babbitt, 2011); only about 60% of the proteins from the same superfamily also have the same function (Hegyi and Gerstein, 2001) and this variability can even span all six Enzyme Commission classes in some enzyme superfamilies (Baier et al., 2016). To summarize, the term superfamily shall be defined within the scope of this thesis as follows: A superfamily is a group of homologous proteins that (i) are derived from a common evolutionary ancestor, (ii) hence share similarities in sequence, structure, and – for enzymes – core reaction mechanism, but (iii) differ in their specific functions.

The first comprehensive implementation of the family-superfamily concept was achieved in the *Structural Classification of Proteins* database (SCOP) (Murzin et al., 1995). This database provides a hierarchical classification of proteins into families and superfamilies, based on evolutionary relationship, and into folds and classes, based on structural features and similarities (**Figure 3**). For example, different glucosidases and glycosyltransferases that all act on α-linked oligosaccharides are summarized in the α-amylase family. This family is, in turn, grouped together with other families like the β-galactosidases, -glucanases, and -amylases to the superfamily of glycosyltransferases. While families are collections of related but very similar enzymes, superfamilies provide an overview of the different functions that have been established on a common protein scaffold and the evolutionary relationship between these different functions.

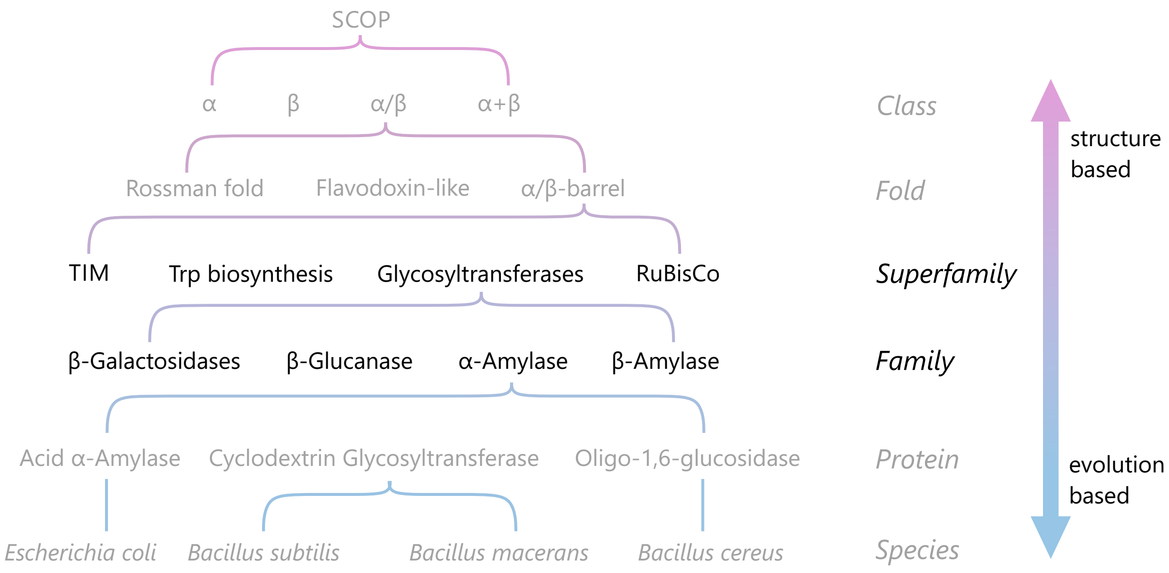


Figure 3. Hierarchical illustration of the SCOP database.
The hierarchical classification of proteins in the SCOP database comprises six levels. At the bottom *species* represent distinct protein sequences, which are grouped together with similar proteins of essentially the same function on the *protein* level. The *family* level summarizes proteins with highly similar sequences and similar functions. The *superfamily* level combines individual families with similar structural features, a common evolutionary origin, but different functions. Above this level *fold* and *class* group superfamilies based on structural features and do not imply homology. The figure is adapted from the SCOP version 1.75 documentation (Andreeva et al., 2008).

A look at frequently used databases that classify proteins into families and superfamilies points out, however, that the definition of a superfamily is not general and that it is often not clear where to draw the line between a family and a superfamily. For example, the latest release of the SCOP database (version 1.75, June 2009) lists 1962 superfamilies and 3902 families. More recent and more sophisticated databases like InterPro (version 60.0, November 2016) and Pfam (version 30.0, June 2016) yet list 1749/19851 and 595/16306 corresponding entries, respectively. These databases have been created using high-performance and high-sensitive algorithms that accurately assign new protein sequences to the correct families and superfamilies (Finn et al., 2016; Mitchell et al., 2014).

An equally challenging task is to put the inherent information of protein superfamilies on evolutionary relationship and functional diversity to use. Traditional phylogenetic approaches to dissect and sub-divide a superfamily are often problematic, because the data-sets frequently contain tens of thousands of sequences and the amino acid similarities between iso-functional sub-families can be very low (Punta et al., 2012). A more

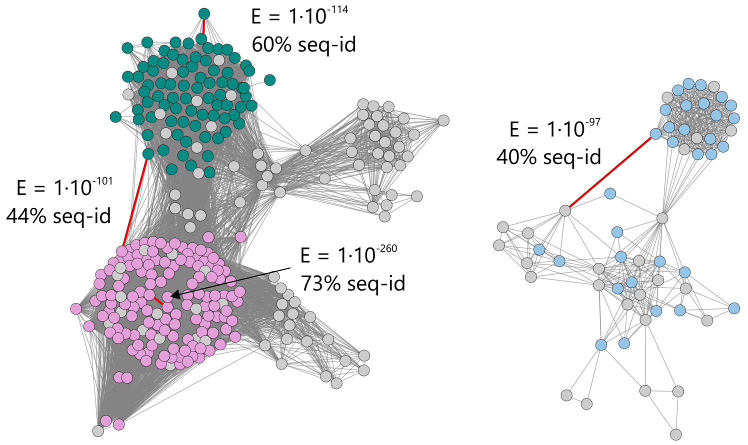


Figure 4. Visualization of pairwise sequence similarity relationships between members of a protein superfamily in a sequence similarity network.
The SSN shows the relationship between selected members of the glycosyl transferase superfamily (InterPro entry IPR000312), which comprises structurally similar but functionally distinct AMP-phosphorylases (green), thymidine phosphorylases (magenta), and anthranilate phosphoribosyltransferases (blue). The network was generated at a similarity cut-off of 10^{-97} (BLAST E-value9 and is shown in a force-directed layout generated with Cytoscape 3.4 (Shannon et al., 2003). Selected edges are marked in red and annotated with their respective E-value and the corresponding sequence identity (seq-id) of the connected sequences.

robust and computationally less demanding alternative nowadays widely used for processing superfamilies are sequence similarity networks (SSNs), developed by Babbitt and co-workers (Atkinson et al., 2009) and made available by the *Enzyme Function Initiative* (Gerlt et al., 2015). SSNs rely on the pre-defined protein family and superfamily data-sets from InterPro or Pfam and are constructed from an all-by-all sequence comparison between all members of a superfamily using the well-established BLAST algorithm (Altschul et al., 1990). The key of SSNs is that these inter-relationships can be visualized as a network graph (**Figure 4**). In such networks, nodes represent individual members of the superfamily (protein sequences) and connecting edges represent the pairwise sequence relation (sequence similarity score or BLAST E-value), if the corresponding proteins are more similar than a pre-defined threshold.

For most thresholds closely related members of a superfamily will be inter-connected more densely than distantly related members. With the threshold chosen appropriately for a specific superfamily, the network can reveal iso-functional groups within a superfamily. For example, an exemplary SSN of the glycosyltransferase superfamily (**Figure 4**) shows that AMP phosphorylases and thymidine phosphorylases are more closely related to each other than to the anthranilate phosphoribosyltransferases. This simple visual analysis tells much about the evolution of this superfamily and the functional relationship between its members. It is thus hardly surprising that superfamily data-sets together with SSNs have been used frequently with great success to get insight about enzyme evolution (Brown and Babbitt, 2014), to discover and predict enzymatic activities of uncharacterized proteins (Baier et al., 2016; Chow et al., 2015; Gerlt et al., 2015; Zhao et al., 2014), and to analyze the sequence-structure-function relationship, the fundamental axiom of modern molecular biology (Mashiyama et al., 2014).

1.5 Aim and scope of this work

Although the number of protein and whole genome sequences as well as the number of elucidated protein structures and functions have soared in the last decade, the evolutionary relationship between proteins and the paths that evolution took from the ancient ancestral entities to the modern ones is still enigmatic in all but few cases. And yet the lessons learned from unraveling, for example, how a specific enzymatic function is distributed across the members of the respective superfamily and how this function might have been established during evolution are the most important; first and foremost for a better understanding of the principles of biology and evolution. But these lessons will also tell us which enzyme scaffolds might be used to alter their catalytic capabilities to a desired function and how the transition to the new function might be achieved in the laboratory or via computer-aided design. Eventually, these lessons will help us to better understand the sequence-structure-function relationship, which in turn could tell what functional aspects or structural features of a protein superfamily might be targeted by drugs that could aid human health.

To study the evolution of protein traits and characteristics, one heavily relies on exploiting homology relationships between proteins. Superfamilies are of great value in this context, because they provide a formal but also biologically sensible framework to sub-divide and dissect groups of related proteins and – with the right tools in hand – to draw far-reaching conclusions and gain valuable insights. In this thesis, these concepts are used together with computational, biochemical, and biophysical techniques to investigate the evolution of enzymatic mono- and bi-functionality, the determinants of substrate specificity in enzyme superfamilies, the evolution of secondary metabolic enzymes, as well as the structural determinants of protein-protein interactions, their evolutionary history, and their physiological significance.

1.6 Guide to the following chapters

The following three chapters each deal with one of the three first-author publications (**A**-**C**) that constitute this thesis. These chapters are not intended to repeat the publications one-to-one. They are rather formulated as synopses that give the reader a profound introduction into the individual concepts and ideas important for the respective topics, highlight the key findings of each publication, and discuss these with attention to the corresponding literature.

In the chapter *Long-term persistence of bi-functionality contributes to the robustness of microbial life through exaptation* SSNs and ancestral sequence reconstruction (ASR) are used to demonstrate that the evolution of modern sugar isomerases with different catalytic functions originated from a bi-functional common ancestor. The chapter further describes that this bi-functionality has persisted to a certain degree in modern members of the corresponding sugar isomerase superfamily.

In the following chapter *Conversion of anthranilate synthase into isochorismate synthase: Implications for the evolution of chorismate-utilizing enzymes* the functional divergence within a superfamily of enzymes is examined and key catalytic residues that determine substrate specificity of these enzymes are identified. It is further demonstrated that this information can be used to modify the function of enzymes from this superfamily and finally the implications of this functional change for the evolution of secondary metabolic enzymes from primary metabolic enzymes are highlighted.

In the final chapter *Evolutionary diversification of protein-protein interactions by interface add-ons* structural determinants of protein-protein interaction specificity are identified from an extensive analysis of heteromeric protein complex structures. A superfamily of glutamine amidotransferases is then used to demonstrate that these structural features prevent the formation of non-physiological complexes and an evolutionary scenario for the emergence of these structural elements is provided.

2 Long-term persistence of bi-functionality contributes to the robustness of microbial life through exaptation *(Synopsis of Publication A)*

2.1 Introduction

2.1.1 The superfamily of ribulose-phosphate binding $(\beta\alpha)_8$-barrels

The *ribulose-phosphate binding barrel* superfamily (SCOP 51366) comprises, among others, several iso-functional families of $(\beta\alpha)_8$-barrel enzymes from histidine and tryptophan biosynthesis. One of its members is the enzyme HisA, which catalyzes the isomerization of the aminoaldose N'-[(5'-phosphoribosyl)-formimino]-5-aminoimidazole-4-carboxamide-ribonucleotide (ProFAR) to the aminoketose N'-[(5'-phosphoribulosyl)-formimino]-5-aminoimidazole-4-carboxamide-ribonucleotide (PRFAR) (Henn-Sax et al., 2002) for the fourth step of the biosynthesis of the essential amino acid histidine (**Figure 5A**). Another prominent member of this superfamily is the enzyme TrpF, which catalyzes a similar isomerization reaction involving the aminoaldose N-(5'-phosphoribosyl)anthranilate (PRA) and the aminoketose 1-(o-carboxyphenylamino)-1-deoxyribulose-5-phosphate (CdRP) in the biosynthesis of tryptophan (Hommel et al., 1995). The common evolutionary history of HisA and TrpF is not only evident from their highly similar structures and their high degree of sequence similarity (List et al., 2011) but is also supported by establishment of PRA isomerase activity on HisA scaffolds through both random mutagenesis (Jürgens et al., 2000) and spontaneous mutations under selective pressure (Näsvall et al., 2012).

The InterPro database associates HisA with the phosphoribosylformimino-5-amino-imidazole-carboxamide-ribotide isomerase superfamily (InterPro entry IPR023016). In the current InterPro release (version 60.0, November 2016) this entry comprises 9565

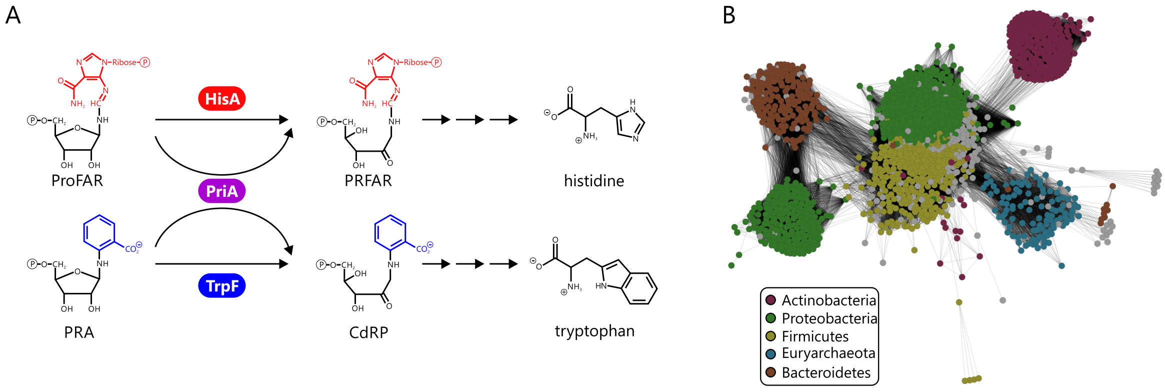


Figure 5. Reactions catalyzed by the homologous enzymes HisA, PriA, and TrpF and a SSN representation of the HisA InterPro family.
(**A**) HisA and TrpF catalyze analogous Amadori rearrangements of the aminoaldoses ProFAR and PRA to the corresponding aminoketoses PRFAR and CdRP in histidine and tryptophan biosynthesis, respectively. The bi-functional enzyme PriA catalyzes both reactions. (**B**) SSN of the phosphoribosylformimino-5-aminoimidazole-carboxamide-ribotide isomerase superfamily. Nodes are colored by the five main bacterial phyla contributing to the superfamily. The figure is adapted from publication A.

members, which mainly correspond to HisA enzymes from bacterial and archaeal species. An SSN shows that the superfamily is basically iso-functional, with a certain sub-cluster structure as a result of phylum specific sequence variations (**Figure 5B**). The most uniform and most separated sub-cluster is that of Actinobacteria. Most Actinobacteria do not possess HisA enzymes, but rely for the biosynthesis of histidine and tryptophan on a homolog of HisA called phosphoribosyl isomerase A (PriA). PriA is a bi-substrate specific enzyme processing ProFAR as well as PRA with equal efficiency (Barona-Gómez and Hodgson, 2003). Actinobacteria that possess a *pri*A gene lack a dedicated *trp*F gene, and consequently PriA is part of both histidine and tryptophan biosynthesis. The genomic neighborhoods of *his*A and *pri*A, i.e. the genes upstream and downstream, are highly similar, further indicating that PriA is a homolog of HisA (Publication E). Structural comparison with HisA makes clear that PriA utilizes the same two phosphate binding sites formed by $\beta\alpha$-loops three and four as well as $\beta\alpha$-loops seven and eight to bind ProFAR (Due et al., 2011; Wright et al., 2008). The C-terminal site used for binding PRA is also identical to the single phosphate binding site of its superfamily relative TrpF (Due et al., 2011; Henn-Sax et al., 2002). Its quite unusual characteristics have made PriA a well-examined member of this superfamily with respect to structure and catalytic mechanism (Due et al., 2011; Küper et al., 2005; Wright et al., 2004) as well as its evolutionary relationship with HisA and TrpF (Noda-García et al., 2013, 2015; Verduzco-Castro et al., 2016).

2.1.2 Promiscuity and bi-functionality

Enzymes are traditionally viewed as remarkably specific catalysts that process only one out of the countless possible substrates present in cells with extremely high catalytic efficiency. Casually speaking, the textbook description of enzyme catalysis is *one enzyme – one substrate – one reaction.* However, it has long been acknowledged that at least some enzymes are actually capable of catalyzing more than one reaction. Early reported examples are carbonic anhydrase (Pocker and Stone, 1965), pepsin (Reid and Fahrney, 1967), chymotrypsin (Nakagawa and Bender, 1969), and L-asparaginase (Jackson and Handschumacher, 1970). In the recent years, more and more enzymes were identified and described in detail that possess in addition to their native function some kind of alternative activity; see the reviews by Hult and Berglund (2007), Nobeli et al. (2009), and Khersonsky and Tawfik (2010).

The term *promiscuity* was first introduced by O'Brien and Herschlag (1999) to describe such alternative activities but without any generally applicable definition to it. However, there has formed an understanding that promiscuity refers to enzymatic activities that (i) are different from the native activity for which an enzyme has evolved and that (ii) are not part of the physiology and metabolic network of a corresponding organism (Copley, 2003; Khersonsky and Tawfik, 2010). In other words, promiscuous enzymes coincidentally catalyze more than one reaction that differ in the third, second, or even first digit of their Enzyme Commission numbers, indicating different classes of substrates or even different reaction categories and mechanisms (Bornscheuer and Kazlauskas, 2004). Also, the catalytic efficiencies for the promiscuous reactions are usually several orders of magnitude lower than those for the native reaction (Khersonsky and Tawfik, 2010). Promiscuous enzymes have been identified in many protein families and superfamilies: The enolase superfamily for instance contains the *o*-succinylbenzoate synthase that has a promiscuous *N*-acylaminoacid racemase activity, albeit with a 1000-fold lower catalytic efficiency (Palmer et al., 1999). To name one more, a member of the amidohydrolase superfamily, the dihydroorotase, has a promiscuous phosphotriesterase activity with a 10^6-fold lower catalytic efficiency (Roodveldt and Tawfik, 2005). Today it is assumed that on average at least half of the members of large and diverse enzyme superfamilies display either substrate or functional promiscuity (Baier et al., 2016).

For some enzymes, on the contrary, there is almost no difference in activities for the different reactions they catalyze, so that the terms promiscuous, native, and alternative are not really justified in these cases. For instance, cytochrome P450s or detoxifying glutathione *S*-transferases can efficiently perform several more or less similar functions (Khersonsky et al., 2006; Zhang et al., 2012b). Such enzymes that catalyze different reactions with two or more substrates are termed *bi-* or *multi-functional.* In analogy to the definition of promiscuity above, bi-functional enzymes have *a priori* evolved to catalyze two reactions with comparable efficiency, which are also part of the physiology of the corresponding organism. Thus PriA is a classical bi-functional enzyme, because it converts ProFAR to PRFAR in histidine biosynthesis and PRA to CdRP in tryptophan biosynthesis with catalytic efficiencies of $10^4 - 10^5\ \mathrm{M}^{-1}\,\mathrm{s}^{-1}$ in both cases (Due et al., 2011; Küper et al., 2005). Also, both reactions are essential for the viability of Actinobacteria like *Streptomyces coelicolor* or *Mycobacterium tuberculosis* and thus neither of the two activities is promiscuous.

Before moving on, two more points should be stressed: First, real bi- or even multi-functional enzymes are rare; although frequently designated as bi-functional (Zhang et al., 2009), enzymes that possess two distinct active sites on the same polypeptide chain – for example as a result of gene fusion – do not meet the definition made above. Second, the terms promiscuity and bi-functionality only apply to naturally occurring enzymes and substrates. Even strictly mono-specific and non-promiscuous enzymes may still transform artificial substrates (Villiers and Hollfelder, 2009); this is because they have never been exposed to these substrates and thus the "artificial" activities have never been selected against.

2.1.3 Ancestral sequence reconstruction

The connection between promiscuity or multi-functionality and protein evolution was first made by Jensen in 1976 when he put forward his *patchwork hypothesis* (Jensen, 1976). He made the claim that, unlike modern enzymes that tend to be specialists, ancient enzymes featured very broad substrate specificities and thus only few multi-functional enzymes acted on multiple substrates, covering a wide range of early metabolic functions. Only later, divergence through gene duplication, mutation, and selection yielded the highly specialized extant enzymes.

Direct experimental evidence supporting or rebutting Jensen's hypothesis has remained elusive for quite some time, not least due to the lack of macromolecular fossils (Nisbet and Sleep, 2001). However, in recent years the technique of ancestral sequence reconstruction (ASR) has matured and now allows one to reconstruct and experimentally characterize extinct proteins. ASR goes back to the idea put forward by Zuckerkandl and Pauling over half a century ago that it should be possible to infer the sequence of an ancient protein from the amino acid composition of its modern descendants (Pauling and Zuckerkandl, 1963). And indeed the basic process of reconstructing the evolutionary ancestors of extant proteins follows their notion (**Figure 6**): Initially a set of sequences of

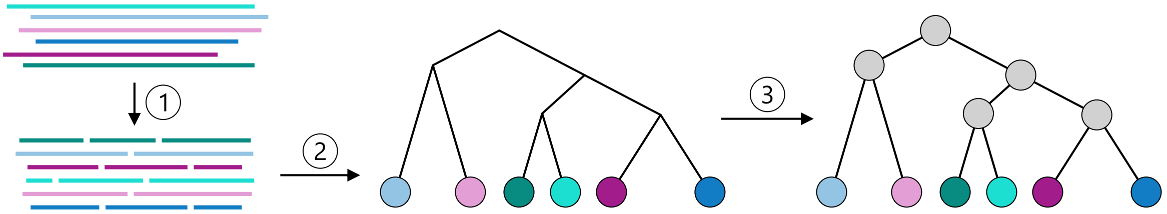


Figure 6. The strategy of ancestral sequence reconstruction.
ASR generally comprises three steps: First, a set of sequences of extant proteins is compiled and aligned. Second, this MSA is used to calculate a phylogenetic tree whose topology and branch lengths represent the evolutionary relationship between the extant proteins (colored circles). Third and last, precursor sequences (grey circles) can be reconstructed.

extant proteins is compiled and used to generate a multiple sequence alignment (MSA). This in turn serves to calculate a phylogenetic tree, which reflects the evolutionary relationship between the extant proteins. In such a tree every node that links two branches corresponds to an individual ancestral sequence and the branch lengths are a measure of the time lying between connected nodes. Computing this tree requires the use of an appropriate evolutionary model, which takes into account the probability of mutations, insertions, and deletions that led from a specific node to its child. Finally, the amino acid sequences at all internal nodes of the tree, back to its root, can be reconstructed. This process is, however, not as simple as that. The initial sequence set has to be carefully compiled and the correct topology of the phylogenetic tree has to be ensured. Also, the evolutionary model used to reconstruct ancestral sequences has to be chosen with caution, as it for instance has to take temporal changes of mutation rates into account. For these technical aspects of ASR the interested reader is referred to reviews on this topic (Liberles, 2007; Merkl and Sterner, 2016).

So far, ASR has been applied to reconstruct numerous protein ancestors that presumably existed millions of years (Wilson et al., 2015; Yokoyama et al., 2014) or billions of years ago in the Paleoarchaean era around the time of the enigmatic LUCA (Perez-Jimenez et al., 2011; Reisinger et al., 2013) or even in the pre-LUCA era (Fournier and Alm, 2015). Among the reconstructed and resurrected ancient proteins are visual (Chang et al., 2002; Shi and Yokoyama, 2003; Yokoyama et al., 2008) and fluorescent pigments (Field and Matz, 2010; Ugalde et al., 2004), elongation factors (Gaucher et al., 2008), steroid receptors (Bridgham et al., 2006; Carroll et al., 2011; Eick et al., 2012; Ortlund et al., 2007), and a number of enzymes; terpene synthases (Guzzetti et al., 2016), alcohol dehydrogenases (Thomson et al., 2005), kinases (Akanuma et al., 2013), glutamine amidotransferases (Reisinger et al., 2013), and tryptophan synthases (Busch et al., 2016) just to name a few.

Even more interesting than the reconstructed proteins themselves are the lessons to be learned from such experiments. For instance, the reconstruction of translation elongation factors (Gaucher et al., 2008) and thioredoxins (Perez-Jimenez et al., 2011) from 0.5-4 billion years ago demonstrated that the thermal stability of these proteins decreased by about 30 °C in that time span, which is in accordance with the temperature trend observed from the deposition of silicone isotopes (Robert and Chaussidon, 2006). Reconstructed ancestors have also been used to address crucial questions of molecular evolution, including the evolution of protein-ligand interactions (Eick et al., 2012; Kuang et al., 2006) and protein-protein interactions (PPIs) (Finnigan et al., 2012), the effect of gene duplication on evolutionary innovation (Voordeckers et al., 2012), the changes of conformational flexibility during evolution (Hart et al., 2014; Wilson et al., 2015), and the sophistication of enzyme complexes (Busch et al., 2016; Perica et al., 2014; Reisinger et al., 2013).

Coming back to Jensen's patchwork hypothesis, ASR has also been applied frequently to shed light on whether primordial enzymes were indeed multi-functional. However, the characterization of reconstructed ancestral enzymes has not yet brought forward a conclusive result. For example, ancestral variants of serine proteases or HisF were specific enzymes that could only process one substrate and did not display activity for other, similar substrates (Chandrasekharan et al., 1996; Reisinger et al., 2013). On the other hand, the reconstructed ancestors of terpene synthases processed two isomers of a cyclic sesquiterpene (Guzzetti et al., 2016). Furthermore, reconstructed Precambrian ancestors of β-lactamases displayed catalytic versatility and hydrolyzed several β-lactam antibiotics with catalytic efficiencies comparable to those of modern enzymes (Risso et al., 2013; Zou et al., 2014). However, also many modern β-lactamases process several antibiotic substrates (Delmas et al., 2008; Tomatis et al., 2008). The far more interesting question whether ancient enzymes were really bi- or multi-functional and could processes different substrates with equal catalytic efficiencies – as it would have been required to sustain any form of simple primordial metabolism – has not been answered to date; none of the reconstructed proteins so far displayed true bi- or multi-functionality.

2.2 Summary and Discussion

2.2.1 Ancient HisA precursors are bi-functional, PriA-like enzymes

Since its first description in 2003, PriA has often been described as a "molecular fossil", or in other words a remnant of ancient, bi- or multi-functional enzymes (Barona-Gómez and Hodgson, 2003). If this holds true, modern HisA and TrpF enzymes would have evolved from an ancient PriA-like ancestor. However, as this assumption could never be demonstrated experimentally, it has also been postulated that PriA is rather a modern development of Actinobacteria and that it has emerged via several intermediates from mono-functional, specialist HisA enzymes (Noda-García et al., 2015). The details of these evolutionary scenarios will be discussed later on.

In order to directly investigate the enzymatic properties of ancient HisA and PriA enzymes, we applied ASR to computationally reconstruct three precursor enzymes. To this end we used a set of 103 sequences of extant HisA and PriA enzymes from Firmicutes, Spirochaetes, Bacteroidetes, Proteobacteria, and Actinobacteria. We constructed a reliable phylogenetic tree and inferred the sequences of the last common ancestors of Actinobacteria (CA-Act-HisA), of Proteobacteria (CA-Prot-HisA), and of all bacteria (CA-Bact-HisA). A representation of the tree is shown in **Figure 7**. It should be stressed that for the reconstruction of the more ancient precursors (CA-Prot-HisA and CA-Bact-HisA) the actinobacterial sequences were removed from the tree to exclude any effect of these sequences – especially their active site motifs – on the reconstruction process.

Notably, all three precursors catalyzed the isomerization of ProFAR (HisA reaction), as well as of PRA (TrpF reaction) with catalytic efficiencies in the order of $10^2 - 10^5\,\mathrm{M^{-1}\,s^{-1}}$ and $10^2 - 10^3\,\mathrm{M^{-1}\,s^{-1}}$, respectively (**Figure 7**). For comparison, the modern PriA enzyme from *S. coelicolor* (scPriA) displays catalytic efficiencies for the HisA and TrpF reaction of $3.2 \times 10^4\,\mathrm{M^{-1}\,s^{-1}}$ and $3.0 \times 10^6\,\mathrm{M^{-1}\,s^{-1}}$, respectively. Thus, these three precursors are, to our knowledge, the first examples of ancestral metabolic enzymes that were shown to be bi-functional. It is appropriate to use the term bi-functional in this context, because we were able to show that all three ancestors could support both histidine and tryptophan biosynthesis in the context of a *his*A and *trp*F deficient *Escherichia coli* strain. Their TrpF activity is thus clearly distinct from mere promiscuity, which would not play a role for the physiology of the *E. coli* host cells (**Table 1**).

A comparison of the sequences of the three precursors with those of extant HisA and PriA enzymes now allows one to determine which residues of PriA are likely to be crucial for its bi-functionality (**Figure 8A**). Several sequence motifs and structural elements have been

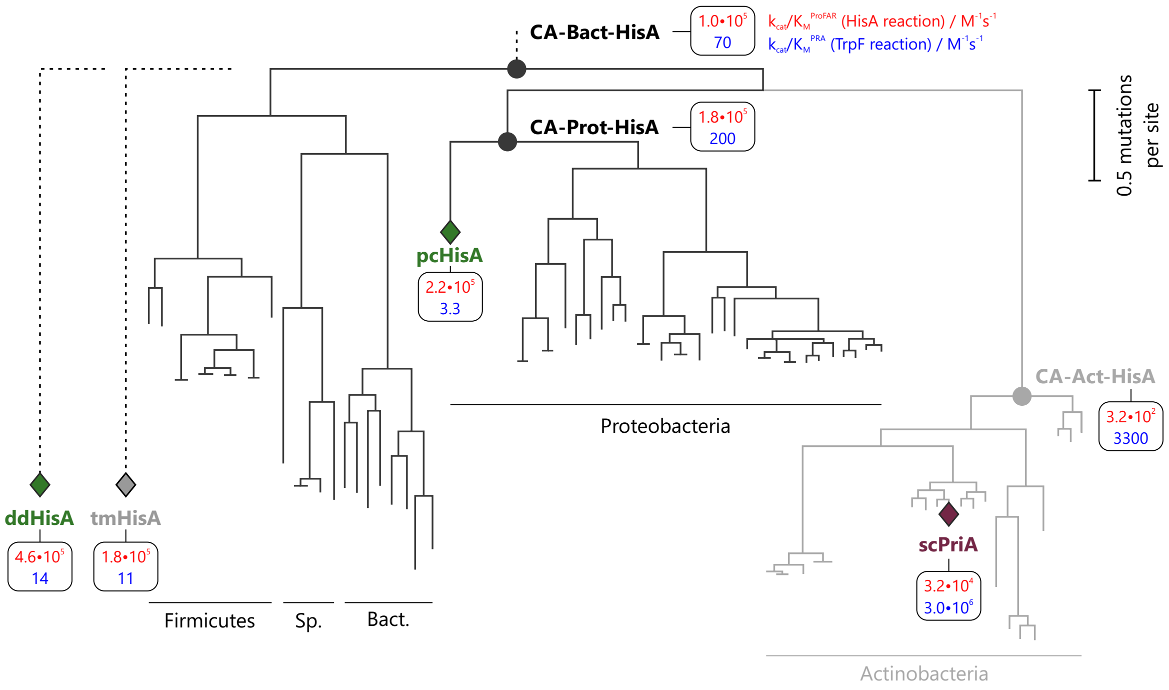


Figure 7. Evolutionary relationship between HisA and PriA enzymes.
The shown phylogenetic tree was constructed from HisA sequences from Firmicutes, Spirochaetes (Sp.), Bacteroidetes (Bact.), and Proteobacteria, as well as PriA sequences from Actinobacteria. The vertical bar corresponds to 0.5 mutations per site. CA-Bact-HisA, CA-Prot-HisA, and CA-Act-HisA refer to the common ancestors of HisA and PriA enzymes from all bacteria, all Proteobacteria, and all Actinobacteria, respectively. For the reconstruction of CA-Bact-HisA and CA-Prot-HisA a pruned tree with removed Actinobacteria sequences was used. Modern HisA enzymes that were characterized are marked by diamonds. For each enzyme its catalytic efficiency for the HisA (red) and TrpF reaction (blue) are given. The figure is taken from publication A.

suggested in the past to play important roles in this context; the N-terminal phosphate binding site, which interacts with the second ribose phosphate moiety of ProFAR, the residues that participate in substrate binding, as well as the residues of $\beta\alpha$-loop five that play a role in catalysis of both the HisA and TrpF reactions (List et al., 2011).

The N-terminal phosphate binding site comprises the motif SGG in PriA and GGG in HisA enzymes (Noda-García et al., 2015). Its importance for PriA function has been inferred from the role of the serine in PriA enzymes for the binding of PRA (Due et al., 2011) and the complete loss of TrpF activity by substitution of the serine with the highly similar amino acid threonine (Noda-García et al., 2010). However, CA-Prot-HisA and CA-Bact-HisA contain the GGG motif but are still able to catalyze the isomerization of

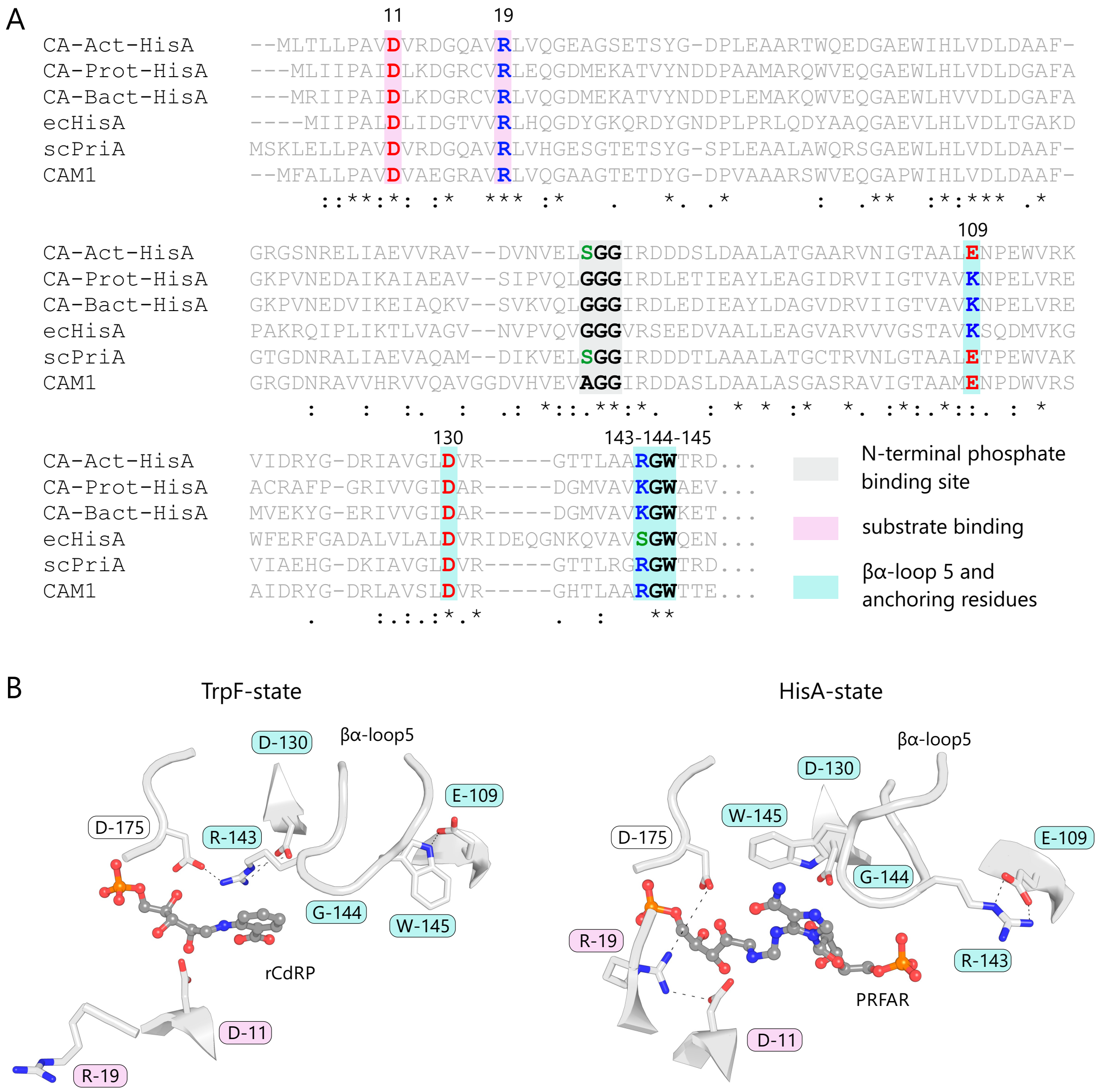

A
11
19
CA-Act-HisA --MLTLLPAVDVRDGQAVRLVQGEAGSETSYG-DPLEAARTWQEDGAEWIHLVDLDAAF-
CA-Prot-HisA ---MLIIPAIDLKDGRCVRLEQGDMEKATVYNDDPAAMARQWVEQGAEWLHLVDLDGAFA
CA-Bact-HisA ---MRIIPAIDLKDGRCVRLVQGDMEKATVYNDDPLEMAKQWVEQGAEWLHVVDLDGAFA
ecHisA ----MIIPALDLIDGTVVRLHQGDYGKQRDYGNDPLPRLQDYAAQGAEVLHLVDLTGAKD
scPriA MSKLELLPAVDVRDGQAVRLVHGESGTETSYG-SPLEAALAWQRSGAEWLHLVDLDAAF-
CAM1 --MFALLPAVDVAEGRAVRLVQGAAGTETDYG-DPVAAARSWVEQGAPWIHLVDLDAAF-
109
CA-Act-HisA GRGSNRELIAEVVRAV--DVNVELSGGIRDDDSLDAALATGAARVNIGTAALENPEWVRK
CA-Prot-HisA GKPVNEDAIKAIAEAV--SIPVQLGGGIRDLETIEAYLEAGIDRVIIGTVAVKNPELVRE
CA-Bact-HisA GKPVNEDVIKEIAQKV--SVKVQLGGGIRDLEDIEAYLDAGVDRVIIGTVAVKNPELVRE
ecHisA PAKRQIPLIKTLVAGV--NVPVQVGGGVRSEEDVAALLEAGVARVVVGSTAVKSQDMVKG
scPriA GTGDNRALIAEVAQAM--DIKVELSGGIRDDDTLAAALATGCTRVNLGTAALETPEWVAK
CAM1 GRGDNRAVVHRVVQAVGGDVHVEVAGGIRDDASLDAALASGASRAVIGTAAMENPDWVRS
130
143-144-145
CA-Act-HisA VIDRYG-DRIAVGLDVR-----GTTLAARGWTRD...
CA-Prot-HisA ACRAFP-GRIVVGIDAR-----DGMVAVKGWAEV...
CA-Bact-HisA MVEKYG-ERIVVGIDAR-----DGMVAVKGWKET...
ecHisA WFERFGADALVLALDVRIDEQGNKQVAVSGWQEN...
scPriA VIAEHG-DKIAVGLDVR-----GTTLRGRGWTRD...
CAM1 AIDRYG-DRLAVSLDVR-----GHTLAARGWTTE...
N-terminal phosphate binding site
substrate binding
βα-loop 5 and anchoring residues
B
TrpF-state
D-130
βα-loop5
E-109
D-175
R-143
G-144
W-145
rCdRP
D-11
R-19
HisA-state
βα-loop5
D-130
W-145
D-175
G-144
E-109
R-19
R-143
PRFAR
D-11

Figure 8. MSA of HisA and PriA enzymes and the two states of the PriA active site. (**A**) The MSA contains the sequences of the reconstructed HisA ancestors CA-Act-HisA, CA-Prot-HisA, and CA-Bact-HisA, as well as those of HisA from *E. coli* (ecHisA), PriA from *S. coelicolor* (scPriA), and CAM1, which is a supposed intermediate between HisA and PriA and was identified from a metagenome sequence set in Noda-García et al. (2015). Residues that are part of the N-terminal phosphate binding site, that are involved in substrate binding, and that are part of the catalytically important $\beta\alpha$-loop 5 are highlighted in gray, red, and green, respectively. The residues are colored according to their chemical properties (red, acidic; blue, basic; green, polar; black, small or hydrophobic). Important residues are numbered as in panel B. (**B**) Schematic view of the active site of PriA from *M. tuberculosis* in its TrpF- and HisA-state. Residue coloring is as in panel A. rCdRP and PRFAR are a product analogue and the product of the TrpF and HisA reactions, respectively. The figure is modified from publication A.

PRA both *in vitro* (**Figure 7**) and *in vivo* (**Table 1**). The loss of TrpF activity upon Ser→Thr mutation in HisA must thus result from other, more intricate effects on the active site.

Instead, we propose the $\beta\alpha$-loop five as the element that determines HisA-like mono-functionality or PriA-like bi-functionality, respectively. In ecHisA this loop comprises the amino acid motif SGW. In fact, only the sub-motif GW is conserved in HisA enzymes, as there is no observable preference for any amino acid on the first position of the motif (Publication A). Now on the other hand, in Actinobacteria this motif is strictly conserved as RGW and all three HisA precursors feature RGW or KGW. The presence of either arginine or lysine in $\beta\alpha$-loop five is essential for bi-functionality, as this loop exhibits significant ligand-induced conformational dynamics, triggered by the presence of either the HisA substrate ProFAR or the TrpF substrate PRA (List et al., 2011; Wright et al., 2008): In the TrpF state, the arginine (Arg143 in **Figure 8B**) is positioned by an aspartate (Asp130) to recruit the catalytic aspartate (Asp175). In the HisA state on the other hand the arginine is drawn away from the substrate by interacting with a glutamate (Glu109) and consequently the catalytic aspartate (Asp175) can be recruited by another arginine (Arg19) in a conformation required for the HisA reaction. A similar differential recruiting of the catalytic aspartate is not possible without a large basic residue in $\beta\alpha$-loop five.

It is known that only few amino acid substitutions are necessary to establish PRA isomerization activity on the HisA scaffold. For instance, the substitution of two aspartate residues in the vicinity of the active site by hydrophobic valines – in combination with two other mutations – relieved electrostatic repulsion of the PRA substrate and thus allowed for PRA isomerization (Claren et al., 2009). Along these lines it would be interesting to see if the single Ser143→Arg substitution would suffice to establish TrpF activity on the ecHisA scaffold. Moreover, the crystal structures of the HisA precursors would be informative whether $\beta\alpha$-loop five can adopt different conformations as in extant PriA enzymes and if the reaction mechanism of the precursors is therefore similar.

Another intriguing issue to be addressed in future work would be the shifts in the degree of bi-functionality from the most ancient CA-Bact-HisA precursor to the modern PriA enzymes. Our data show that the catalytic efficiency for isomerization of ProFAR decreased by about one order of magnitude, whereas that for the isomerization of PRA increased by about three orders of magnitude. Following the strategy successfully applied for identifying binding hot-spots in the imidazole glycerole phosphate synthase complex (Holinski et al., 2017) the characterization of evolutionary intermediates (representing internal nodes in the tree leading from its root to the modern PriA enzymes) might be informative on which residues contribute to the observed changes in catalytic efficiencies and thus modulate bi-functionality in modern PriA enzymes.

Table 1. *In vivo* complementation of auxotrophic *E. coli* strains by HisA-ancestors as well as modern HisA and PriA enzymes. The complementation time tells when colonies of *E. coli* strains lacking either *his*A, *trp*F, or both, after transformation with plasmids carrying the genes coding for the indicated enzymes, appeared on selective agar plates.

	complementation time in hours		
	Δ*his*A strain	Δ*trp*F strain	Δ*his*AΔ*trp*F strain
CA-Act-HisA	48	23	47
CA-Prot-HisA	16	33	28
CA-Bact-HisA	16	45	39
scPriA	22	22	23
ddHisA	16	153	181
pcHisA	15	70	63
tmHisA	16	114	144

2.2.2 Ancient bi-functionality persists in modern HisA enzymes

Enzymes, especially those that are part of highly coordinated and regulated metabolic pathways, tend to be specialized for their single task in this pathway and are not known to perform side reactions with other substrates, as this might impair differential regulation of the pathways. It was therefore quite exceptional that we found the RGW motif, which we have linked to PriA-typical bi-functionality, also in about 5% of modern HisA enzymes from organisms that have dedicated histidine and tryptophan biosynthetic pathways and contain a *trp*F gene. We selected and characterized three RGW-HisA enzymes from the Proteobacteria *Desulfovibrio desulfuricans* (ddHisA) and *Pelobacter carbinolicus* (pcHisA), as well as from *T. maritima* (tmHisA). All three did not only catalyze their native reaction but could also isomerize PRA, albeit with about 10^5-fold lower catalytic efficiency than modern TrpF enzymes (**Figure 7**). Nevertheless, these HisA enzymes are the first ones that have been shown to possess also TrpF activity.

Although we could show that all three tested HisA enzymes were able to rescue the growth of *his*A and *trp*F deficient *E. coli* strains (**Table 1**), it is unlikely that their TrpF activity is physiologically relevant, because the K_M^{PRA} values are 10-170-fold higher than the K_M^{ProFAR} values and the catalytic efficiencies k_{cat}/K_M^{PRA} are no higher than $14\,\mathrm{M}^{-1}\,\mathrm{s}^{-1}$ (**Figure 7**). Presumably, these very low efficiencies make the TrpF side-activities tolerable and prevent competition with the HisA reaction.

Our results are not conclusive on whether all modern HisA enzymes possess a TrpF side-activity. We also tested several HisA enzymes that do not contain the PriA-typical

RGW motif. And indeed, these enzymes were not able to complement *his*A and *trp*F deficient *E. coli* strains, although in some experiments after long incubation times very small colonies could be observed (Publication A). It is likely that the residual TrpF activity of inherent *E. coli* enzymes like PurF is responsible cell growth (Patrick and Matsumura, 2008). Nonetheless, these observations further support our view of the RGW motif as the decisive element that determines PriA-typical bi-functionality. Along these lines the bi-functional PriA enzymes from a sub-clade of the genus Corynebacterium became mono-functional HisA enzymes after the horizontal acquisition of a whole-pathway tryptophan operon (including a dedicated *trp*F gene) from a member of the γ-Proteobacteria and a change of the RGW motif to NGW (Noda-García et al., 2013). It is therefore plausible to assume that the mutational barrier between bi-functional and mono-functional HisA variants did not pose a real hurdle during evolution and that mono- or bi-functionality are easily accessible under corresponding selective pressure.

To put this rather complicated matter in a nutshell: Where did bi-functional PriA enzymes come from and why do Actinobacteria rely on them? The latter issue can be answered relatively straightforward. Actinobacteria like Streptomycetes have lost their *trp*F gene and thus simply have to rely on PriA for histidine and tryptophan biosynthesis. Also, they do not face any drawbacks from this strategy, because their amino acid biosynthesis is not regulated by feedback inhibition and the expression of biosynthetic enzymes is not coordinated in a pathway-specific manner (Hodgson, 2000). Thus, enzymes with multiple functions in multiple pathways are more likely to occur in Actinobacteria than in organisms with strict regulation of biosynthetic pathways.

Much, however, has been speculated about the evolutionary origins of PriA. It has been suggested that modern PriA enzymes are a remnant of ancient, bi-functional enzymes that supported histidine and tryptophan metabolism and that diverged into modern HisA and TrpF enzymes (Barona-Gómez and Hodgson, 2003). Only recently, another origin story was proposed from the analysis of marine meta-genome samples (Noda-García et al., 2015). In these samples, sequences were identified that bridged the gap between PriA-like and HisA-like sequences with respect to the sequence motif of the N-terminal phosphate binding site. It was proposed that the transition from the HisA motif SGG to the PriA motif GGG proceeded via intermediary enzymes possessing an AGG motif that display PriA-like bi-functionality (compare the transition enzyme CAM1 in **Figure 8A**). Thus, PriA might have developed from mono-functional HisA enzymes after the formation of Actinobacteria. However, with our data at hand, this scenario is rather unlikely considering the following two arguments: First, the study by Noda-García et al. (2015) only considered marine metagenome samples and Actinobacteria are not known to be prevalent in marine environments, which argues against a relation of these transition enzymes and modern Actinobacteria. Second, our characterization of the HisA ancestors shows that the N-terminal phosphate binding site motif is not relevant for PriA-like bi-functionality. On the contrary, CAM1 already possesses the PriA-typical RGW motif in $\beta\alpha$-loop 5, arguing against a role for transition from HisA to PriA.

We propose a different scenario for the evolution of HisA, TrpF, and PriA: Our results show that HisA is an enzyme whose ancestral bi-functionality has pertained for over two billion years of evolution, most likely without immediate and obvious benefit and without any evolutionary pressure. Its bi-functionality may have played an important role in the Precambrian era to maintain early metabolism but became expendable upon the integration of dedicated *trp*F genes into genomes. In the following the TrpF side-activity of the HisA enzymes evolved in a way that it was not lost completely but just went down to a physiologically not harmful level. Although the remaining bi-functionality of HisA was neither actively selected for, nor against, it was still a factor of metabolic flexibility not to be neglected. Consequently, the loss of the *trp*F gene in Actinobacteria could be compensated by the presence of a bi-functional HisA enzyme.

Such innovations that originate non-adaptively are called exaptations (Gould and Vrba, 1982). In other words, these are traits whose function has shifted during evolution and that did not originally form for their eventual purpose. A vivid example for exaptation are feathers, which were adopted to assist flight only after they had evolved for temperature regulation of primordial reptiles (Gould and Vrba, 1982). In the same way, the bi-functionality of ancient HisA enzymes did no longer serve its original purpose after the emergence of TrpF (in analogy, keeping the reptiles warm) but was exploited when there was the necessity due to the loss of *trp*F in Actinobacteria (in analogy, using feathers for flying). Although exaptation has not frequently been observed, there are some other examples: The light-refracting crystallins in the lens of the human eye stem from glutathione *S*-transferase (Tomarev and Piatigorsky, 1996) and a single point mutation in human β-globin causes sickle-cell anaemia but has in recent evolutionary timescales also been selected for in African populations as it confers resistance against malaria (Friedman and Trager, 1981; Ingram, 1959).

3 Conversion of anthranilate synthase into isochorismate synthase: Implications for the evolution of chorismate-utilizing enzymes *(Synopsis of Publication B)*

3.1 Introduction

3.1.1 Chorismate-utilizing enzymes and the MST superfamily

The shikimate pathway is a major metabolic route in bacteria, fungi, and plants. It is estimated that up to 20-50% of the photosynthetically fixated carbon is channeled into it in plants (Corea et al., 2012; Weiss, 1986). As the shikimate pathway is absent in animals it has sparked considerable interest as a target for potential herbicides and antibiotics. For instance, one enzyme of the pathway is the target of glyphosate, the most widely used herbicide in the 20th century (Duke and Powles, 2008). The pathway comprises seven enzymatic steps that convert D-erythrose-4-phosphate and phosphoenolpyruvate to chorismate (CH), the common precursor of the three proteinogenic aromatic amino acids tryptophan, tyrosine, and phenylalanine, folate coenzymes, benzoid and napthoid quinones, iron chelators, antibiotics, and innumerable other, mostly aromatic secondary metabolites like alkaloids, flavonoids, or lignins (Haslam, 2014; Tohge et al., 2013). CH has thus often been referred to as a central metabolic node molecule (Dosselaere and Vanderleyden, 2001).

The enzymes that catalyze the initial steps in the biosynthesis of the numerous CH-derived metabolites, i.e. chorismate-utilizing enzymes (CUEs), can be grouped into four

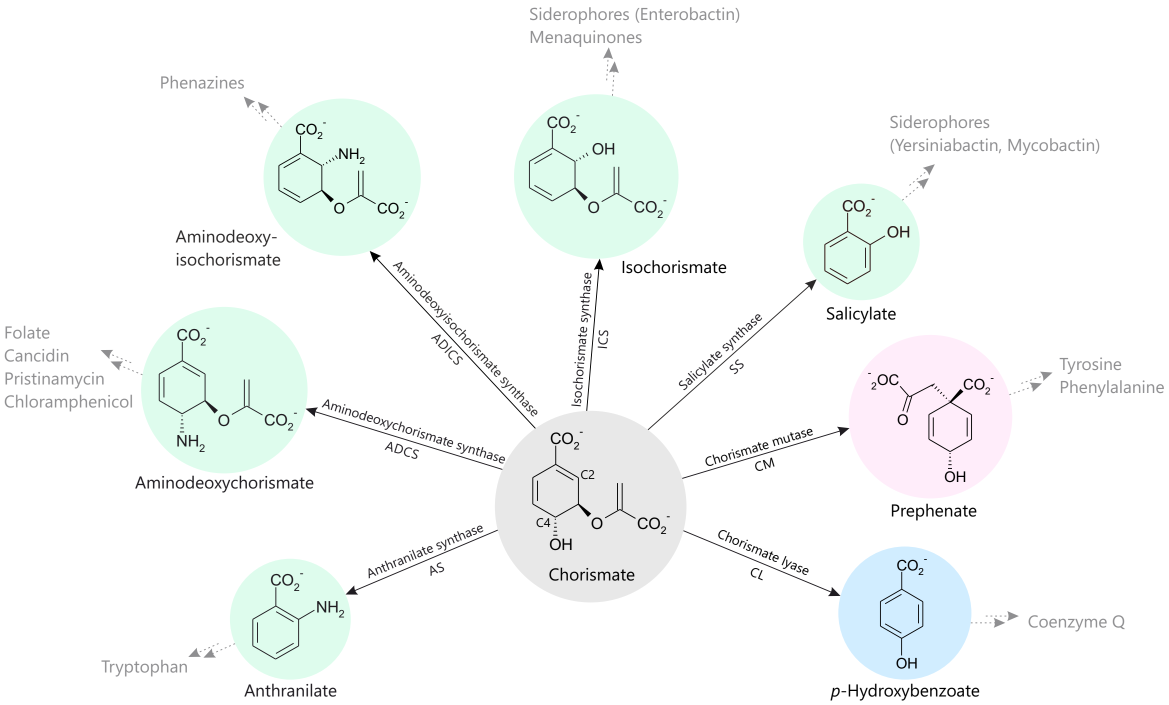


Figure 9. Reactions catalyzed by chorismate-utilizing enzymes.
The central metabolic node molecule chorismate is converted to several products that are the starting points for the biosynthesis of a wide range of primary and secondary metabolites in plants, fungi, and bacteria. The MST-superfamily comprises AS, ADCS, ADICS, ICS, and SS, which catalyze the formation of the metabolites highlighted in green. CMs belong to two different superfamilies and catalyze the formation of prephenate, highlighted in red. CLs convert chorismate to *para*-hydroxybenzoate, highlighted in blue.

different superfamilies (**Figure 9**). Anthranilate synthases (ASs), aminodeoxychorismate synthases (ADCSs), aminodeoxyisochorismate synthases (ADICSs), isochorismate synthases (ICSs), and salicylate synthases (SSs) belong to the SCOP superfamily 56322 (*ADC synthase*), which was dubbed MST-superfamily, referring to the CH derivatives menaquinone, siderophores, and tryptophan (Kolappan et al., 2007). Chorismate mutases (CMs) are known to have two different folds and are thus grouped into the SCOP superfamilies 48600 (*chorismate mutase II*) and 55298 (*YjgF-like*). Finally, chorismate lyases (CLs) belong to SCOP superfamily 64288 (*chorismate lyase-like*). In the following paragraphs the various CUEs are described with respect to structure and function.

The most heterogeneous group of CUEs are CMs. Based on their folds, they are grouped in the AroH class (trimeric, pseudo-($\alpha\beta$)-barrel fold), which is prevalent in Gram-positive bacteria (Chook et al., 1994; Gray et al., 1990) and the AroQ class (dimeric, all-helix bundle fold), which is prevalent in Gram-negative bacteria (Goerisch and Lingens, 1974). CMs catalyze the committed step in the biosynthesis of the aromatic amino acids tyrosine and phenlyalanine by converting CH in a 3,3-sigmatropic Claissen-rearrangement of its enolpyruvyl sidechain to prephenate (PA) (Dosselaere and Vanderleyden, 2001). CLs comprise a 6-stranded anti-parallel β-sheet core with few surrounding, short α-helices (Gallagher et al., 2001). They catalyze a 1,2-elimination of the enolpyruvyl moiety of CH to yield *para*-hydroxybenzoate (PHB), which is a key precursor of quinones like coenzyme Q (Walsh et al., 1990).

The remaining CUEs are paralogs and belong to the MST-superfamily. The close evolutionary relationship between them is evident from several facts. Pairwise alignments of consensus AS, ADCS, ADICS, ICS, and SS sequences reveal similarities between 25 and 50%. The C-terminal, catalytic halves of individual MST-enzymes share between 24 and 32% identical residues, with this value going up to 70% for the active site and surrounding residues. The structures of MST-enzymes superimpose with root-mean-square deviation (RMSD) values of $0.9 - 1.6\,\text{Å}$ over all Cα-atoms (**Figure 10**). Moreover, all MST-enzymes depend on Mg^{2+} as a cofactor and work according to the same general catalytic mechanism (He et al., 2004). In typical S_N2'' manner, their reactions involve a nucleophilic attack on C2 of CH and concomitant loss of water at C4. Nevertheless, the different MST-enzymes catalyze considerably different transformations of CH: Based on the nucleophile employed, ammonia-utilizing MST-enzymes (AMEs) – AS, ADICS, and ADCS – can be differentiated from water-utilizing MST-enzymes (WMEs) – ICS and SS. Alternatively, based on the nature of the product, the superfamily can be grouped into the enzymes that form aromatic products (AS and SS) and such that form non-aromatic products, which still contain the enolpyruvyl sidechain of CH (ADICS, ADCS, and ICS). The combination of the same fold, similar sequences, the same substrate, and the overall same catalytic mechanism, but subtle differences in reaction chemistries and formed products makes the MST-superfamily an ideal model to study the sequence-structure-function relationship and the evolutionary links between members of protein superfamilies.

The best characterized family of MST-enzymes are ASs, which belong to the group of glutamine amidotransferases (GATases). These enzyme complexes utilize glutamine to introduce amino-substituents in numerous anabolic pathways leading to nucleotides, amino acids, coenzymes, or antibiotics (Zalkin and Smith, 2009). AS catalyzes the committed step of the tryptophan biosynthetic pathway; the only metabolic pathway in nature that generates an aromatic indole ring from non-aromatic precursor molecules (**Figure 9**).

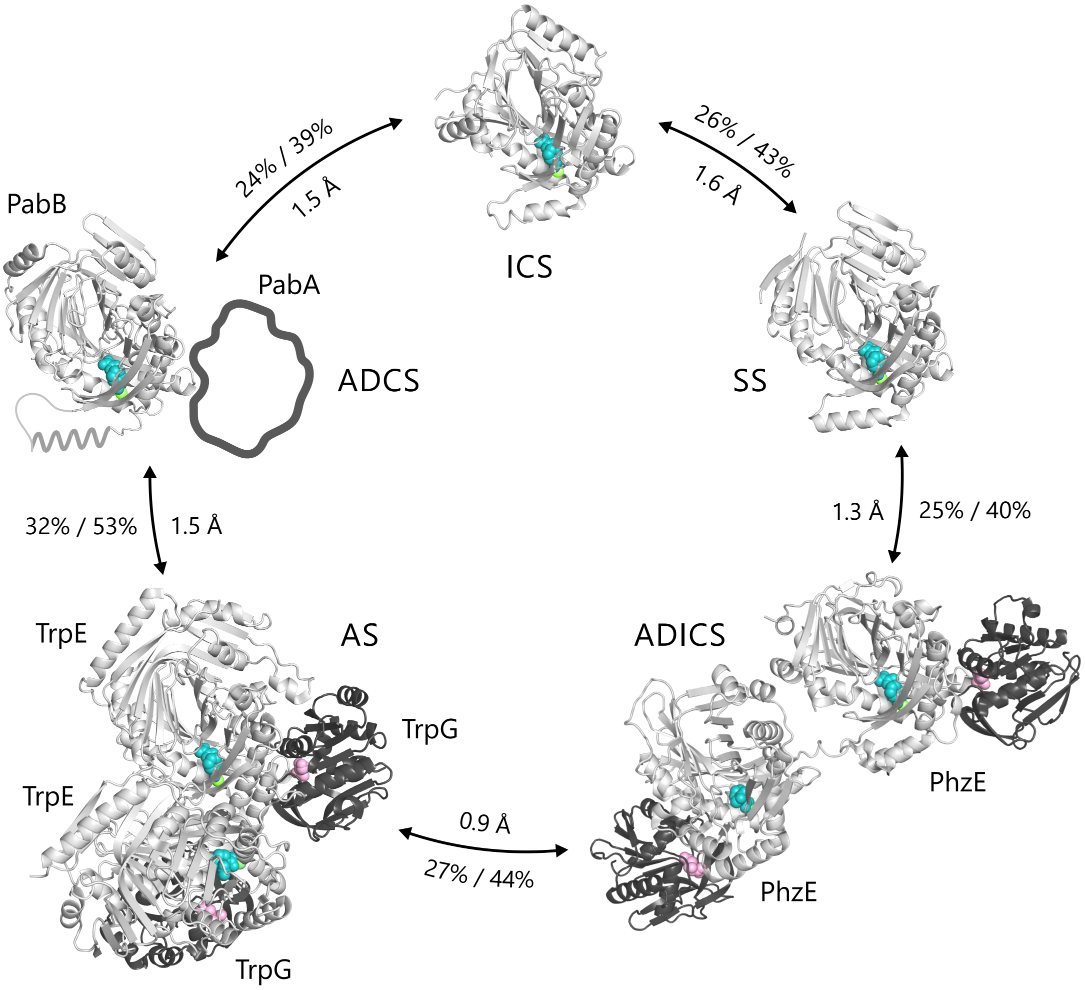


Figure 10. Structure and sequence relationship in the MST-superfamily.
The proteins and complex subunits that are members of the MST-superfamily are colored in light gray. Structurally not related glutaminase subunits are colored in dark gray. The active sites of the MST-enzymes are indicated by sphere-representations of CH (cyan) and a Mg^{2+} ion (green); those of glutaminase subunits by glutamine (magenta). Each MST-enzyme is arranged such that it is next to its two closest homologs from the superfamily. Sequence identities/similarities between the MST-enzymes are given as percentages, RMSD values over all Cα-atoms are given in Å. The values were determined for the following proteins: ICS from *E. coli* (PDB ID 3hwo), SS from *M. tuberculosis* (PDB ID 2fn1), ADICS from *Burkholderia lata* (PDB ID 3r76), AS from *Serratia marcescens* (PDB ID 1i7q), and ADCS from *E. coli* (PDB ID 1k0e; note that the glutaminase subunit and an α-helix are not resolved in the structure and are thus sketched).

So far, the crystal structures of three ASs have been solved and reveal heterotetrameric complexes with a_2b_2 stoichiometry (Knöchel et al., 1999; Morollo and Eck, 2001; Spraggon et al., 2001) (**Figure 10**). Only the larger subunit – commonly referred to as the *synthase* or TrpE – is a member of the MST-superfamily and shows the typical MST-fold. This fold comprises a complex α/β-sandwich with two anti-parallel β-sheets forming a core grove, surrounded by α-helices. The active site is located near a prominent, kinked α-helix, which coordinates the Mg^{2+} ion essential for CH binding and catalysis (Parsons et al., 2008; Sridharan et al., 2010). The smaller subunit – referred to as the *glutaminase* or TrpG – has a class I GATase fold, which is characterized by a catalytic triad comprising a cysteine, a histidine, and a glutamate residue (Zalkin and Smith, 2009). In the AS from *Salmonella typhimurium* and *S. marcescens* the two dimers associate via their TrpE subunits, whereas in the AS from *Sulfolobus solfataricus* they associate via their TrpG subunits.

The reaction catalyzed by AS comprises the amination of CH at C2 and the subsequent aromatization to yield anthranilate (AA). This step is highly regulated through various genetic repression and attenuation mechanisms (Dosselaere and Vanderleyden, 2001; Merino et al., 2008; Yanofsky et al., 1984) and feedback inhibition of AS by tryptophan (Caligiuri and Bauerle, 1991; Spraggon et al., 2001). The reaction itself proceeds in three steps: Initially, glutamine is hydrolyzed in the active site of TrpG and nascent ammonia is channeled to the active site of TrpE, where it reacts with CH in a reversible *syn*-1,5-substitution of the C4 hydroxyl group to an aminodeoxyisochorismate (ADIC) intermediate (Kozlowski et al., 1995; Walsh et al., 1990). The final *cis*-elimination of pyruvate from ADIC renders the total reaction irreversible and yields AA (Walsh et al., 1990).

ADCSs convert CH to aminodeoxychorismate (ADC) in the committed step of folate biosynthesis (Dosselaere and Vanderleyden, 2001). Folate is a precursor of tetrahydrofolate, an important one-carbon donor molecule that takes part in various biosynthetic pathways leading for example to purines, methionine, glycine, and pantothenic acid (Dosselaere and Vanderleyden, 2001). The biosynthesis of tetrahydrofolate is targeted by two of the most common drugs used against infectious diseases, trimethoprim and sulfamethoxazole (Masters et al., 2003). In Streptomycetes, ADC additionally serves as the precursor for several antibiotics, including cancidin, chloramphenicol, and pristinamycin (Blanc et al., 1997; Brown et al., 1996; Criado et al., 1993).

Similar to ASs, ADCSs are GATases and comprise a large synthase subunit, commonly named PabB, which shows the typical MST-fold (**Figure 10**), and a small glutaminase subunit, named PabA, which is structurally similar to the glutaminase subunit of AS. Although only the structure of PabB has been solved so far (Bera et al., 2012; Parsons et al., 2008), it is assumed that the quaternary structure of ADCSs resembles that of one AS dimer (Parsons et al., 2008). Interestingly, the PabB subunit contains the same tryptophan binding site as TrpE, despite no inhibition of ADCS by tryptophan has been observed. This, however, is another hint at the common evolutionary history of the MST-enzymes.

In contrast to AS, ADCS catalyzes the amination of CH under retention of the original regio- and stereochemistry in ADC (Walsh et al., 1987) (**Figure 9**). For this, ADCS employ a lysine side chain nucleophile to covalently link the ADIC-like reaction intermediate to the enzyme. This covalent complex is subsequently resolved in a second S_N2'' reaction with nascent ammonia as the nucleophile, which has been generated from the hydrolysis of glutamine at the PabA subunit (He et al., 2004). Interestingly, some ADCSs, mainly those from Firmicutes, do not rely on the enzyme-provided lysine nucleophile but utilize two molecules of free ammonia for the two S_N2'' reactions with release and re-uptake of the ADIC reaction intermediate (Schadt et al., 2009). The obvious difference between AS and ADCS catalyzed reactions – the fate of the enolpyruvyl moiety – has recently been traced back to two factors: In lyase-active MST-enzymes like AS the reaction intermediate is bound with three-fold higher affinity than in lyase-inactive ones like ADCS (Meneely et al., 2016). Furthermore, the lyase-active enzymes bind the reaction intermediate in an equatorial conformation, which favors elimination of pyruvate by presenting the C2-hydrogen to a lysine catalytic base (Culbertson et al., 2015).

ADICSs are homologs of AS and ADCS and are unique to Pseudomonas species. These enzymes, also called PhzE, occur as fusion proteins of TrpE and TrpG domains. The structure of PhzE from *Burkholderia lata* shows that two fused TrpE-TrpG dimers associate via their TrpE domains (Li et al., 2011) (**Figure 10**). Their quaternary structure thus differs from canonical AS quaternary structures. ADICSs aminate CH at C2 like ASs but do not eliminate pyruvate and thus form a non-aromatic product like ADCSs (**Figure 9**). This characteristic is exploited in the biosynthesis of ADIC-based, blue-green, redox-active pigments called phenazines (Blankenfeldt, 2013), which have recently been discovered as a source for novel antibiotics against multi-drug resistant *Staphylococcus aureus* strains (Borrero et al., 2014).

ICSs also show the typical MST-fold but are, unlike the previously discussed AMEs, monomers or homo-dimers and do not contain other, non-MST-type subunits (**Figure 10**). ICSs provide a fork from CH towards the metabolic pathways leading to dihydroxybenzoate-based siderophores and menaquinone, by converting CH with regio- and stereo-specificity to isochorismate (IC) (Liu et al., 1990; Walsh et al., 1990) (**Figure 9**). Siderophores are iron-chelator molecules that are produced by bacteria to assist in iron-uptake (Crosa, 1997; Zwahlen et al., 2007), because iron is mainly present as the highly insoluble ferric hydroxide form in the environment and is thus severely limited in its biological availability. Siderophores bind ferric iron with extremely high affinities of up to 10^{52} M^{-1} (Harris et al., 1979) and thus allow for the uptake of external iron (Neilands, 1995) or enable infectious pathogens to scavenge iron from iron-containing host-proteins (Domagalski et al., 2013). The efficiency of iron-uptake through siderophores is directly related to virulence in many pathogenic bacteria (Ratledge, 2004; Ratledge and Dover, 2000). Consequently, siderophore biosynthetic enzymes and siderophore carrier proteins have been suggested as targets for novel antimicrobials (Lamb, 2015). IC is also the precursor for menaquinone and the derived compounds menaquinol, phylloquinone, and phylloquinol. Under anaerobic conditions these quinones are required for electron transport and ATP synthesis (Dosselaere and Vanderleyden, 2001) and some bacteria like the pathogen *M. tuberculosis* solely rely on menaquinone for oxidative phosphorylation (Dhiman et al., 2009). As vitamins of the K group, menaquinone and its derivatives play a important roles for blood clotting in humans and must be obtained from the intestinal flora or from dietary sources (DiNicolantonio et al., 2015; Olson, 1984). Due to these considerably different utilizations of IC most bacteria possess two distinct ICS isozymes (EntC-type and MenF-type), which are specific for siderophore and quinone biosynthesis, respectively (Dahm et al., 1998).

SSs are structurally and functionally closely related to ICSs. They are also monomers or homo-dimers and show the typical MST-fold (**Figure 10**). SSs provide a fork from CH towards the biosynthesis of salicylate-based siderophores like yersiniabactin and mycobatin (Dosselaere and Vanderleyden, 2001) (**Figure 9**). These compounds are essential for pathogenic bacteria like *M. tuberculosis*, *Yersinia pestis*, or several Pseudomonas species for survival under iron-deficient conditions and for establishment and maintenance of infection (Cornelis and Matthijs, 2007; De Voss et al., 2000).

The most intriguing difference between ICS and SS is the fate of the enolpyruvyl moiety of CH. SS eliminates pyruvate from the IC reaction intermediate to give the aromatic product salicylate (SA), with the consequence that the derived siderophores carry monohydroxyl substituted phenyl rings. ICS, on the other hand, does not eliminate pyruvate and releases the non-aromatic IC, which leads to di-hydroxyl substituted siderophores. This subtle difference is attributed to different conformations of IC in the respective active sites. In SS, it is bound such that the pi-orbitals of the enolpyruvyl methenyl-group overlap with the orbital of the hydrogen at C2, allowing for a pericyclic, sigmatropic elimination without the involvement of solvent protons (Lamb, 2011; Zwahlen et al., 2007). In ICS, however, IC is bound in a pseudo-axial conformation that precludes a sigmatropic elimination due to unfavorable dihedral angles of the enolpyruvyl moiety (Sridharan et al., 2010).

3.1.2 Primary and secondary metabolism

It has been mentioned in the preceding section that the MST-superfamily can be grouped into AMEs (AS, ADCS, ADICS) and WMEs (ICS, SS), depending on whether ammonia or water is utilized as a nucleophile in the first step of the S_N2'' reaction. Interestingly, this subdivision of the MST-superfamily is also reflected in the assignment of AMEs to primary metabolism and of WMEs to secondary metabolism. The two terms *primary* and *secondary* have entered general usage in the scientific community to differentiate between the branches of metabolism that produce essential metabolites and such that produce metabolites that are less relevant for an organism, e.g. that are needed only under certain conditions. The terms date back to the first coarse-grained classification of plant metabolism by physiological chemist Albrecht Kössel in the 19th century (Kössel, 1891).

Since then primary metabolism is understood as the collection of enzymatic reactions that produce metabolites, i.e. chemical compounds, that are directly involved in growth, development and reproduction of an organism; recall the essential amino acid tryptophan, which is produced in the metabolic pathway starting from the AME AS. Typically, carbohydrate, lipid, amino acid, and nucleotide metabolism are core parts of primary metabolism. Usually, primary metabolic pathways are target-oriented and only make a single product and no side products (Fischbach and Clardy, 2007). In other words, they are stream-lined in order to maximize the yield of the essential metabolites they produce and to minimize the waste of metabolic resources. Also, the enzymes that build up such pathways are highly optimized and coordinated; for instance, only seven enzymatic activities are necessary to generate the complex aromatic amino acid tryptophan from its precursor CH (Yanofsky and Crawford, 1987).

Using the principle of exclusion, secondary metabolism is defined to comprise all enzymatic reactions that yield products that are not directly involved in growth, development, and reproduction, but rather play different roles in the ecology and physiology of organisms. Such secondary metabolites could be competitive agents (antibiotics, iron chelators), agents of symbiosis, sexual hormones, signaling molecules among microorganisms (Demain and Fang, 2000; Straight et al., 2007), or mediators of interaction between microbial and multi-cellular organisms (Engel et al., 2002; Strobel and Daisy, 2003). Generally speaking, secondary metabolites confer some sort of competitive advantage to their producers under certain environmental conditions (Stone and Williams, 1992). Secondary metabolites are also a rich source of antibiotics (Sosio et al., 2004), chemotherapeutics (Tang et al., 2000), immunosuppressants (Schwecke et al., 1995), or cholesterol-lowering agents (Hendrickson et al., 1999) and thus play an enormous role in medicine and economy. Modern genetic and microbiological methods even allow for the tailoring of custom, artificial secondary metabolic pathways (Leitão and Enguita, 2016).

In contrast to the few hundreds of primary metabolites, hundreds of thousands of secondary metabolites have been identified so far (Hartmann, 2007) that also have a much wider range of chemical structures and functions than primary metabolites (Vining, 1992). For instance, flavonoids can act as UV screens in plants and as signaling molecules or defense compounds in bacteria (Firn and Jones, 2009). Gibberilins, a class of over hundred structurally related terpenes, are plant hormones that regulate growth and developmental processes (Hedden et al., 2001) but are also synthesized by fungi to render plants more susceptible to fungal infection (Tudzynski, 2005) and are even produced by some bacteria, albeit with unknown function (Fischbach and Clardy, 2007). Secondary metabolic pathways are thus not target- but diversity-oriented and are usually organized as complex networks of enzymatic reactions. Some pathways contain more than 40 enzymes (Trefzer et al., 2002) that are encoded in large clusters on continuous DNA stretches of up to 100.000 base pairs (McAlpine et al., 2005), which are often referred to as biosynthetic gene clusters (BGCs) (Fischbach and Walsh, 2006; Medema et al., 2015).

Why did such a complexity evolve and what selective pressures shaped secondary metabolic pathways? Two theoretical models have been put forward that aim to answer this issue. The *target-based model* assumes that evolutionary pressure led to the formation of pathways that yield metabolites, which specifically act against competitors and thus confer selective advantages. One assumed mechanism, for example, is high-affinity binding to cellular targets of other organisms in the same environment (Williams et al., 1989). For instance rapamycin, produced by Streptomycetes, forms high-affinity complexes with cis/trans peptidylprolyl isomerases and thus inhibits their function as protein folding chaperones (Heitman et al., 1991; Koltin et al., 1991). It is one of the most potent anti-fungal agents known and confers Streptomycetes with growth advantages against competing organisms like Candida fungi (Sehgal, 2003).

The *diversity-based model* was put forward by Firn and Jones and assumes two key points: First, potent biological activity as displayed by rapamycin is rare. Second, selective pressure would favor organisms that could generate and retain chemical diversity at low genetic and metabolic costs (Firn and Jones, 2000, 2009). It suggests that the higher the number of secondary metabolites produced by an organism, the higher the chances are that some possess potent biological activity. This model is supported by the fact that although most secondary metabolic pathways form hundreds of different products, like the gibberilins mentioned above, only few of these metabolites are highly biologically active and bind to their targets with low nanomolar affinity. Moreover, the organization of

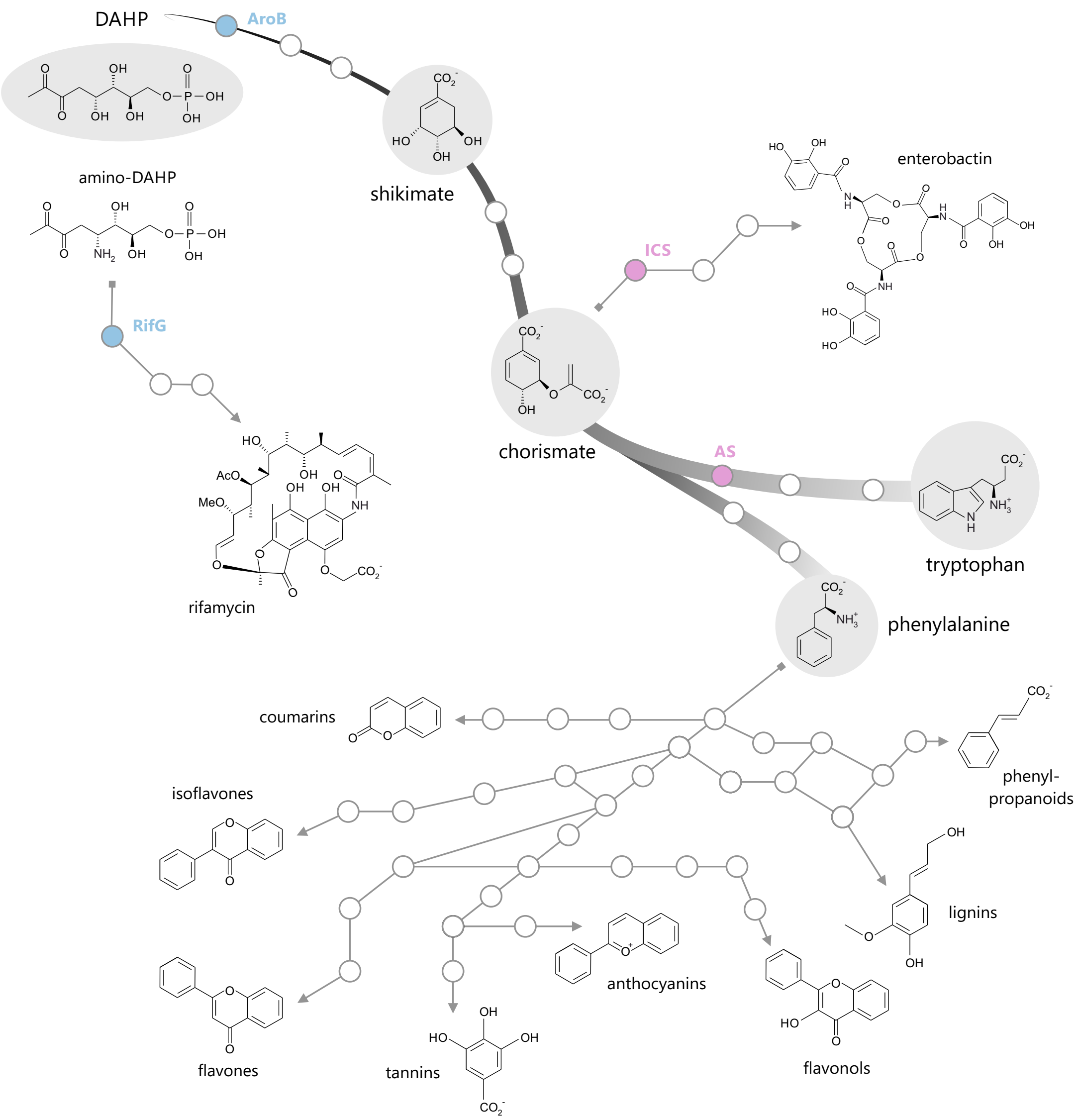


Figure 11. Links between primary and secondary metabolism.
The path in the center of the figure symbolizes the primary metabolic route leading from deoxyarabino-heptulonate-phosphate (DAHP) over shikimate and CH to phenylalanine and tryptophan. The secondary metabolic pathway leading to the antibiotic rifamycin (top left corner) involves the enzyme RifG, which acts on amino-DAHP. RifG is a homolog of AroB, which is part of the central primary pathway (blue circles). The secondary metabolic pathway leading to enterobactin (top right corner) starts with ICS, which acts on chorismate, similar to its homolog AS in the primary pathway leading to tryptophan (red circles). The primary metabolite phenylalanine is the precursor of a wide variety of phenylpropanoid secondary metabolites that are synthesized in a complex network of interdependent reactions (bottom part). It is assumed that some of the genes coding for the respective secondary metabolic enzymes have evolved from genes of the primary metabolic shikimate pathway; for details see Tohge et al. (2013).

secondary metabolic pathways in network- or matrix-style manner reduces genetic and metabolic costs while feeding metabolites to as many enzymes as possible (Fischbach and Clardy, 2007). A typical example of such a network is the phenylpropanoid pathway that starts from phenylalanine (**Figure 11**). A highly-interconnected network of enzymatic reactions converts phenylalanine into a range of structurally and functionally diverse metabolites including coumarins, isoflavones and flavones, tannin derivatives, anthocyanins, flavonols, lignin derivatives, and phenylpropanoids (Tohge et al., 2013).

Although these theories might explain the evolutionary forces that shaped secondary metabolic pathways, they do not provide an answer where the genes that code for secondary metabolic enzymes originally stem from. However, the fact that all secondary metabolites are ultimately derived from a primary metabolic precursor (Wink, 2011) already points in the direction of the answer. For example, in addition to the phenylpropanoids mentioned above, terpene building blocks originate from intermediates of glycolysis (Kirby and Keasling, 2009) and nitrogen-containing alkaloids are formed from purine nucleoside and amino acid precursors (Facchini, 2001). It has thus been acknowledged that also the enzymes of secondary metabolism must ultimately have their origin in primary metabolic pathways (Firn and Jones, 2000; Vining, 1992; Weng, 2014). And indeed, secondary metabolic pathways often comprise enzymes highly similar to enzymes from primary metabolism. For instance, the enzyme RifG catalyzes the cyclization of amino-deoxyarabino-heptulonate-phosphate (amino-DAHP) in the secondary metabolic pathway that leads to the antibiotic rifamycin (August et al., 1998; Rascher et al., 2003; Yu et al., 2001) (**Figure 11**). A homolog of RifG, AroB, catalyzes a nearly identical cyclization of DAHP in the primary metabolic shikimate pathway (Carpenter et al., 1998). It has consequently been speculated that RifG evolved from AroB through gene duplication and functional divergence (Fischbach et al., 2008). Duplicated genes coding for primary metabolic enzymes can be recruited for secondary metabolism via multiple paths, including mutational divergence, changes in transcriptional, translational, or allosteric regulation as well as changes in protein structure and protein-protein interactions (Moghe and Last, 2015). Another example for such recruitment events is the phenylpropanoid pathway. Many of its enzymes are assumed to have originated from the enzymes of the closely related shikimate and phenylalanine biosynthetic pathways (Tohge et al., 2013). To conclude, and to come back to the MST-superfamily, also AMEs and WMEs provide a link between primary and secondary metabolism. AS catalyzes the conversion of CH to IC in the primary metabolic pathway leading to tryptophan, while the homologous ICS catalyzes a very similar reaction in the secondary metabolic pathway leading to the siderophore enterobactin (**Figure 11**).

3.2 Summary and Discussion

3.2.1 Establishing isochorismate-synthase activity on an anthranilate-synthase scaffold

Based on the assumptions about the evolution of secondary metabolic enzymes outlined in the prior section, we hypothesized that a transition in nucleophile specificity from ammonia to water was the basis for the evolution of secondary metabolic WMEs from ancestral, primary metabolic AMEs. To retrace this putative evolutionary path we first identified the residues in AMEs and WMEs that are crucial for preference of ammonia or water as a nucleophile and subsequently established ICS activity on an AS scaffold.

An SSN of the MST-superfamily shows that AMEs and WMEs can be distinguished with high certainty at the sequence level (**Figure 12**). The ammonia-utilizing ASs and ADCSs are mostly grouped together in a large cluster, but are well separated from the water-utilizing SSs and ICSs. From the SSN we extracted the sequences of ASs, ADCSs, SSs, and ICSs and generated MSAs for each group, which in turn were used to compute a phylogenetic tree of AMEs and WMEs and finally sequence logos of the individual families (**Figure 13**). Also in the phylogenetic tree, primary metabolic AMEs and secondary metabolic WMEs are well separated with the posterior probabilities of important branches above 80%. To focus the search for the residues that determine nucleophile specificity, we made use of the fact that the AMEs AS and ADCS are heteromeric GATase complexes, in which the ammonia nucleophile is channeled from the active site of the glutaminase subunit to the active site of the synthase subunit (Raushel et al., 2003). We identified a 30 Å long channel that connects the active sites in the AS from *S. typhimurium* (stAS), similar to the channel observed in the crystal structure of a homologous ADICS (Li et al., 2011). Three residues principally shape the channel near the active site of the synthase where CH is bound: Gln263 in β-strand 11 as well as Met364 and Leu365 in α-helix 12 of stAS (**Figure 13**, inset). This trio of residues is conserved as QML in AS and QMI in ADCS. Although the WMEs ICS and SS are monomers and neither channel a nucleophile nor a reaction intermediate, it stands to reason, due to homology and similar active sites, that the hydroxide nucleophile reaches the CH electrophile via the same route as the ammonia does in AMEs. The three-residue motif is predominant as KLV in ICS and KIS in SS. The lysine residue in these motifs was shown to act as the catalytic base necessary for generation of the reactive hydroxide ion from water in the reactions of ICS and SS (Kolappan et al., 2007; Zwahlen et al., 2007). However, a simple substitution of Gln263 by lysine in AMEs is not sufficient to establish ICS or SS activity (Kerbarh et al., 2006).

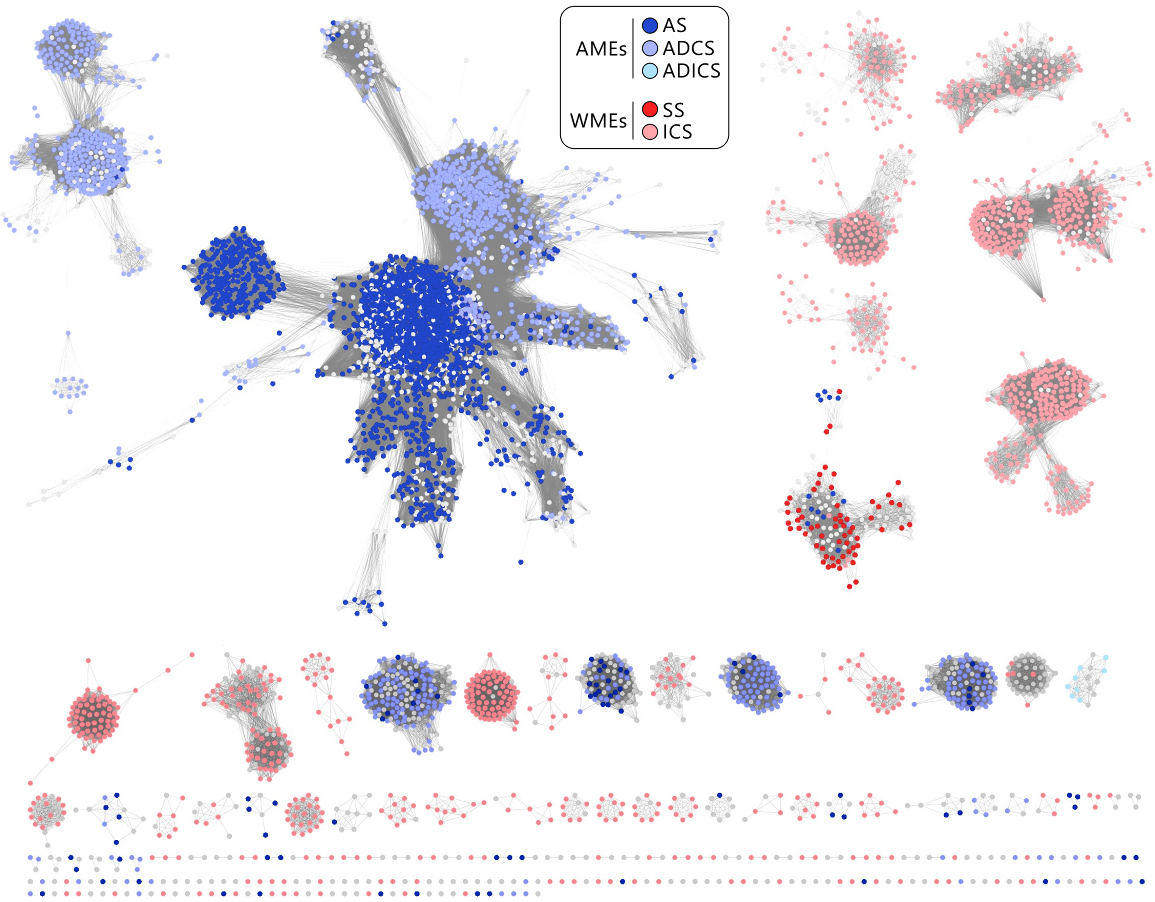


Figure 12. Sequence similarity network of the MST-superfamily.
The SSN of the MST-superfamily was generated from the InterPro data-set IPR019999 (anthranilate synthase component I-like) at an E-value threshold of 10^{-80}. Nodes corresponding to AMEs and WMEs, according to the InterPro annotation, are colored in shades of blue and red, respectively. Gray nodes refer either to ambiguously annotated sequences or to proteins with unknown function.

However, by combining the Gln263→Lys substitution with other substitutions at positions 364 and 365 in α-helix 12 of stAS, we were able to generate twelve enzyme variants that formed IC in amounts comparable to the native ICS EntC from *E. coli* (**Figure 14A**). On average, these stAS variants converted 20% of the supplied CH to IC within three hours, with the most active KIA variant, reaching about 37% IC, which falls in the range of EntC. The incomplete conversion is the result of an equilibrium between CH and IC, as has been described for EntC (Liu et al., 1990). Importantly, upon mutation of the supposed catalytic lysine of the triple-motif in the KLS variant, the resulting ALS variant lost all of its ICS activity, demonstrating that the introduced lysine acts as a catalytic base.

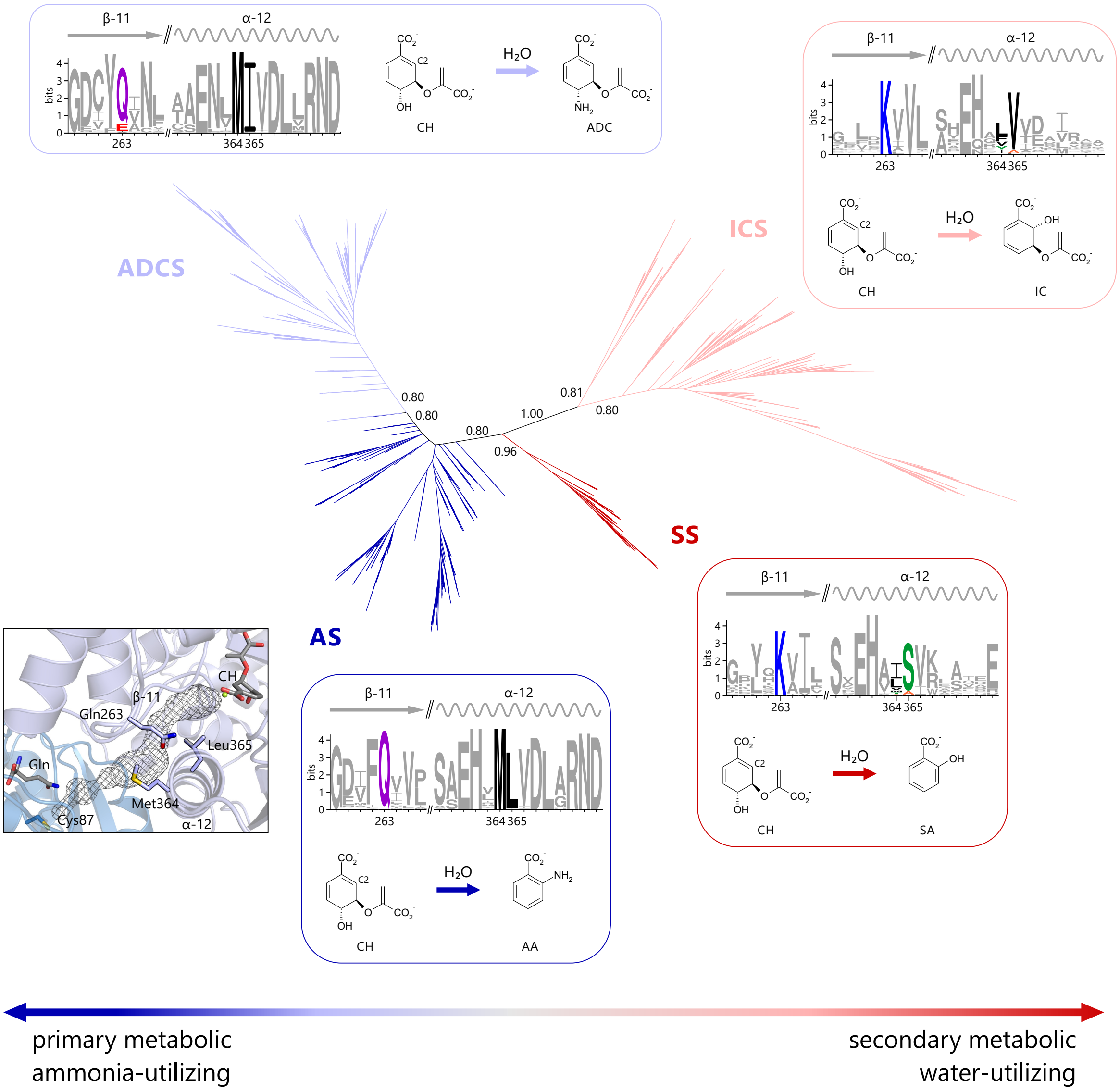


Figure 13. Sequence-structure-function relationship between ammonia- and water-utilizing MST-enzymes.
The primary metabolic AMEs AS and ADCS (blue) catalyze the formation of the aminated CH-derivatives AA and ADC. The small inset on the left shows the putative ammonia channel that connects the active sites of the AS glutaminase subunit (indicated by a Gln ligand and Cys87) and the AS synthase subunit (indicated by a CH ligand). The sequence logos show the residues of β-strand 11 and α-helix 12, which mostly shape the channel as it approaches the synthase active site. The secondary metabolic WMEs ICS and SS (red) catalyze the formation of the hydroxylated CH-derivatives IC and SA. The sequence logos show the same residues as for the AMEs. The figure is modified from publication B.

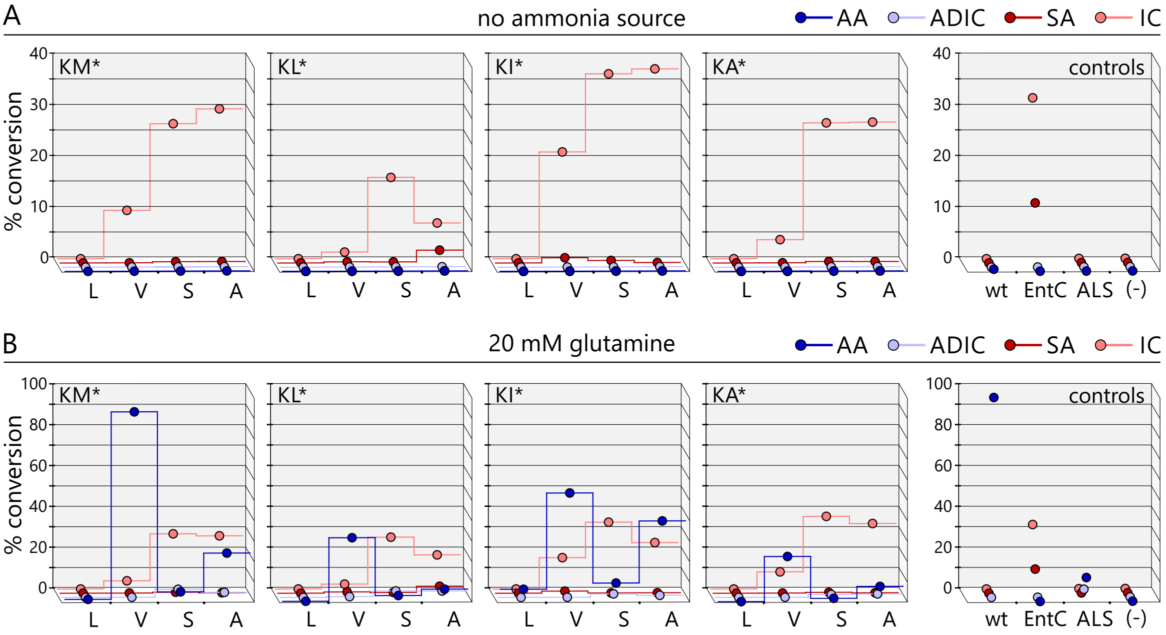


Figure 14. HPLC analysis of product distribution after reaction of chorismate with water or ammonia catalyzed by AS wild-type and variants.
For the four products AA, ADIC, SA, and IC their fraction out of all products formed is shown (% conversion). All analyzed variants share Lys263 and are grouped by residue 364 (e.g. KM* describes the four stAS variants with Lys263, Met364, and either Leu, Val, Ser, or Ala at position 365). All variants were assayed in the presence of the glutaminase stTrpG. "wt" refers to the product distribution of the native AS from *S. typhimurium*, "EntC" to that of the ICS from *E. coli*, and "(-)" to that of a control reaction in the absence of any enzyme. Each data point represents the mean of at least three independent experiments. Error bars are omitted for clarity, because the average and maximum absolute errors were only 0.3% and 3.7%, respectively. Step lines connecting data points were added to aid visual tracking of product-specific values. (**A**) Product distributions in the absence of any ammonia source. (**B**) Product distributions in the presence of glutamine as an ammonia source. The figure is modified from publication B.

Interestingly, apart from the catalytic lysine in β-strand 11, ICS activity only depends on the occupation of position 365. The nature of the residue at position 364 is irrelevant for IC formation; all tested amino acids (Met, Leu, Ile, and Ala) lead to active KM*, KL*, KI*, and KA* variants (**Figure 14A**). These results argue against a role of the residue at position 364 in positioning the catalytic lysine via hydrophobic interactions, as has been suggested previously (Ziebart and Toney, 2010).

Molecular dynamics simulations predicted that the occupation of position 365 with a relatively small amino acid is crucial for ICS activity: The presence of a residue with a larger sidechain than that of valine blocks the access of hydroxide ions to CH bound in the active site. In accordance with this prediction, all tested variants with a leucine at this position were inactive (**Figure 14A**). Moreover, virtually none of the ICS and SS sequences in our MST data-set contained a residue larger than valine at this crucial position. The chemical properties of the residue at position 365 seem to be less important; variants with Val, Ser, or Ala all displayed ICS activity (**Figure 14A**). Therefore, ICS activity is not dependent on a hydrogen-bond acceptor close to the catalytic lysine, as it has been suggested (Ziebart and Toney, 2010). Moreover, a hydrogen bond acceptor residue like serine is not conserved in SS and is not present in ICS.

3.2.2 Secondary-metabolic MST-enzymes might have evolved via bi-functional intermediates

Notably, all stAS variants that displayed ICS activity were still able to utilize ammonia and form AA, implying a broadened nucleophile specificity (**Figure 14B**). In other words, these variants are bi-functional and formed substantial amounts of AA and IC at the same time (see for example the stAS variant with the KIA motif, which forms about 40% AA and 20% IC). This demonstrates that a lysine catalytic base is required for the activation of water as a nucleophile but still allows for the access of ammonia to the active site.

Does this bi-functionality hint at a possible evolutionary path leading from AMEs to WMEs? It is generally acknowledged that secondary metabolic pathways have been assembled through the acquisition of genes coding for primary metabolic enzymes (Fischbach et al., 2008; Vining, 1992). The thought process behind this assumption involves the duplication of a gene coding for a primary metabolic enzyme and the shaping of its biological activity through mutations (neo-functionalization) (Firn and Jones, 2000; Fischbach et al., 2008; Ohno, 1970). Although the novel metabolite that is formed by the mutated enzyme may not immediately be beneficial for its producer organism, it might add to the overall chemical and metabolic potential by generating additional new metabolites through spontaneous transformations or through conversion by enzymes with broad substrate spectra (Vining, 1992). In the end, mutations that confer some kind of evolutionary advantage, are fixated in the genome. This could for example be a novel biomolecular activity, like high-affinity binding to cellular receptors of competing organisms, or a novel physiochemical property, like the diversification of cell wall and membrane components, making them less susceptible to enzymatic degradation (Firn and Jones, 2009). Later in evolution the duplicated genes are step by step integrated into BGCs that diverge through loss or gain of genes and through differentiating mutations in their enzymes (Fischbach et al., 2008; Traitcheva et al., 2007).

Most of such recruitment events have been described for enzymes from plant secondary metabolism (Moghe and Last, 2015); most likely because these events happened much more recently in evolutionary time-scales than comparable events in prokaryotic organisms and are thus easier to track. Just to give two examples, an isopropylmalate synthase from acylsugar biosynthesis has recently been traced back to its homolog from leucine biosynthesis (Ning et al., 2015) and a fatty acid-binding protein from primary metabolism was identified as the origin of a chalcone isomerase that is part of flavonoid biosynthesis (Ngaki et al., 2012). For bacteria, polyketid synthases and non-ribosomal peptide-synthethases have been directly and experimentally traced back to their primary metabolic homologs (Jenke-Kodama and Dittmann, 2009; Jenke-Kodama et al., 2008; Ziemert et al., 2014).

The notion of *duplication-divergence-integration* might, however, be a gross simplification, because it is not completely reasonable to assume that during evolution thousands and thousands of duplicate genes just emerged, mutated, and eventually led to hitherto unknown metabolites that also conferred a selective advantage to their producing organisms. Three key points argue against this assumption: First, as mentioned in the introduction to this chapter, decent biological activity with physiological relevance of any chemical compound is very rare (Firn and Jones, 2009). Second, gene duplication comes at a significant cost. Aside from the immediate physiological costs including the expenditure of energy for replicating, transcribing, and translating the extra genes, duplication may also lead to deleterious changes in gene dosages (Papp et al., 2003; Veitia, 2004). For example, an increased flux of important primary metabolites into a specific amino acid biosynthetic pathway, due to a duplication of the gene that codes for the enzyme of the committed step, would lead to a depletion of other amino acids, with the consequences at hand (Freeling and Thomas, 2006). Third, the rare mutations necessary for altering the phenotype of the gene towards novel metabolic diversity can only accumulate if the duplicated gene remains in the genome or population for a sufficient time (Bergthorsson et al., 2007). To ensure this, some kind of selective pressure would have to act on the redundant gene to be maintained, despite conferring no immediate benefit to its owner. This selective pressure would in turn restrict the ability of the gene copy to accumulate mutations; the cat evidently catches its tail here. Consequently, the overwhelming majority of duplicated genes are lost (Hughes, 1994; Lynch and Conery, 2000), especially duplicates of genes that code for primary metabolic enzymes (Chae et al., 2014).

Evolution should thus favor such organisms that can produce chemical diversity at low genetic and metabolic costs, which means, not necessarily through an obligatory gene duplication. This is possible by exploiting promiscuity and multi-functionality of enzymes. Richard Firn described this idea in the following example (Firn and Jones, 2009): Consider an organism, which contains an enzyme that produces an orange pigment. Now a mutation occurs in the gene coding for that enzyme, leading to either a relaxed substrate specificity so that the enzyme accepts a structurally related but chemically different substrate or to a slightly altered reaction chemistry. The result is a new pigment with similar physicochemical properties that still absorbs in the orange but now also in the yellow part of the visible spectrum. This mutant might thus exert no drawbacks on the organism (i.e. loss of absorption of orange light) and there will be no selective pressure to return the mutant enzyme to its original state. Moreover, there would also be no selective pressure due to increased genetic or metabolic costs as described above, because this scenario does not require a preceding gene duplication; the phenotype of the original gene is adequately retained in the new, mutated version.

Such multi-functional enzymes are often called metabolic generalists and are supposed to be an important part of metabolic evolution (Weng et al., 2012). An example of such sub-functionalization of a metabolic generatlist enzyme is the biosynthesis of pyrrolizidine alkaloids in plants (Moghe and Last, 2015): The enzyme deoxyhypusine synthase is essential in primary metabolism for covalently modifying a lysine residue of an eukaryotic translation initiation factor. It has also gained the ability to catalyze a similar reaction with putrescine as a substrate, which mimics the four-carbon lysine sidechain with its ϵ-amino group (Ober et al., 2003). This reactivity is important in the biosynthesis of the pyrrolizidine alkaloids. Modern plants however, contain two homologs of the deoxyhypusine synthase, whereby the secondary metabolic homolog has lost the ability to modify the translation initiation factor and only accepts putrescine as a substrate (Kaltenegger et al., 2013). Of course, such side-functions must not negatively impact the original function (Copley, 2015). Consequently, generalist enzymes are often inefficient compared to highly specialized primary metabolic enzymes, possibly to avoid a competition for metabolic resources (Bar-Even and Tawfik, 2013).

We have identified an example for the transition from a primary metabolic enzyme to a secondary metabolic one that fits Firn's generalist allegory remarkably well. Two mutations are sufficient to convert the primary metabolic AS into a generalist enzyme that accepts both ammonia and water as nucleophiles and thus forms AA and IC. An evolutionary intermediate similar to our bi-functional stAS variants could have generated novel metabolic diversity by opening the route to hydroxyl-substituted CH derivatives while at the same time maintaining the formation of AA for tryptophan biosynthesis. At this stage, duplication may have increased the efficiency of such a generalist through gene dosage effects (Näsvall et al., 2012) and sub-functionalization through gradual adaptation led to an increase in catalytic efficiency and specificity for the formation of IC. It can be speculated that some inherent property of a generalist AS/ICS intermediate may have influenced the retention of the duplicates in the genome, which would be an important factor for sub-functionalization. It has been suggested that promiscuity or multi-functionality and the possibility to explore a diverse chemical reaction space are such properties that favor retention of duplicate genes (Moghe and Last, 2015; Weng and Noel, 2012).

The question remains of course, how such metabolic novelty as in the case of IC was conserved and integrated into modern secondary metabolites. From IC only two additional enzymatic activities, exerted by EntB and EntA, are required to yield 2,3-dihydroxybenzoic acid (2,3-DHB), a molecule that itself can act as a siderophore, albeit with lower affinity than enterobactin (Lopez-Goñi et al., 1992). EntB and EntA are frequently found together in small genomic cluster with the ICS EntC (Crosa, 1989; May et al., 2001). We recently traced back the evolutionary history of EntB and EntA to their primary homologs from fatty acid and nucleotide metabolism (Publication D), which further adds to the notion that the enterobactin biosynthetic pathway was recruited from primary metabolic enzymes. It can thus be assumed that the enterobactin pathway evolved piece by piece: Initially a rudimentary pathway for the biosynthesis of 2,3-DHB was established from EntC, EntB, and EntA and later three iron-chelating 2,3-DHB molecules were linked together via a serine peptide scaffold through the recruitment of non-ribosomal peptide-synthetase-like enzymes into the pathway (Fischbach et al., 2008).

Although the acquisition of primary metabolic enzymes for secondary metabolism has long been known and numerous examples have been characterized, the true extent of these recruitment processes was not clear. We have recently identified primary metabolic homologs for 55% of the secondary metabolic enzymes listed in the MIBiG database (Publication D). It may thus be safe to say that about half of all known secondary metabolic enzymes can be traced back to their primary metabolic ancestors. It will be interesting to see whether experiments similar to those described in this chapter will enable one to transform primary metabolic homologs into their secondary metabolic relatives and thus to uncover more and more of the intricate evolutionary relationships between primary and secondary metabolism. One obvious challenge would be to reconstruct the enterobactin pathway from the identified primary homologs. The results presented in this chapter might also guide other metabolic engineering approaches involving siderophores, like the alteration of the biosynthetic capabilities of human pathogenic *Burkholderia* strains to produce different types of hydroxamate siderophores (Franke et al., 2014).

4 Evolutionary diversification of protein-protein interactions by interface add-ons *(Synopsis of Publication C)*

4.1 Introduction

4.1.1 Protein-protein interactions

Life in its full complexity would not be possible without protein-protein interactions (PPIs) or in other words the specific and mutual recognition of individual proteins and the non-covalent formation of larger complexes. In fact, interacting proteins are at the very basis of virtually every cellular function, including but not limited to replication, transcription, translation, membrane transport, signal transfer, and cell-cell communication (Keskin et al., 2016). With such critical roles to play in key cellular processes, it is not surprising that a number of severe diseases including cancer (Zinzalla and Thurston, 2009) and neurodegenerative diseases (Menche et al., 2015) are linked to dis-regulated PPIs (Meyer et al., 2014; Schuster-Böckler and Bateman, 2008).

In contrast to the old doctrine that phenotypic complexity directly relates to genetic complexity it is now clear that instead the extent and connectivity of interacting and inter-playing proteins determine the developmental stage of organisms. In this context, the sequencing of the first complete genomes has shown that not only the number of base pairs (around 3×10^9) but also the number of genes (around 6000 to 30 000) is comparable between lower and higher species (Blattner et al., 1997; Goffeau et al., 1996; Venter et al., 2001). However, for uni-cellular organisms like *Saccharomyces cerevisiae* the number of binary PPIs is projected to be around 18 000 to 75 000 (Yu et al., 2008; Zhang et al., 2012a), whereas for higher species like humans, this number is at least 10-fold higher, with

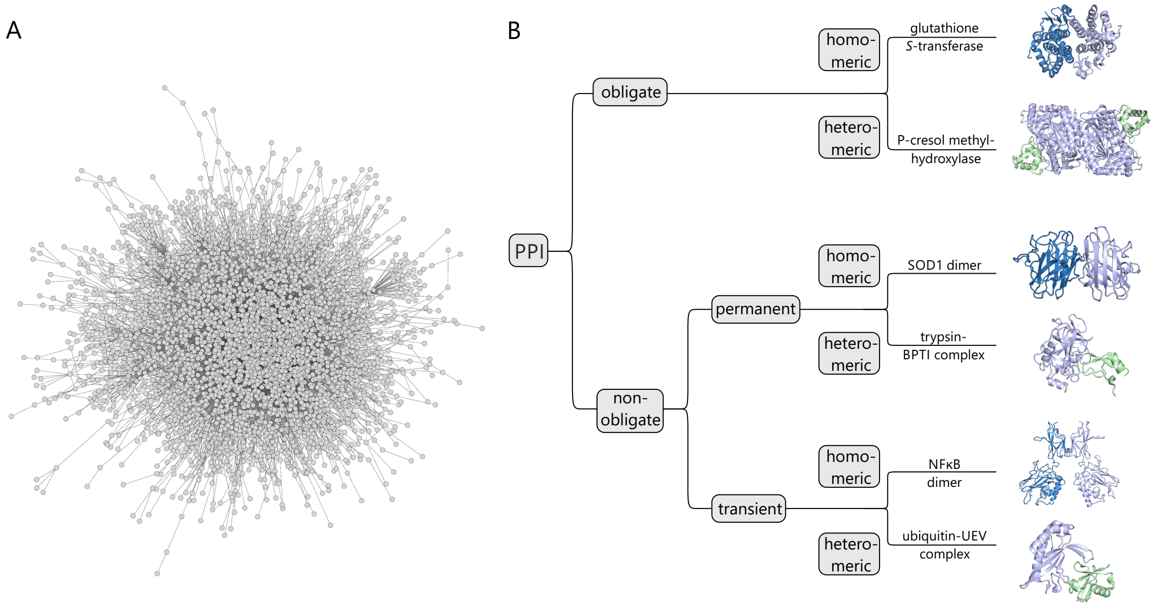


Figure 15. The human interactome and a classification of PPIs.
(**A**) Network representation of a part of the human interactome showing 14 000 binary PPIs derived from small-scale and systematic high-throughput experiments (Rolland et al., 2014). Each node represents a protein and connecting edges represent PPIs. (**B**) Classification of PPIs as proposed by Nooren and Thornton (2003a) and Keskin et al. (2016). PPIs can be organized based on the stability of the resulting complexes (obligate vs. non-obligate), based on the affinity between the subunits (transient vs. permanent), and based on the subunit composition of the complexes (homomeric vs. heteromeric). Identical subunits of the example complexes are colored in different shades of blue, structurally different subunits are colored green. The structures were taken from the PDB.

about 200 000 to 600 000 projected PPIs (Stumpf et al., 2008; Venkatesan et al., 2009). The resulting highly inter-connected networks of PPIs are often referred to as interactomes (**Figure 15A**). The complexity of interactomes is also evident from the observation that in *S. cerevisiae*, for example, over 70% of all proteins undergo at least one PPI (Gavin et al., 2006; Krogan et al., 2006). In other words, the interaction of proteins is much rather the norm that the exception and only few proteins act isolated. Our understanding of PPIs and interactomes is, however, still in its early stages. Only about 10% of all human PPIs have been experimentally validated and characterized so far (Hart et al., 2006). Moreover, structural information of protein complexes is often inaccessible. For example, of the 75 000 projected binary PPIs in *S. cerevisiae*, only 300, or about 0.5%, are structurally characterized (Zhang et al., 2012a).

The different types of PPIs are as diverse as the tasks they are involved in. Nooren and Thornton (2003a) proposed a classification scheme based on three important aspects of PPIs: stability, affinity, and composition of the resulting complexes (**Figure 15B**). Concerning stability, PPIs are divided into obligate and non-obligate interactions. Obligate complexes are exclusively present in oligomeric form and the corresponding monomers are not found in free or isolated form *in vivo*. Their association is often compulsory due to simultaneous gene expression and translation (Nooren and Thornton, 2003a) and their function usually critically depends on complex formation. Examples for obligate complexes are ribosomes (Yusupov et al., 2001), chaperones (Xu et al., 1997), glutathione *S*-transferase (Strange et al., 2001), or *P*-cresol methylhydroxylase (Kim et al., 1995).

In contrast, non-obligate complexes exist in an equilibrium between the oligomeric form and the free, monomeric subunits (Jones and Thornton, 1996; Levy and Teichmann, 2013). Non-obligate interactions can further be classified based on the affinity between the subunits into permanent and transient PPIs (**Figure 15B**). Permanent interactions are characterized by high affinities between complex subunits and corresponding dissociation constants are usually in the low nanomolar range or below (Jones and Thornton, 1996). A typical example are antibody-antigen complexes, which form with high affinity and usually dissociate slowly. Other examples are the human superoxide dismutase (SOD1) that forms highly stable dimers but is also active in its monomeric form (Banci et al., 1998) or complexes between proteases like trypsin and corresponding high-affinity inhibitors like the bovine pancreatic trypsin inhibitor, which binds to trypsin with a dissociation constant of 10^{-14} M (Vincent and Lazdunski, 1972). In contrast, transient complexes are characterized by a much weaker interaction between their subunits with dissociation constants generally in the micromolar range or higher (Nooren and Thornton, 2003b). Usually, key components of cellular signaling pathways like NFκB-target complexes or complexes between ubiquitin and corresponding ubiquitin-binding proteins are transient. Their dynamic, continuous association and dissociation allows the formation of many diverse effector complexes, depending on temporal and spatial conditions (Plowman and Hancock, 2005; Verma et al., 1995).

Finally, PPIs can be classified based on the composition of the resulting complexes. Binary or dimeric complexes comprise only two subunits, whereas higher-order or multimeric complexes like ribosomes, chaperons, or virus particles can comprise up to 1200 individual subunits (Klose and Rossmann, 2014). Homomeric complexes are formed through the association of identical subunits, whereas heteromeric complexes are formed through the association of non-identical subunits. It should be noted at this point that, no matter how well such classification schemes perform, they are just simplifications that make it easier to comprehend the continuum of PPIs that results in the most diverse protein complexes, reflecting specific physiological conditions, the presence of effector molecules and inhibitors, or cellular co-localization.

The variety of PPIs and protein complexes is the result of evolutionary processes that have shaped each interaction in a way to optimize their associated function (Janin et al., 2008; Marsh and Teichmann, 2015). For instance, obligate complexes have possibly evolved for reasons of stability or because of the need for a tight and permanent interaction that allows the formation of a shared active site (Dobson et al., 2004; Wente and Schachman, 1987). Just as well, transient PPIs were shaped that way for reasons of multi-specificity or for the ability to interact with numerous different partners in signal transduction pathways (Nooren and Thornton, 2003a).

The evolution of protein complexes most likely proceeded via step-by-step assembly of individual subunits to higher-order complexes. These processes have been retraced in the laboratory for a number of protein complexes (Levy et al., 2008; Marsh et al., 2013; Marsh and Teichmann, 2014). Moreover, it has recently been demonstrated that complex evolution most likely only involved three basic steps: subunit dimerization, cyclization, and addition (Ahnert et al., 2015). However, due to the complexity of PPIs and interaction networks, their evolutionary history still lies mostly in the dark. This is furthermore aggravated by the fact that only a minor fraction of protein complexes are accessible by traditional structural biology methods like crystallography, nuclear magnetic resonance spectroscopy, or electron microscopy. Also, small-scale but high-quality *in vitro* analyses of PPIs are often limited to indirect and labor-intensive methods like gel filtration chromatography or mass spectrometry that do not yield immediate three-dimensional structure information on protein complexes. Thus, mapping of whole interactome networks heavily relies on high-throughput but low-quality methods like two-hybrid screenings (Brückner et al., 2009; Fields and Sternglanz, 1994; Warbrick, 1997) and tandem affinity purifications coupled to mass spectrometry (Kaiser et al., 2008; Völkel et al., 2010). Two hybrid and related split systems are based on the genetic fusion of suspected interaction partners to DNA-binding and transcription-activation domains of transcription factors or to two domains of an enzyme in a way that the intact reporter construct (full transcription factor or full active enzyme) is only formed when the two fused proteins interact. Aside from unraveling whole organism interactomes (Parrish et al., 2006) such systems have been successfully applied to detect over 300 novel phospho-tyrosine based PPIs related to cancer signaling pathways (Grossmann et al., 2015) or for the analysis of extra-cellular, cross-synaptic PPIs (Martell et al., 2016). One of the newest methods for high-throughput analysis of PPIs are protein-based micro-arrays (Hall et al., 2007), which helped, for instance, to identify inhibitors of protein complexes associated with the development of breast cancer (Na et al., 2014). However, the big caveats of these high-throughput approaches are the high rate of falsely identified (Huang et al., 2007; Williamson and Sutcliffe, 2010) or unspecific PPIs (Hayes et al., 2016) and the frequent loss of transient interactions (Yu et al., 2008).

Consequently, computational approaches have long played a key role in investigating PPIs. One task is to predict possible interaction partners as exactly as possible based only on non-structural information like sequence homology, gene co-expression, and phylogenetic profiling (Shoemaker and Panchenko, 2007). Importantly, the combination of such methods with experimentally derived structural information, homology modeling, or docking procedures has proven to yield more accurate results and more coverage of interactomes compared to experimental high-throughput methods (Zhang et al., 2012a). In the last decade, a lot of effort has been put into integrating the results of experimental and computational approaches into PPI-databases like STRING (Szklarczyk et al., 2011), BioGRID (Chatr-Aryamontri et al., 2015), or DIP (Salwinski et al., 2004). Computational tools can also further assist in the characterization of PPIs. For instance, the technique of *in silico* alanine scanning (Kortemme et al., 2004) allows for the replacement of every amino acid of a protein complex by alanine and the computation of the resulting change in binding free energy. Modern applications like mCSM (Pires et al., 2014) or FoldX (Schymkowitz et al., 2005) even predict the actual experimental change in binding free energy with high confidence. For details of the computational approaches targeted at PPIs and protein complexes see the reviews of Keskin et al. (2016), Shoemaker and Panchenko (2007), and Wetie et al. (2014).

4.1.2 Protein interfaces

So far, one important aspect of PPIs has been left out: What is their molecular basis? Eventually, every PPI comes down to three steps. First, the subunits have to mutually approach each other by diffusion, possibly supported by cellular co-localization. Second, geometrically and electrostatically complementary surfaces allow the subunits to arrange each other in a precise orientation, which, third, enables the formation of hydrophobic, ionic, or hydrogen-bond contacts between amino acid backbone and side-chain groups across the interface between the individual subunits. These interfaces are as diverse as PPIs themselves and often reflect the type of interaction that is conveyed (Conte et al., 1999; Jones and Thornton, 1997). For example, the interfaces of proteins that form non-obligate, transient complexes are often much smaller and more planar in geometry than the interfaces of proteins that form highly stable, permanent homo-dimers (Nooren and Thornton, 2003a). **Figure 16** illustrates the concept of geometric complementarity of interfaces at the example of an AS dimer from *S. typhimurium.*

Nevertheless, there is a consensus that interface regions significantly differ from the remaining protein surface (Glaser et al., 2001; Jones and Thornton, 1996; Ofran and Rost, 2003). Interfaces generally comprise more hydrophobic and aliphatic residues, whereas the remaining surface contains more polar and charged ones (Hamer et al., 2010). The interfaces of the two subunits of the AS dimer reflect this preference as both are relatively hydrophobic (**Figure 16**, dark red). Additionally, serine, alanine, and glycine residues are under-represented and arginine, cysteine, and histidine are over-represented in interfaces (Hamer et al., 2010). Also, the propensity of interface residues to form salt-bridges is generally higher (Sheinerman et al., 2000). However, as mentioned above, the precise characteristics of interfaces vary considerably between different types of protein complexes; for instance, the interfaces of obligate, homomeric complexes have a much higher proportion of hydrophobic residues than that of non-obligate, heteromeric complexes, which possibly reflects a hydrophobic effect-like principle during the simultaneous folding and association of the homomers (Ofran and Rost, 2003).

Interfaces differ from the remaining protein surface also with respect to amino acid conservation. In general, interface residues tend to be more conserved than other surface residues (Caffrey et al., 2004; Valdar and Thornton, 2001). On closer inspection, it is evident that most interfaces can also be sub-divided into a highly conserved central *core* and a less conserved, peripheral *rim* (Bouvier et al., 2009; Chakrabarti and Janin, 2002; Guharoy and Chakrabarti, 2005). The interfaces of the AS dimer also reflect this tendency; notice the almost completely conserved central core in both the synthase and glutaminase interface (**Figure 16**, dark magenta). The interface of the synthase subunit also shows that the highly conserved core regions are often more hydrophobic than the peripheral regions (Chakrabarti and Janin, 2002).

The common properties of protein interfaces have been exploited for the development of over 70 different computational procedures (Esmaielbeiki et al., 2016) that aim to predict interfaces from the primary sequence alone (Chen and Li, 2010; Yu et al., 2010a) or from a combination of sequence and structure information (Dong et al., 2014; Liu et al., 2014; Zellner et al., 2012). However, such methods strongly depend on the quality of available training data and their prediction quality is still limited (Yu et al., 2010b). Thus, most protein interface databases like PISA (Krissinel and Henrick, 2007) and PDBsum (de Beer et al., 2014) still rely on experimentally determined protein structures for the identification of interfaces .

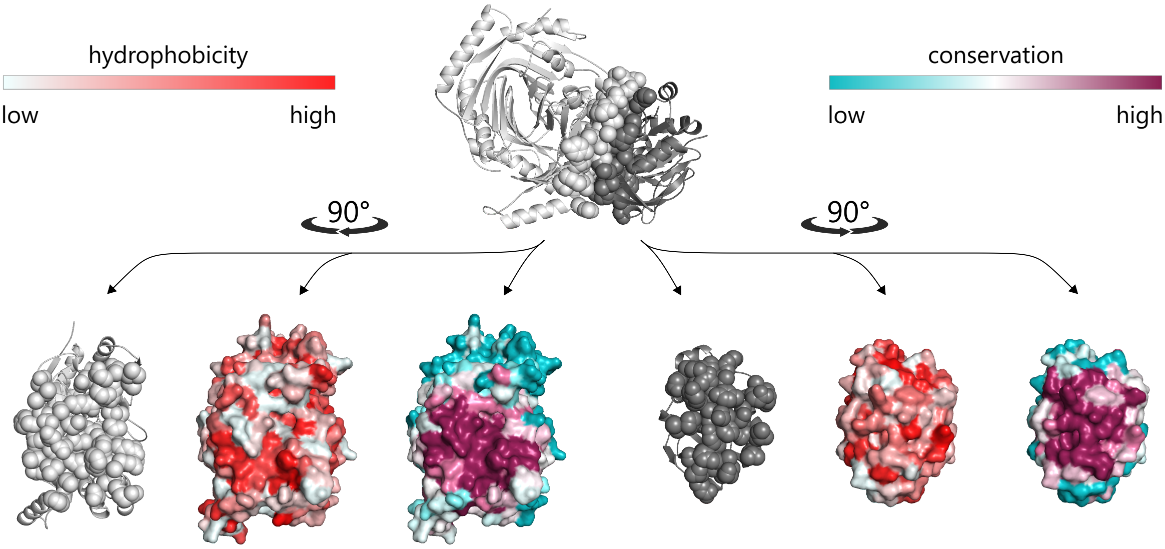


Figure 16. Illustration of common interface properties.
The interface between a synthase (light gray) and a glutaminase subunit (dark gray) of the AS complex from *S. typhimurium* demonstrates the shape-complementarity typical for protein interfaces. Interface residues are shown in sphere representation. In the lower half of the figure the synthase and glutaminase subunits are rotated by 90° clock-wise and counter clock-wise, respectively, to allow a plan view of the interfaces. The subunits are also shown in surface representation, either colored by hydrophobicity or amino acid conservation. Especially the synthase interface makes clear that protein interfaces comprise a highly conserved, mostly hydrophobic core region and a less conserved, more polar peripheral rim region. Hydrophobicity was calculated according to the scale from Eisenberg et al. (1984). Amino acid conservation in the interface was adapted from publication C.

Despite the overall conservation and the common properties of interfaces, it has become more and more evident that only very few amino acids in interfaces are actually important for forming and sustaining interactions between proteins. These so-called binding *hot-spots* are often responsible for the majority of binding free energy (Ma et al., 2001; Ofran and Rost, 2007; Reichmann et al., 2007) and are defined such that a substitution by alanine decreases this energy by at least 2 kcal mol^{-1}, thus significantly destabilizing the corresponding complex (Bogan and Thorn, 1998; Thorn and Bogan, 2001). These hot-spots are usually identified by site-directed mutagenesis (Fowler and Fields, 2014; Hinkley et al., 2011; Van Petegem et al., 2008) or *in silico* alanine-scanning (Kortemme et al., 2004; Robin et al., 2014).

4.1.3 Glutamine amidotransferases

A vivid example for the importance of PPIs are GATases. As mentioned in section **3.1.1** these non-obligate complexes are essential for the incorporation of nitrogen into a variety of metabolites. Interactions between their glutaminase and synthase subunits are important to shield the generated ammonia from the solvent and to couple its formation to a biosynthetic reaction (**Figure 17**). Hence, an unnecessary hydrolysis of glutamine and the formation of nascent, possibly toxic ammonia is prevented.

GATases can adopt various quaternary structures, ranging from heterodimers like the imidazole glycerolphosphate synthase (Douangamath et al., 2002) or ADCS (Parsons et al., 2002), to tetrameric complexes like AS (Knöchel et al., 1999; Morollo and Eck, 2001; Spraggon et al., 2001) and dodecamers like the pyridoxalphosphate synthase (Strohmeier et al., 2006; Zein et al., 2006). The synthase subunits are as diverse as the respective reactions they catalyze (Massiere and Badet-Denisot, 1998). The glutaminases on the other hand can be grouped into two classes based on the chemistry applied to hydrolyze glutamine. Class I glutaminases display an α/β-hydrolase like fold (Ollis et al., 1992) and possess a catalytic triad comprising cysteine, histidine, and glutamate for the hydrolysis of glutamine via a glutamyl thioester intermediate (Thoden et al., 1998). This class includes AS (**Figure 17**), ADCS, imidazole glycerolphosphate synthase (Douangamath et al., 2002), carbamoylphosphate synthase (Thoden et al., 1997), formylglycinamide synthase (Anand et al., 2004), guanosine-5'-monophosphate synthase (Tesmer et al., 1996), cytidine-5'-triphosphate synthase (Goto et al., 2004), and pyridoxalphosphate synthase (Strohmeier et al., 2006). Class II glutaminases also posses an α/β-hydrolase like fold (Ollis et al., 1992) but contain an N-terminal nucleophilic cysteine instead of a catalytic triad and are thus often referred to as N-terminal nucleophile (NTN) glutaminases (Zalkin and Smith, 2009). This class includes asparagin (Nakatsu et al., 1998) and glutamate synthase (Binda et al., 2000), glutamine phosphoribosylpyrophosphate amidotransferase (Smith, 1998), and glucosamine-6-phosphate synthase (Milewski, 2002).

Despite the different classes of glutaminases and the considerable structural and functional variability of synthases, most GATases act in a similar, highly coordinated manner (Bera et al., 1999; Goto et al., 1976; Miles et al., 1998). Usually, glutamine hydrolysis is allosterically triggered by binding of the synthase and the acceptor substrate. Several cases can be discerned in this context: The glutaminase activity of the pyridoxalphosphate synthase from *Bacillus subtilis* critically depends on complex formation but is not further increased by the presence of an acceptor substrate in the synthase active site (Raschle et al., 2005). In a certain way similar, interaction with the synthase subunit is mandatory for glutaminase activity in AS (List et al., 2012) and imidazole glycerol phosphate synthase

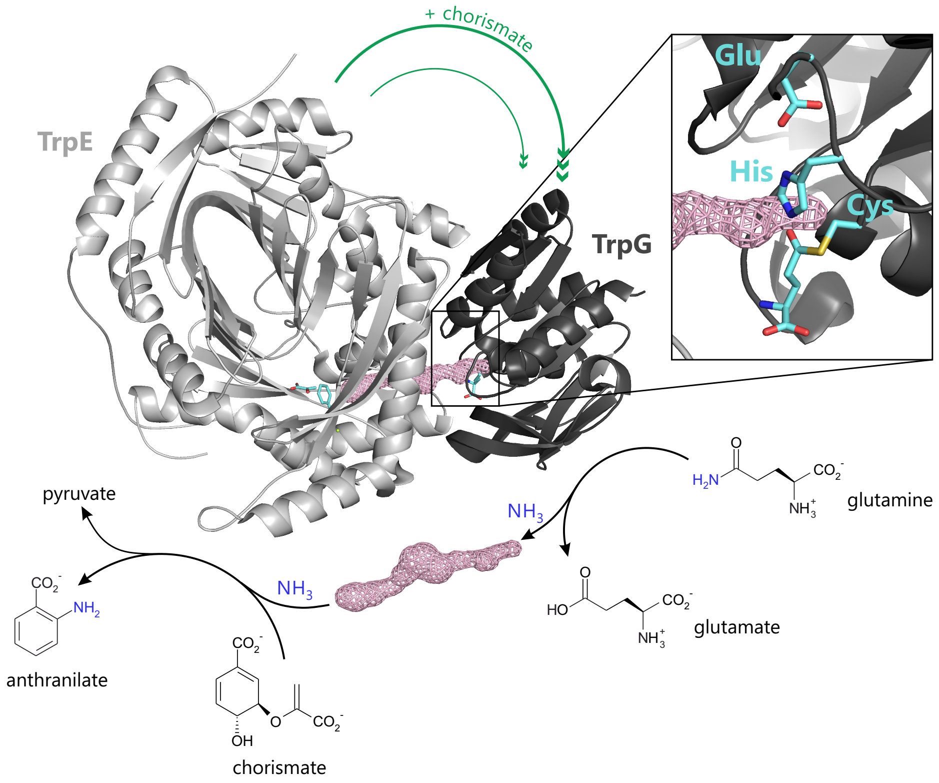


Figure 17. Anthranilate synthase as an example of glutamine amidotransferases. AS catalyzes the formation of anthranilate from chorismate and glutamine. Here, one of the two TrpE:TrpG dimers of the AS from *S. marcescens* is shown (PDB ID 1i7q). The glutaminase subunit TrpG is a class I glutaminase, which utilizes a catalytic triad comprising cysteine, histidine, and glutamate (see cut-out) for the hydrolysis of glutamine into glutamate and nascent ammonia. The ammonia is subsequently channeled through a roughly 40 Å long channel (pink mesh object) to the active site of the synthase subunit TrpE (represented by cyan benzoate and pyruvate stick models). After elimination of pyruvate, the amino-substituted product anthranilate is released. The activity of the glutaminase TrpG is dependent on the binding to TrpE (thin green arrow) and the additional presence of chorismate in the active site of TrpE leads to a further allosteric activation of TrpG (thick green arrow).

(Beismann-Driemeyer and Sterner, 2001). In these complexes, however, the additional binding of the acceptor substrates PRFAR or CH (**Figure 17**) drastically increases the reaction rate of glutamine hydrolysis 30-fold and up to 4900-fold, respectively (Klem et al., 2001; Klem and Davisson, 1993; Zalkin and Smith, 2009). Although similar activation mechanisms are observed for ADCS, their glutaminases also display basal, albeit weak, glutamine hydrolysis activity. This indicates that the functional coupling between glutamine hydrolysis and ammonia consumption is not always fully guaranteed.

4.2 Summary and Discussion

4.2.1 The role of interface add-ons for protein-protein interaction specificity

PPI-networks are often highly inter-connected and of complex topology. The *E. coli* and human interactomes deposited in the STRING database comprise hundreds of thousands of interactions, inferred from high-throughput experiments, gene-neighborhoods, co-expression, other databases, and text mining (Szklarczyk et al., 2011). This high connectivity suggests that many proteins not only interact with one specific but with multiple, different partners. One example is the well-characterized p53 protein, the key player in response to DNA damage and oncogene activation. p53 assuredly interacts with at least 30 different proteins (Dartnell et al., 2005). Just the same, the NFκB transcription factor forms complexes with a multitude of different partners in various signaling pathways (Nooren and Thornton, 2003a).

However, PPI-networks tend to overstate the number of multi-specific proteins, as they often contain a high number of falsely identified and non-specific interactions (Hayes et al., 2016; Williamson and Sutcliffe, 2010). The *E. coli* interactome is drastically simplified to about 2200 interactions when only high-confidence, experimental evidence is considered (**Figure 18A**). Moreover, this reduction makes clear that over two thirds of all proteins that undergo any kind of interaction only bind to one or two other proteins. Thus, most interactions are highly specific, which means that many proteins are engaged in only a single or a very limited number of complexes (Schreiber and Keating, 2011). Such specific PPIs are crucial for many essential physiological functions like the assembly of catalytic enzyme complexes or the formation of toxin-antitoxin (Fiebig et al., 2010) or antibody-antigen complexes (Van Regenmortel, 2014).

But how is interaction specificity achieved and which conditions may influence it? First of all, proteins do not exist in isolation but rather in the crowded environment of a cell. For example, *E. coli* expresses over 4000 different proteins (Karp et al., 2002). The resulting macromolecular crowding is extensive – up to 30% of cellular dry mass is composed of proteins (Elcock, 2010) – and it can even limit protein diffusion (Wang et al., 2010). Proteins thus logically come into contact with a wide variety of others inside a cell (Levy et al., 2012; Sarkar et al., 2013). Interaction specificity consequently depends on that proteins not only bind their native partners but also disfavor non-native interactions and thus avoid the formation of non-physiological complexes (Sikosek and Chan, 2014). The affinity between and the local concentration of the protomers as well as possible sub-cellular

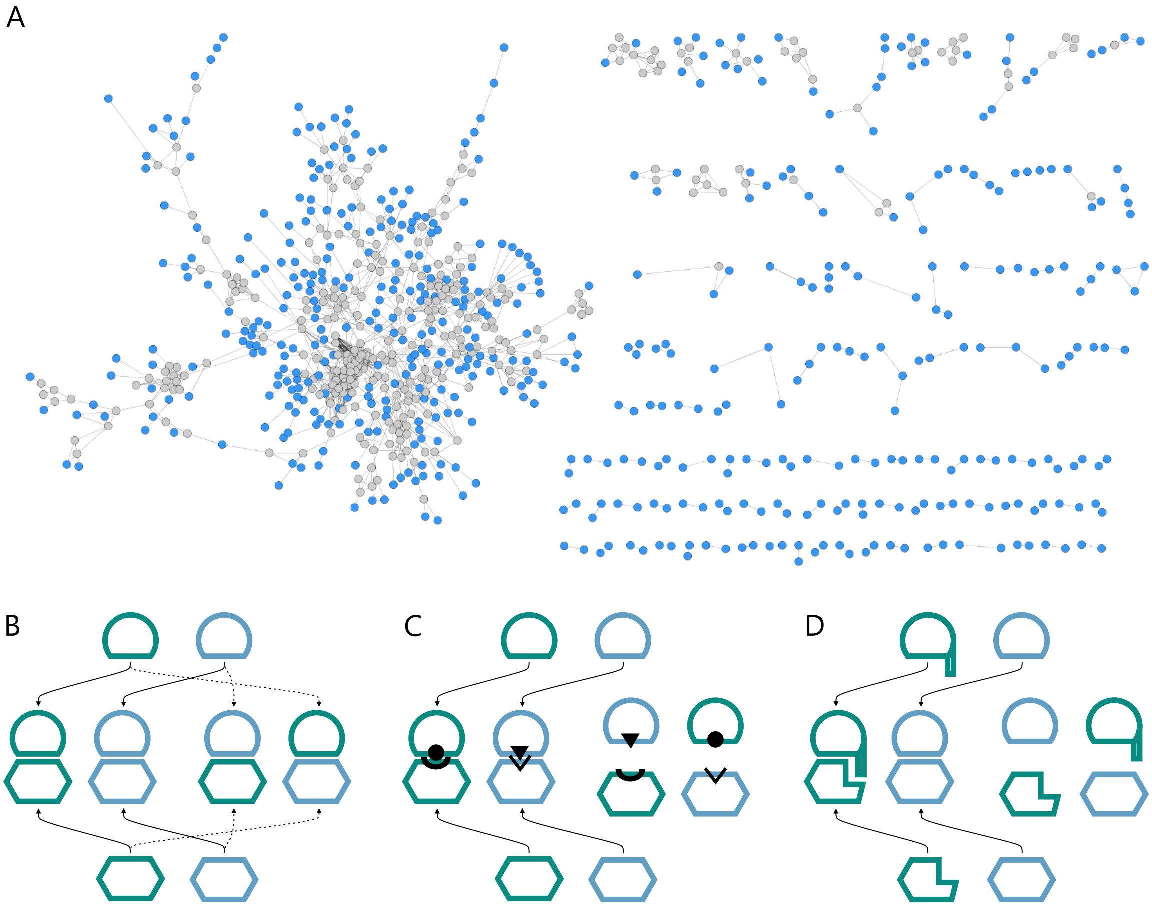


Figure 18. Protein-protein interaction specificity.
(**A**) Network representation of a part of the *E. coli* interactome derived from high-confidence, experimental data deposited in the STRING database. The network shows the 878 proteins (nodes) that interact with at least one other protein and the resulting 2285 interactions (edges). Proteins with only one or two interaction partners are represented by blue nodes (**B-D**) Illustration of PPI specificity. (**B**) In principle, several homologous subunits with similar interface geometries (circle segments and hexagons, respectively) create the possibility of erroneous interactions (dashed arrows) in addition to native interactions (solid arrows). (**C**) Specificity determining mutations (black) are a possible means to enforce native interactions. (**D**) An interface add-on (green extension) as an alternative means to prevent erroneous interactions without modifying the conserved interface core.

co-localization have been described to play important roles in this context (Schreiber and Keating, 2011). The most important point for interaction specificity, however, is the geometric and electrostatic complementarity between matching protein interfaces and the lack of it between non-related interfaces (Nooren and Thornton, 2003a). But the necessary variability in interface structures or geometries has its limitations: First, the protein structure space is finite (Levitt, 2009; Skolnick et al., 2009; Zhang and Skolnick, 2005) and comprises not more than 1000-1500 different folds (Chothia, 1992; Gao and Skolnick, 2010), most likely due to the physical constraints resulting from the need for efficient and compact packing of hydrogen-bonded secondary-structure elements during protein folding (Finkelstein and Ptitsyn, 1987; Zhang et al., 2006). Consequently, the same biophysical constraints limit the number of possible interface geometries to around one thousand (Gao and Skolnick, 2010). Second, the overall relatively flat architecture of protein interfaces as well as functional constraints additionally limit larger structural variations. For example, several serine protease-inhibitor complexes utilize highly similar interfaces in spite of structurally different inhibitors (Gao and Skolnick, 2010). A direct consequence of these limitations is the relatively low number of possible protein quaternary structures of supposedly only 4000 (Garma et al., 2012). Taken together, it is not surprising that the detailed molecular basis of interaction specificity is not well understood.

The problem of assuring interaction specificity becomes even more challenging, if several homologous interaction partners exist (**Figure 18B**). In this case, similarities in sequence, structure, and consequently interface geometry and electrostatics create the risk of erroneous interactions and formation of non-physiological complexes. This has been recognized as a particular problem in large, paralogous enzyme families (Schreiber and Keating, 2011). One obvious solution is the diversification of interfaces by adaptational mutations (**Figure 18C**) (Capra et al., 2012). In this scenario, complementary mutations in the native interaction partners prevent non-native cross-interactions. Although interfaces possess such variability (DePristo et al., 2005; Harms and Thornton, 2013), there are several restraints limiting mutational diversification. First, the alphabet of proteinogenic amino acids is limited and hence also the set of different physicochemical properties of interfaces (Sikosek and Chan, 2014). Second, interfaces often convey a function beyond binding; e.g. the formation of a shared active site (Giroux et al., 1994; Wente and Schachman, 1987), the propagation of allosteric signals (Beismann-Driemeyer and Sterner, 2001; Rivalta et al., 2012), or the transfer of reaction intermediates (Huang et al., 2001). These constraints are reflected in the lower mutational rate of interface residues relative to non-interface ones and the higher number of functionally equivalent and neutral mutations (Ames et al., 2016; Mintseris and Weng, 2005).

We hypothesized that another concept exists, which avoids the limitations described above. PPIs require the geometric complementarity of protein interfaces; local variations in interface geometries themselves might thus contribute to the specification of interactions. We speculated that such geometric variability might preferentially exist on the peripheral rim regions of interfaces that are not as conserved as the central core

(Bouvier et al., 2009; Guharoy and Chakrabarti, 2005). In a comprehensive computational survey of heteromeric, bacterial protein complexes, we identified large structural insertions at protein interfaces that significantly change interface geometry and named them accordingly interface add-ons (**Figure 18D**). Interface add-ons typically comprise 10-20 amino acids and form well-defined secondary structure elements. They are also crucial for PPIs in the respective complexes, because they contain at least one and mostly three or more binding hot-spot residues, which means that a substitution of these amino acids by alanine significantly destabilizes the corresponding protein complexes (Bogan and Thorn, 1998; Thorn and Bogan, 2001).

A detailed description of this survey is available in publication C. In brief, we compiled a reference set of bacterial, heteromeric protein complex structures and compared them with their homologs of the corresponding superfamilies. By that we were able to identify in the reference complexes large insertions that are not present in other superfamily homologs. Vice versa, we identified in some superfamily homologs large insertions that are missing in the reference complexes. Protein superfamilies are well suited for this approach, because they provide a high level of classification and group proteins with a broader structural and functional relationship than the more narrowly defined, often iso-functional protein families. Several filter routines made sure that only such insertions were accepted as interface add-ons that conform with the points made above. In total, we found interface add-ons in about 10% of the analyzed complexes. Moreover, they are not limited to certain phyla and we detected them in complexes from Actinobacteria, the Deinococcus-Thermus group, Firmicutes, Thermotogae, and γ-Proteobacteria.

In the following, four examples are be described that illustrate why interface add-ons may contribute to PPI-specificity. Cysteine desulfurases (**Figure 19A**) are enzyme complexes that catalyze the transfer of sulfur from cysteine or selenocysteine to various acceptor substrates in different biosynthetic pathways (Loiseau et al., 2005; Mihara and Esaki, 2002). In *E. coli*, the CsdA$_2$:CsdE$_2$ complex contains interface add-ons in the CsdA subunits, which form parts of the interface with CsdE. Importantly, *E. coli* also contains the very similar IscS$_2$:TusA$_2$ complex, in which IscS is homologous to CsdA but TusA is structurally and functionally different from CsdE. The IscS$_2$:TusA$_2$ complex does not contain the interface add-ons, which suggests that they might contribute to the differentiation of the CsdA and IscS homologs and their respective biosynthetic pathways.

A second example are the isocitrate- and isopropylmalate dehydrogenase complexes from *E. coli* (**Figure 19B**). The former contains α-helical interface add-ons in its catalytic IcdA subunits that are part of the interface with the isocitrate-dehydrogenase-kinase/phosphatase AceK. Binding of the IcdA catalytic to the AceK regulatory subunits is

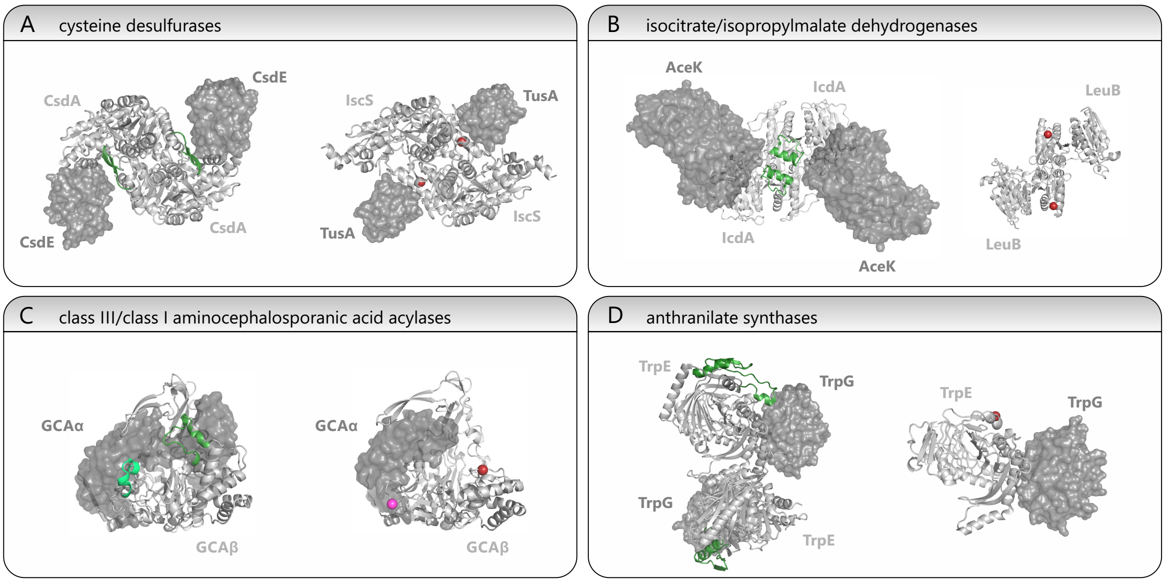


Figure 19. Examples of heteromeric protein complexes with interface add-ons. Subunits with interface add-ons are shown in cartoon representation; other subunits in surface representation. Interface add-ons are colored green and their location is marked by red spheres in complexes that lack the interface add-ons. The examples are two cysteine desulfurases from *E. coli* (**A**), the isocitrate and isopropylmalate dehydrogenases from *E. coli* (**B**), two aminocephalosporanic acid acylases from Pseudomonas species (**C**), and two anthranilate synthases from *S. typhimurium* and *S. solfataricus* (**D**).

important for tuning isocitrate flux between the citric acid and the glyoxylate cycle (Laporte and Koshland, 1982; Laporte, 1993). The isopropylmalate synthase LeuB, which is part of leucine biosynthesis in *E. coli*, does not contain the interface add-on. The interface add-on in IcdA might allow for a specific recognition by the regulatory AceK subunits and thus differentiate the isocitrate dehydrogenase from the isopropylmalate dehydrogenase for which no comparable phosphorylation/dephosphorylation mechanism is known.

Another example of homologous protein complexes that are differentiated by an interface add-on are the glutaryl-7-aminocephalosporanic acid acyclases from Pseudomonas species (**Figure 19C**). These enzyme complexes are involved in penicillin and cephalosporin biosynthesis (Kim et al., 1999). Class III representatives contain in their β-subunit two interface add-ons, which form large parts of the interface with the α-subunit (Golden et al., 2013). In contrast, class I representatives lack both interface add-ons and the interface between the two subunits is less extended.

The final example shall be ASs. These GATase complexes catalyze the committed step in tryptophan biosynthesis (Zalkin and Smith, 2009). Tetrameric ASs like the one from *S. typhimurium* contain a large interface add-on in the TrpE synthase subunit, which

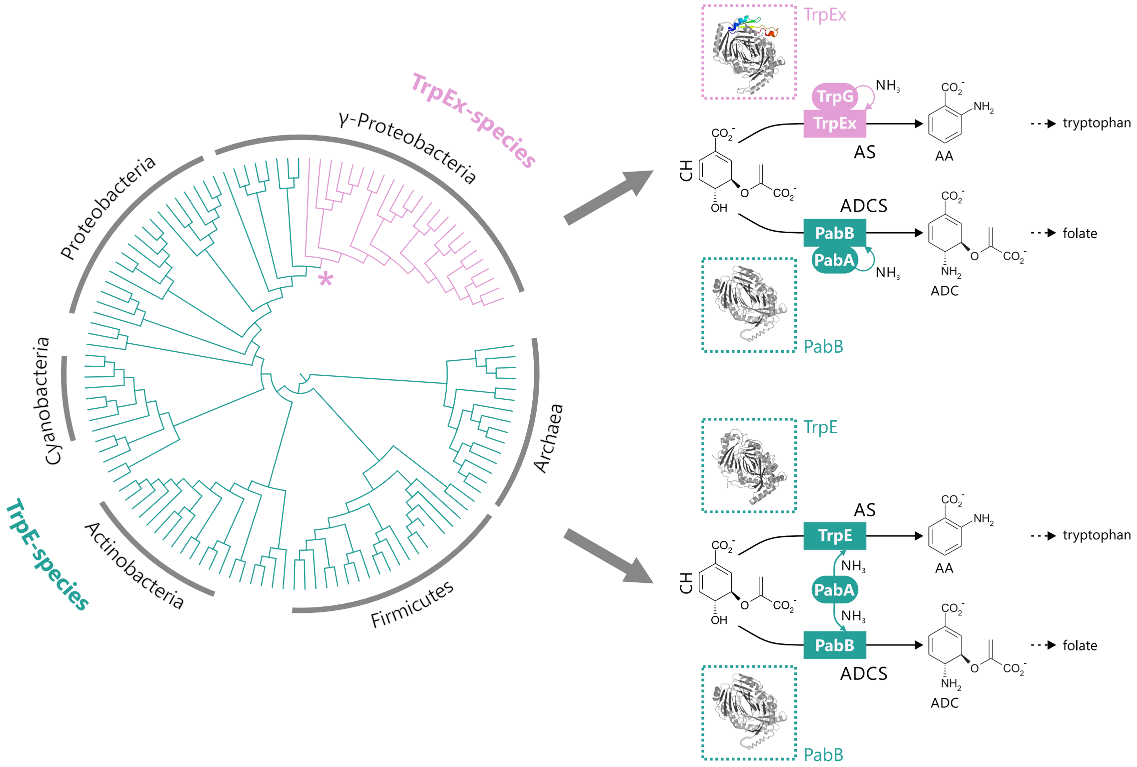


Figure 20. Genetic profiling of prokaryotic species for TrpEx, TrpE, PabB, TrpG, and PabA homologs and the functional consequences.
The phylogenetic tree is representative for the over 15 000 profiled archaeal and bacterial genomes. It shows that all TrpEx-species (magenta) are γ-Proteoacteria that arose from the split between *Pseudomonas* and *Shewanella* species (indicated by an asterisk). TrpEx-species possess two paralogous AS and ADCS complexes. The interface add-on in TrpEx is colored in a rainbow gradient. All Archaea and the remaining Bacteria are TrpE-species (green), which possess only a single PabA glutaminase for both AS and ADCS complexes.

forms parts of the interface to the TrpG glutaminase subunit (**Figure 19D**). Dimeric ASs like the one from *S. solfataricus*, in contrast, do not contain the interface add-on. In order to avoid any possibility of confusion, TrpE homologs that contain the interface add-on are hereafter denoted by "TrpEx" (extended).

A genetic profiling of over 15 000 archaeal and bacterial species revealed that TrpEx proteins are limited to a group of γ-Proteobacteria. **Figure 20** shows a representative phylogenetic tree with the branches leading to these TrpEx-species colored in magenta. TrpEx-species also possess the homologous ADCS complex, which catalyzes a reaction similar to that of AS as the committed step of folate biosynthesis (**Figure 20**, top right). Both synthase subunits of the complexes (TrpEx, PabB) as well as both glutaminase subunits (TrpG, PabA), are homologs and share high sequence and structure similarity.

The majority of prokaryotic species, in contrast, possess ASs with TrpE-subunits that lack the interface add-on; the branches to these TrpE-species are colored in green in **Figure 20**. Not surprisingly, also TrpE-species possess the folate biosynthetic pathway. However, our profiling showed that they only contain a PabA glutaminase and no TrpG glutaminase. This suggests that both AS and ADCS complexes share the same glutaminase subunit for their respective reactions in tryptophan and folate biosynthesis. These results confirmed early reports that Firmicutes like *B. subtilis* rely on a single, amphibolic glutaminase for the two biosynthetic pathways (Kane, 1977). Given the high sequence and structure similarity between the two synthase subunits TrpE and PabB it is not surprising that the single PabA glutaminase is re-used for both AS and ADCS complexes; a tendency that has also been observed for other protein complex subunits (Kühner et al., 2009). The presence of just one glutaminase in TrpE-species but two different glutaminases in TrpEx-species suggests that the interface add-on in TrpEx acts as a negative design element to differentiate the otherwise highly similar AS and ADCS complexes and to ensure specific TrpEx-TrpG and PabB-PabA interactions.

We experimentally tested this assumption by heterologously expressing and purifying several TrpEx, TrpE, and PabB synthases, as well as several TrpG and PabA glutaminases and characterizing their ability to form AS and ADCS complexes by chromatographic methods, light scattering, mass spectrometry, and enzyme kinetic measurements (**Figure 21**). Such a combination of several orthologous methods is important to minimize the probability of detecting false-positive interactions or missing interactions due to false-negative results (Hayes et al., 2016). As expected, we only observed interactions between TrpEx synthases and TrpG glutaminases on the one hand and between TrpE or PabB synthases and PabA glutaminases on the other hand (**Figure 21A**). Interestingly, the conservation of the interfaces is so extensive that even glutaminases and synthases from different species bind to each other. Moreover, practically all observed AS and ADCS complexes were enzymatically active and formed AA or ADC in a glutamine-dependent manner, indicating functional, GATase-typical ammonia channeling (**Figure 21B**). Moreover, these complexes also displayed allosteric communication between the subunits, because the activity of the glutaminases was stimulated by binding of the different synthases (**Figure 21C**). In summary, these results indicate that the interface add-on determines interaction specificity in AS and ADCS complexes. We could further experimentally affirm this by partially deleting the interface add-on in TrpEx and by showing that the resulting TrpEx-variant was able to form a complex with PabA and to stimulate its glutaminase activity (Publication C).

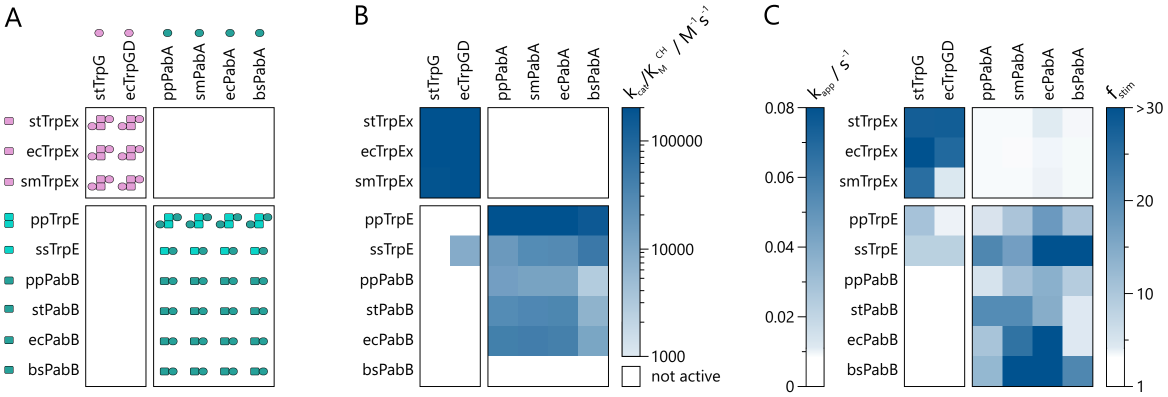


Figure 21. Structural and functional characterization of interactions between TrpEx, TrpE, and PabB synthases and TrpG and PabA glutaminases.
(**A**) Oligomeric states of glutaminases, synthases, and complexes determined by size-exclusion chromatography, static light scattering, and native mass spectrometry. (**B**) Catalytic efficiencies of the glutamine-dependent conversion of CH to AA (for complexes with TrpEx or TrpE) and ADC (for complexes with PabB). Complexes with PabB from *B. subtilis* (bsPabB) were inactive under the applied experimental conditions. (**C**) Apparent turnover rates of glutamine hydrolysis by TrpG (k_{app}) and stimulation of PabA glutamine hydrolysis by the presence of the different synthases (f_{stim}). The figure is modified from publication C.

Other studies investigating the role of insertions and deletions for PPI-specificity to the same computational and experimental extent are rare. Panchenko and co-workers investigated the tendency of proteins to form homo-oligomers (Hashimoto and Panchenko, 2010). Based on *in silico* analyses, the authors showed that insertions mediating homo-oligomerization have a significant tendency to be located at protein interfaces. However, although over 100 protein complexes with such insertions were identified, most of them comprised only one to four amino acids and were merely extensions of already existing secondary structure elements of protein interfaces. Moreover, the interfaces of homo-oligomers additionally differ from other types of protein interfaces in terms of amino acid composition and residue-residue contact preferences (Ofran and Rost, 2003) and are thus a relatively special case. Although the authors speculated that the insertions might differentiate homo-dimeric and monomeric paralogs and thus prevent the formation of undesired interactions and assure the separation of functional pathways, no experimental evidence for such characteristics of insertions in protein interfaces are available to date. The following section describes our experimental approach to show that interface add-ons display such characteristics.

4.2.2 Biological relevance of interface add-ons

The observation that TrpEx and TrpG have been retained in all descendants of the split between Pseudomonas and Shewanella species approximately 950 million years ago (Battistuzzi and Hedges, 2009), suggests that the specific formation of AS and ADCS complexes confers a selective advantage, or at least does not impair the fitness of the respective organisms. Given the central roles of the two complexes in the biosynthetic pathways leading to the essential metabolites tryptophan and folate, we assumed that such a selective advantage could be the prevention of cross-interactions between AS and ADCS synthases and glutaminases and thus of potentially harmful biosynthetic cross-talk.

Such cross-talk is in principle possible in TrpE-species, which contain only PabA that interacts with both TrpE and PabB. Firmicutes have evolved sophisticated regulatory mechanisms acting on the transcriptional and translational level to assure that the intracellular concentration of PabA is always properly adjusted to the demands of tryptophan and/or folate biosynthesis (Babitzke, 1997; Yakhnin et al., 2007; Yanofsky, 2007). However, the participation of PabA in both AS and ADCS complexes fundamentally links tryptophan and folate biosynthesis together and creates the risk of potentially harmful metabolic cross-talk. We simulated such a situation by over-expressing either *pab*B or *trp*E in *B. subtilis* cells and monitoring their growth on minimal medium plates (**Figure 22**). We hypothesized that high intracellular amounts of PabB would take up all available PabA glutaminase and thus deduct it from folate biosynthesis. Similarly, the over-expression of *trp*E could impair folate biosynthesis. Although no effect was observed for over-expression of *pab*B, the over-expression of *trp*E significantly impaired cell growth (**Figure 22A**). We could also demonstrate that this effect is based on compromised folate biosynthesis, because the supplementation of folate or an intermediate of its biosynthesis completely compensated the effect of *trp*E over-expression (**Figure 22B**). A similar metabolic conflict is impossible in TrpEx-species, because their glutaminases are highly specific for the respective synthases and consequently two independent, paralogous AS and ADCS complexes exist. It remains to be shown if similar effects can be observed for other paralogous complexes. Interesting targets would be the cysteine desulfurases or the isocitrate/isopropylmalate dehydrogenase complexes (**Figure 19A, B**). In both cases *E. coli* contains one complex with and one complex without an interface add-on that are part of different metabolic pathways.

One other aspect of the TrpEx interface add-on that has possible practical applications should be mentioned at this point. Many TrpEx-species are highly pathogenic, causing cholera, severe gastroenteritis, respiratory tract infections, the bubonic plague, haemolytic

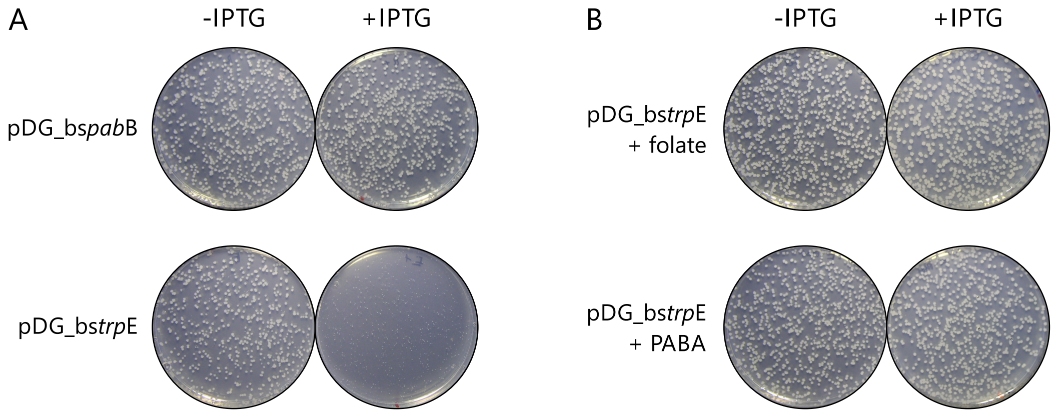


Figure 22. *B. subtilis* growth experiments.
(**A**) Minimal medium plates with *B. subtilis* colonies grown from cells transformed with plasmids carrying the genes for PabB or TrpE from *B. subtilis*. The plates, either lacking or containing IPTG for induction of protein expression, were incubated at 37° for 48 hours. (**B**) The supplementation of the minimal medium with either folate or an intermediate of its biosynthesis (*p*-aminobenzoic acid, PABA) offsets the detrimental effect of bs*trp*E over-expression. The figure is modified from publication C.

syndromes, or typhoid (Berman, 2012). In the light of more and more prevalent multi-drug resistant bacterial strains, the development of alternatives to traditional antibiotics is among the most important future tasks for research and medicine (Czaplewski et al., 2016). Inhibitors of PPIs are one possible alternative and act by preventing the formation of protein complexes that are either essential for pathogenicity or viability of the pathogens. Some small molecule- and antibody-based PPI-inhibitors have already shown great therapeutic and clinical potential (Petta et al., 2015). However, such inhibitors often have a limited cell-delivery, structural flexibility, or chemical variability, which is why peptide-based PPI-inhibitors have come into focus more and more (Araghi and Keating, 2016). For example, the peptide enfuvirtide, which is derived from a C-terminal fragment of the HIV envelope glycoprotein gp41, blocks membrane fusion of the virus particles and thus inhibits viral cell entry (Matthews et al., 2004). Moreover, peptides have been developed that target Bcl-2 family complexes, which are involved in the formation and progression of several cancer types (Oltersdorf et al., 2005), and p53-binders, which impede the guarding role of p53 in the cell cycle (Liu et al., 2010).

Pathogenic TrpEx-species might be similarly targeted with a selective PPI-inhibitor. One possibility would be to design a peptide that mimics the 3D-structure of the interface add-on in the TrpEx subunit of AS and thus compete with TrpEx for binding to the TrpG subunit. Vice versa, the interface part of TrpG that contacts the TrpEx interface add-on might serve as a starting scaffold for the grafting of a peptide that tightly binds to the interface add-on and thus prevents TrpG from binding to TrpEx. Targeting the TrpEx-type AS to engage pathogenic Proteobacteria is a feasible approach for several reasons: First, AS is essential for tryptophan biosynthesis and its blockage is lethal for many pathogenic bacteria (Leonhardt et al., 2007; Wang et al., 2014; Zhang et al., 2012c). Second, AS is absent from eukaryotic cells, which reduces the possibility of adverse side effects of peptide-based inhibitors on human cells. Third, all bacteria other than Proteobacteria contain TrpE instead of TrpEx. Thus a TrpEx-specific inhibitor would not impair many of the beneficial bacterial species in the human gut microbiome, of which only about 2% are Proteobacteria (D'Argenio and Salvatore, 2015).

4.2.3 Evolution of protein-protein interactions in AS and ADCS complexes: The role of the interface add-on

As we have shown, a possible cross-talk between AS and ADCS complexes and hence tryptophan and folate biosynthesis can impair cellular fitness. Such a cross-talk is highly unlikely in TrpEx-species, which have evolved specific interactions between TrpEx and TrpG in the AS complex and between PabB and PabA in the ADCS complex. It can be assumed that the resulting effective separation of these two essential biosynthetic pathways and the concomitant possibility for differential and independent regulation conferred selective advantages to TrpEx-species, because TrpEx and TrpG have been retained since their first emergence in Shewanella species.

The phylogenetic distribution of TrpEx, TrpE, and PabB suggests that TrpEx-species have evolved from TrpE-species (**Figure 20**). This transition requires at least two events: The incorporation of the interface add-on into TrpE and the emergence of the TrpG glutaminase. The origin of the interface add-on cannot reliably be traced back; we assume that the corresponding DNA sequence has been integrated into the *trp*E gene via homologous recombination or events related to horizontal gene transfer. TrpG has most likely evolved via duplication of an ancestral PabA glutaminase and a subsequent change in interaction specificity from TrpE/PabB to TrpEx.

But how do protein complexes evolve such changes in interaction specificity? Computational studies have demonstrated that interacting proteins typically co-evolve with similar mutational rates (Fraser et al., 2002; Ovchinnikov et al., 2014). The implicit assumption underlying this idea is that a mutation in one complex subunit that would disrupt the interaction on its own is balanced by a compensating mutation in the other subunit. The interplay of these two (or more) mutations assures that the resulting complex is specific

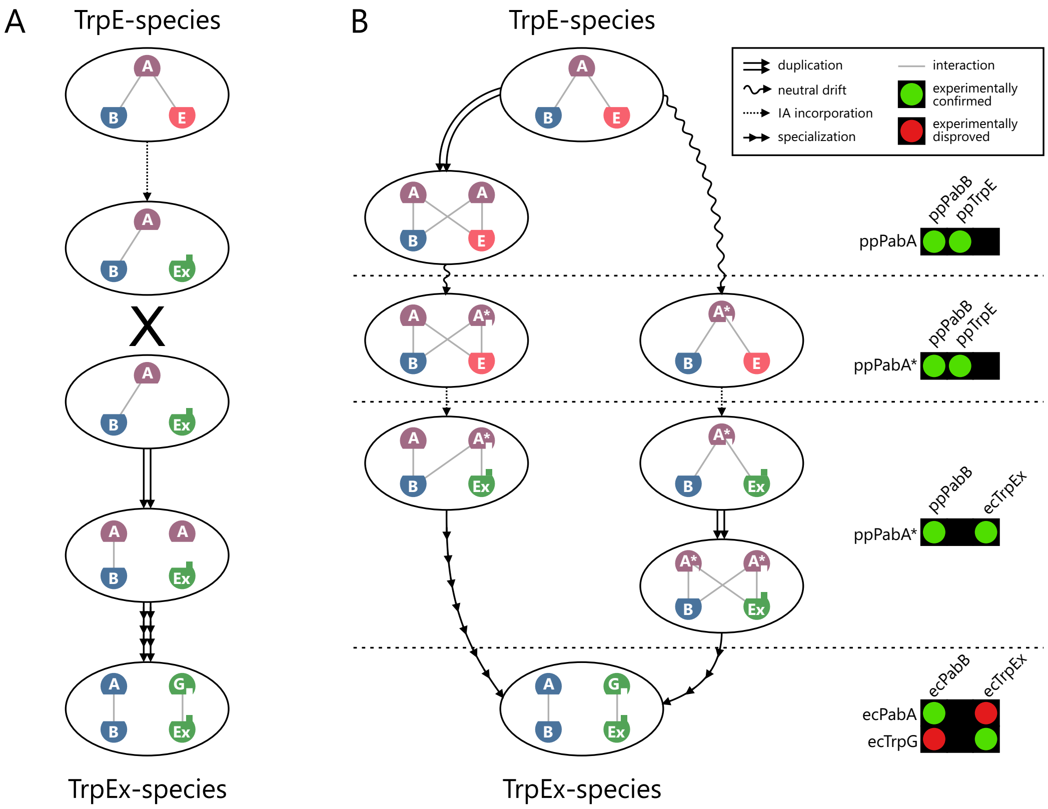


Figure 23. Possible scenarios for the evolution of TrpEx and TrpG.
The scenarios describe the evolution leading from TrpE-species (single PabA glutaminase and the two synthases PabB and TrpE) to TrpEx-species (two specific glutaminases PabA and TrpG for the two synthases PabB and TrpEx). (**A**) Compensatory model. The emergence of TrpEx in an organism with just a PabA glutaminase creates a non-functional intermediate (X). Granted that this dead-end has been overcome somehow, after duplication of the gene for PabA, compensatory mutations lead to the development of TrpG. (**B**) Promiscuous model. The two illustrated paths both involve a promiscuous PabA* glutaminase that interacts with TrpE/PabB synthases as well as with TrpEx synthases. On the right hand side experimentally characterized interactions are shown that represent the respective evolutionary phases. The figure is modified from publication C.

and that its subunits no longer bind to their predecessor partners. Applied to the evolution of TrpEx-species, this *compensatory* model (Aakre et al., 2015) required a duplication of the *pab*A gene after the emergence of TrpEx and the subsequent accumulation of mutations in the duplicate to shape its interaction specificity towards TrpEx (**Figure 23A**). However, the implicit flaw of the compensatory model in this case is clear: After the emergence of TrpEx an organism would be non-viable due to missing interactions between TrpEx and PabA and thus non-functional tryptophan biosynthesis. In general terms, the compensatory model requires that the evolution of a protein complex passes through a non-functional state, when the first mutation (incorporation of the interface add-on) has not yet been compensated by the second one (adaptation to the interface add-on).

Given the critical role of AS complexes in metabolism, non-functional states in their evolution are highly unlikely. We thus propose another scenario (**Figure 23B**), inspired by recent findings of Laub an co-workers (Aakre et al., 2015). This *promiscuity* model assumes that the "adaptation" of PabA to TrpEx has preceded the emergence of the latter. This pre-adaptation requires mutations in the interface region of PabA that maintain productive interactions with PabB and TrpE but at the same time tolerate the binding of a TrpEx synthase with an interface add-on. How easily is such a promiscuous state reached from a PabA scaffold? We could show by rational protein design that only five mutations in PabA from the TrpE-species *Pseudomonas putida* are sufficient to achieve such interaction promiscuity. The resulting ppPabA* variant was able to form functional AS complexes with both TrpE and TrpEx synthases, as well as functional ADCS complexes with a PabB synthase (**Figure 23B**, traffic-light visualizations). Importantly, the catalytic efficiencies of the resulting complexes were only one order of magnitude lower than those of native AS and ADCS complexes. This suggests that such a promiscuous glutaminase could have supported both tryptophan and folate biosynthesis *in vivo*.

In principle, two different evolutionary trajectories are in line with the promiscuity model. The neo-functionalization trajectory (**Figure 23B**, left) involves an initial duplication of the gene for PabA, allowing one copy to accumulate mutations that lead to a relaxed interaction specificity. After the emergence of TrpEx this promiscuous PabA* copy became essential for maintaining tryptophan biosynthesis and mutational drift and co-evolution eventually yielded the specific AS and ADCS complexes of TrpEx-species. Alternatively, in the sub-functionalization trajectory (**Figure 23B**, right) mutational drift led to a relaxed interaction specificity in the single PabA glutaminase, which tolerated the emergence of TrpEx. Duplication of the promiscuous PabA* glutaminase was probably favored by gene dosage effects and sub-functionalization of the copies eventually led to the specific AS and ADCS complexes of TrpEx-species.

How likely is the promiscuity model for explaining the evolution of the TrpEx-TrpG interaction? One important aspect is that it avoids any non-functional evolutionary intermediate. However, it heavily relies on gene duplication and mutational drift to explain the development of the promiscuous PabA* glutaminase and the sub- or neo-functionalization to contemporary TrpG. It has recently been shown that the rate of sequence divergence after gene duplication of protein complex subunits is relatively high with about 6×10^{-3} per PPI and per million years (Ames et al., 2016). Moreover, work with yeast has demonstrated that gene duplicates evolve up to three times faster than single copy genes and that duplication is usually followed by a burst in sequence divergence (Conant and Wagner, 2003; Scannell and Wolfe, 2008). Taken into account that only five mutations separate the specific PabA and the promiscuous PabA* states, it can be assumed that relaxed interaction specificity could be easily reached during evolution.

Similar low hurdles between interaction specificity and promiscuity have been observed for other protein complexes. Bacterial toxins are stable and globular proteins that inhibit cell growth or impair cell viability unless bound and sequestered by their cognate antitoxin (Hallez et al., 2010). New toxin-antitoxin pairs often arise from gene duplication and subsequent divergence of interaction specificities (Yamaguchi et al., 2011). During this process it is critical that toxins can evolve new specificities while never losing interaction with their cognate antitoxin to prevent self-inflicted toxicity. It has recently been shown that only four mutations are sufficient to achieve interaction promiscuity in these toxins, allowing the antitoxins to follow them during sequence divergence (Aakre et al., 2015). Similarly, the transition from one selective pair of the SOS-induced colicin DNA-nuclease and its cognate inhibitor to another functionally insulated pair proceeded in the laboratory through inhibitors with interaction promiscuity (Levin et al., 2009).

Our design of the PabA* variant demonstrates that relaxed interaction specificity can easily be reached not only in toxin-antitoxin or DNA-nuclease-inhibitor complexes but also in essential metabolic enzyme complexes. These findings are important for an understanding how prevalent such promiscuous evolutionary intermediates are and whether they can be easily reached from specific starting complexes (Aakre et al., 2015). Moreover, the PabA* glutaminase is, to our knowledge, the first example of a protein with interaction promiscuity that has been generated by rational protein design. Last but not least, this promiscuous glutaminase demonstrates that the concept of PPI evolution through promiscuous intermediates is not only applicable to protein complexes whose interaction specificity is determined by individual amino acids in the interface but also to complexes whose specificity is defined by large changes in interface geometry caused by interface add-ons.

5 Abbreviations

2,3-DHB	2,3-dihydroxybenzoic acid
AA	anthranilate
ADC	aminodeoxychorismate
ADCS	aminodeoxychorismate synthase
ADIC	aminodeoxyisochorismate
ADICS	aminodeoxyisochorismate synthase
AME	ammonia-utilizing MST-enzyme
AS	anthranilate synthase
ASR	ancestral sequence reconstruction
B. lata	*Burkholderia lata*
B. subtilis	*Bacillus subtilis*
BGC	biosynthetic gene cluster
CdRP	1-(o-carboxyphenylamino)-1-deoxyribulose-5-phosphate
CH	chorismate
CL	chorismate lyase
CM	chorismate mutase
CUE	chorismate-utilizing enzyme
D. desulfuricans	*Desulfovibrio desulfuricans*

DNA	deoxyribonucleic acid
E. coli	*Escherichia coli*
GATase	glutamine amidotransferase
HisA	1-(5-phosphoribosyl)-5-[(5-phosphoribosylamino)-methylideneamino]-imidazole-4-carboxamide isomerase
HisF	cyclase subunit of imidazole glycerol phosphate synthase
IC	isochorismate
ICS	isochorismate synthase
LUCA	last universal common ancestor
M. tuberculosis	*Mycobacterium tuberculosis*
MSA	multiple sequence alignment
P. carbinolicus	*Pelobacter carbinolicus*
P. putida	*Pseudomonas putida*
PA	prephenate
PHB	*para*-hydroxybenzoate
PPI	protein-protein interaction
PRA	N-(5'-phosphoribosyl)anthranilate
PRFAR	N'-[(5'-phosphoribulosyl)-formimino]-5-aminoimidazole-4-carboxamide-ribonucleotide
PriA	phosphoribosyl isomerase A
ProFAR	N'-[(5'-phosphoribosyl)-formimino]-5-aminoimidazole-4-carboxamide-ribonucleotide
RMSD	root-mean-square deviation
RNA	ribonucleic acid

S. aureus	*Staphylococcus aureus*
S. cerevisiae	*Saccharomyces cerevisiae*
S. coelicolor	*Streptomyces coelicolor*
S. enterica	*Salmonella enterica*
S. marcescens	*Serratia marcescens*
S. solfataricus	*Sulfolobus solfataricus*
S. typhimurium	*Salmonella typhimurium*
SA	salicylate
SS	salicylate synthase
SSN	sequence similarity network
T. maritima	*Thermotoga maritima*
TrpF	phosphoribosylanthranilate isomerase
WME	water-utilizing MST-enzyme
Y. pestis	*Yersinia pestis*

6 References

Aakre, C. D., Herrou, J., Phung, T. N., Perchuk, B. S., Crosson, S., and Laub, M. T. (2015). Evolving new protein-protein interaction specificity through promiscuous intermediates. *Cell*, 163:594–606.

Ahnert, S. E., Marsh, J. A., Hernández, H., Robinson, C. V., and Teichmann, S. A. (2015). Principles of assembly reveal a periodic table of protein complexes. *Science*, 350:aaa2245.

Akanuma, S., Nakajima, Y., Yokobori, S., Kimura, M., Nemoto, N., Mase, T., Miyazono, K., Tanokura, M., and Yamagishi, A. (2013). Experimental evidence for the thermophilicity of ancestral life. *Proceedings of the National Academy of Sciences*, 110:11067–11072.

Almonacid, D. E. and Babbitt, P. C. (2011). Toward mechanistic classification of enzyme functions. *Current Opinion in Chemical Biology*, 15:435–442.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215:403–410.

Alva, V., Söding, J., and Lupas, A. N. (2015). A vocabulary of ancient peptides at the origin of folded proteins. *Elife*, 4:e09410.

Ames, R. M., Talavera, D., Williams, S. G., Robertson, D. L., and Lovell, S. C. (2016). Binding interface change and cryptic variation in the evolution of protein-protein interactions. *BMC Evolutionary Biology*, 16:1.

Anand, R., Hoskins, A. A., Stubbe, J., and Ealick, S. E. (2004). Domain organization of *Salmonella typhimurium* formylglycinamide ribonucleotide amidotransferase revealed by x-ray crystallography. *Biochemistry*, 43:10328–10342.

Andreeva, A., Howorth, D., Chandonia, J.-M., Brenner, S. E., Hubbard, T. J., Chothia, C., and Murzin, A. G. (2008). Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Research*, 36:419–425.

Araghi, R. R. and Keating, A. E. (2016). Designing helical peptide inhibitors of protein–protein interactions. *Current Opinion in Structural Biology*, 39:27–38.

Atkinson, H. J., Morris, J. H., Ferrin, T. E., and Babbitt, P. C. (2009). Using sequence similarity networks for visualization of relationships across diverse protein superfamilies. *PLoS One*, 4:e4345.

August, P. R., Tang, L., Yoon, Y. J., Ning, S., Müller, R., Yu, T.-W., Taylor, M., Hoffmann, D., Kim, C.-G., Zhang, X., Hutchinson, C. R., and Floss, H. G. (1998). Biosynthesis of the ansamycin antibiotic rifamycin: deductions from the molecular analysis of the *rif* biosynthetic gene cluster of *Amycolatopsis mediterranei* S699. *Chemistry & Biology*, 5:69–79.

Babitzke, P. (1997). Regulation of tryptophan biosynthesis: Trp-ing the trap or how *Bacillus subtilis* reinvented the wheel. *Molecular Microbiology*, 26:1–9.

Bada, J. L. (2013). New insights into prebiotic chemistry from Stanley Miller's spark discharge experiments. *Chemical Society Reviews*, 42:2186–2196.

Baier, F., Copp, J. N., and Tokuriki, N. (2016). Evolution of enzyme superfamilies: Comprehensive exploration of sequence–function relationships. *Biochemistry*, 55:6375–6388.

Banci, L., Benedetto, M., Bertini, I., Del Conte, R., Piccioli, M., and Viezzoli, M. S. (1998). Solution structure of reduced monomeric Q133M2 copper, zinc superoxide dismutase (SOD). Why is SOD a dimeric enzyme? *Biochemistry*, 37:11780–11791.

Bar-Even, A. and Tawfik, D. S. (2013). Engineering specialized metabolic pathways—is there a room for enzyme improvements? *Current Opinion in Biotechnology*, 24:310–319.

Barlow, M. and Hall, B. G. (2002). Phylogenetic analysis shows that the OXA β-lactamase genes have been on plasmids for millions of years. *Journal of Molecular Evolution*, 55:314–321.

Barona-Gómez, F. and Hodgson, D. A. (2003). Occurrence of a putative ancient-like isomerase involved in histidine and tryptophan biosynthesis. *EMBO Reports*, 4:296–300.

Battistuzzi, F. and Hedges, S. (2009). Eubacteria. In *The Timetree of Life*. Oxford University Press, New York.

Beismann-Driemeyer, S. and Sterner, R. (2001). Imidazole glycerol phosphate synthase from *Thermotoga maritima*. Quaternary structure, steady-state kinetics, and reaction mechanism of the bienzyme complex. *Journal of Biological Chemistry*, 276:20387–20396.

Bera, A. K., Atanasova, V., Dhanda, A., Ladner, J. E., and Parsons, J. F. (2012). Structure of aminodeoxychorismate synthase from *Stenotrophomonas maltophilia*. *Biochemistry*, 51:10208–10217.

Bera, A. K., Chen, S., Smith, J. L., and Zalkin, H. (1999). Interdomain signaling in glutamine phosphoribosylpyrophosphate amidotransferase. *Journal of Biological Chemistry*, 274:36498–36504.

Bergthorsson, U., Andersson, D. I., and Roth, J. R. (2007). Ohno's dilemma: evolution of new genes under continuous selection. *Proceedings of the National Academy of Sciences*, 104:17004–17009.

Berman, J. J. (2012). *Taxonomic Guide to Infectious Diseases: Understanding the Biologic Classes of Pathogenic Organisms.* Academic Press, New York.

Bernhardt, H. S. (2012). The RNA world hypothesis: the worst theory of the early evolution of life (except for all the others). *Biology Direct*, 7:1–10.

Binda, C., Bossi, R. T., Wakatsuki, S., Arzt, S., Coda, A., Curti, B., Vanoni, M. A., and Mattevi, A. (2000). Cross-talk and ammonia channeling between active centers in the unexpected domain arrangement of glutamate synthase. *Structure*, 8:1299–1308.

Blanc, V., Gil, P., Bamas-Jacques, N., Lorenzon, S., Zagorec, M., Schleuniger, J., Bisch, D., Blanche, F., Debussche, L., Crouzet, J., and Thibaut, D. (1997). Identification and analysis of genes from *Streptomyces pristinaespiralis* encoding enzymes involved in the biosynthesis of the 4-dimethylamino-l-phenylalanine precursor of pristinamycin I. *Molecular Microbiology*, 23:191–202.

Blankenfeldt, W. (2013). The biosynthesis of phenazines. In *Microbial Phenazines*. Springer, Berlin, Heidelberg.

Blattner, F. R., Plunkett, G., Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K., Mayhew, G. F., Gregor, J., Davis, N. W., Kikrpatrick, H. A., Goeden, M. A., Rose, D. J., Mau, B., and Shao, Y. (1997). The complete genome sequence of *Escherichia coli* K-12. *Science*, 277:1453–1462.

Bogan, A. A. and Thorn, K. S. (1998). Anatomy of hot spots in protein interfaces. *Journal of Molecular Biology*, 280:1–9.

Bornscheuer, U. T. and Kazlauskas, R. J. (2004). Catalytic promiscuity in biocatalysis: using old enzymes to form new bonds and follow new pathways. *Angewandte Chemie International Edition*, 43:6032–6040.

Borrero, N. V., Bai, F., Perez, C., Duong, B. Q., Rocca, J. R., Jin, S., and Huigens III, R. W. (2014). Phenazine antibiotic inspired discovery of potent bromophenazine antibacterial agents against *Staphylococcus aureus* and *Staphylococcus epidermidis*. *Organic & Biomolecular Chemistry*, 12:881–886.

Boundy-Mills, K. L., de Souza, M., Mandelbaum, R. T., Wackett, L. P., and Sadowsky, M. J. (1997). The atzB gene of *Pseudomonas sp.* strain ADP encodes the second enzyme of a novel atrazine degradation pathway. *Applied and Environmental Microbiology*, 63:916–923.

Bouvier, B., Grünberg, R., Nilges, M., and Cazals, F. (2009). Shelling the Voronoi interface of protein–protein complexes reveals patterns of residue conservation, dynamics, and composition. *Proteins: structure, Function, and Bioinformatics*, 76:677–692.

Boyanova, L., Kolarov, R., and Mitov, I. (2015). Recent evolution of antibiotic resistance in the anaerobes as compared to previous decades. *Anaerobe*, 31:4–10.

Bridgham, J. T., Carroll, S. M., and Thornton, J. W. (2006). Evolution of hormone-receptor complexity by molecular exploitation. *Science*, 312:97–101.

Brown, M. P., Aidoo, K. A., and Vining, L. C. (1996). A role for *pab*ab, a *p*-aminobenzoate synthase gene of *Streptomyces venezuelae* ISP5230 in chloramphenicol biosynthesis. *Microbiology*, 142:1345–1355.

Brown, S. D. and Babbitt, P. C. (2014). New insights about enzyme evolution from large scale studies of sequence and structure relationships. *Journal of Biological Chemistry*, 289:30221–30228.

Brückner, A., Polge, C., Lentze, N., Auerbach, D., and Schlattner, U. (2009). Yeast two-hybrid, a powerful tool for systems biology. *International Journal of Molecular Sciences*, 10:2763–2788.

Busch, F., Rajendran, C., Heyn, K., Schlee, S., Merkl, R., and Sterner, R. (2016). Ancestral tryptophan synthase reveals functional sophistication of primordial enzyme complexes. *Cell Chemical Biology*, 23:709–715.

Caffrey, D. R., Somaroo, S., Hughes, J. D., Mintseris, J., and Huang, E. S. (2004). Are protein–protein interfaces more conserved in sequence than the rest of the protein surface? *Protein Science*, 13:190–202.

Caligiuri, M. and Bauerle, R. (1991). Identification of amino acid residues involved in feedback regulation of the anthranilate synthase complex from *Salmonella typhimurium*. evidence for an amino-terminal regulatory site. *Journal of Biological Chemistry*, 266:8328–8335.

Capra, E. J., Perchuk, B. S., Skerker, J. M., and Laub, M. T. (2012). Adaptive mutations that prevent crosstalk enable the expansion of paralogous signaling protein families. *Cell*, 150:222–232.

Carpenter, E. P., Hawkins, A. R., Frost, J. W., and Brown, K. A. (1998). Structure of dehydroquinate synthase reveals an active site capable of multistep catalysis. *Nature*, 394:299–302.

Carroll, S. M., Ortlund, E. A., and Thornton, J. W. (2011). Mechanisms for the evolution of a derived function in the ancestral glucocorticoid receptor. *PLoS Genetics*, 7:e1002117.

Chae, L., Kim, T., Nilo-Poyanco, R., and Rhee, S. Y. (2014). Genomic signatures of specialized metabolism in plants. *Science*, 344:510–513.

Chakrabarti, P. and Janin, J. (2002). Dissecting protein–protein recognition sites. *Proteins: Structure, Function, and Bioinformatics*, 47:334–343.

Chandrasekharan, U. M., Sanker, S., Glynias, M. J., Karnik, S. S., and Husain, A. (1996). Angiotensin II-forming activity in a reconstructed ancestral chymase. *Science*, 271:502.

Chang, B. S., Jönsson, K., Kazmi, M. A., Donoghue, M. J., and Sakmar, T. P. (2002). Recreating a functional ancestral archosaur visual pigment. *Molecular Biology and Evolution*, 19:1483–1489.

Chatr-Aryamontri, A., Breitkreutz, B.-J., Oughtred, R., Boucher, L., Heinicke, S., Chen, D., Stark, C., Breitkreutz, A., Kolas, N., O'Donnell, L., Reguly, T., Nixon, J., Ramage, L., Winter, A., Sellam, A., C, C., Hirschman, J., Theesfeld, C., Rust, J., Livstone, M. S., Dolinski, K., and Tyers, M. (2015). The BioGRID interaction database: 2015 update. *Nucleic Acids Research*, 43:470–478.

Chen, P. and Li, J. (2010). Sequence-based identification of interface residues by an integrative profile combining hydrophobic and evolutionary information. *BMC Bioinformatics*, 11(402).

Chook, Y. M., Gray, J. V., Ke, H., and Lipscomb, W. N. (1994). The monofunctional chorismate mutase from *Bacillus subtilis*: structure determination of chorismate mutase and its complexes with a transition state analog and prephenate, and implications for the mechanism of the enzymatic reaction. *Journal of Molecular Biology*, 240:476–500.

Chothia, C. (1992). Proteins. one thousand families for the molecular biologist. *Nature*, 357:543.

Chow, J.-Y., Tian, B.-X., Ramamoorthy, G., Hillerich, B. S., Seidel, R. D., Almo, S. C., Jacobson, M. P., and Poulter, C. D. (2015). Computational-guided discovery and characterization of a sesquiterpene synthase from *Streptomyces clavuligerus*. *Proceedings of the National Academy of Sciences*, 112:5661–5666.

Ciccarelli, F. D., Doerks, T., Von Mering, C., Creevey, C. J., Snel, B., and Bork, P. (2006). Toward automatic reconstruction of a highly resolved tree of life. *Science*, 311:1283–1287.

Claren, J., Malisi, C., Höcker, B., and Sterner, R. (2009). Establishing wild-type levels of catalytic activity on natural and artificial $(\beta\alpha)_8$-barrel protein scaffolds. *Proceedings of the National Academy of Sciences*, 106:3704–3709.

Conant, G. C. and Wagner, A. (2003). Asymmetric sequence divergence of duplicate genes. *Genome Research*, 13:2052–2058.

Conte, L. L., Chothia, C., and Janin, J. (1999). The atomic structure of protein-protein recognition sites. *Journal of Molecular Biology*, 285:2177–2198.

Copley, S. D. (2003). Enzymes with extra talents: moonlighting functions and catalytic promiscuity. *Current Opinion in Chemical Biology*, 7:265–272.

Copley, S. D. (2009). Evolution of efficient pathways for degradation of anthropogenic chemicals. *Nature Chemical Biology*, 5:559–566.

Copley, S. D. (2015). An evolutionary biochemist's perspective on promiscuity. *Trends in Biochemical Sciences*, 40:72–78.

Corea, O. R., Ki, C., Cardenas, C. L., Kim, S.-J., Brewer, S. E., Patten, A. M., Davin, L. B., and Lewis, N. G. (2012). Arogenate dehydratase isoenzymes profoundly and differentially modulate carbon flux into lignins. *Journal of Biological Chemistry*, 287:11446–11459.

Cornelis, P. and Matthijs, S. (2007). Pseudomonas siderophores and their biological significance. In *Microbial siderophores*. Springer, Berlin, Heidelberg.

Criado, L. M., Martín, J. F., and Gil, J. (1993). The *pab* gene of *Streptomyces griseus*, encoding *p*-aminobenzoic acid synthase, is located between genes possibly involved in candicidin biosynthesis. *Gene*, 126:135–139.

Crosa, J. H. (1989). Genetics and molecular biology of siderophore-mediated iron transport in bacteria. *Microbiological Reviews*, 53:517–530.

Crosa, J. H. (1997). Signal transduction and transcriptional and posttranscriptional control of iron-regulated genes in bacteria. *Microbiology and Molecular Biology Reviews*, 61:319–336.

Culbertson, J. E., Chung, D. h., Ziebart, K. T., Espiritu, E., and Toney, M. D. (2015). Conversion of aminodeoxychorismate synthase into anthranilate synthase with janus mutations: Mechanism of pyruvate elimination catalyzed by chorismate enzymes. *Biochemistry*, 54:2372–2384.

Czaplewski, L., Bax, R., Clokie, M., Dawson, M., Fairhead, H., Fischetti, V. A., Foster, S., Gilmore, B. F., Hancock, R. E., Harper, D., et al. (2016). Alternatives to antibiotics—a pipeline portfolio review. *The Lancet Infectious Diseases*, 16:239–251.

Dahm, C., Müller, R., Schulte, G., Schmidt, K., and Leistner, E. (1998). The role of isochorismate hydroxymutase genes *ent*C and *men*F in enterobactin and menaquinone biosynthesis in *Escherichia coli*. *Biochimica et Biophysica Acta - General Subjects*, 1425:377–386.

Dartnell, L., Simeonidis, E., Hubank, M., Tsoka, S., Bogle, I. D. L., and Papageorgiou, L. G. (2005). Robustness of the p53 network and biological hackers. *FEBS Letters*, 579:3037–3042.

Darwin, C. (1859). The origin of species. In *The Harvard Classics. Vol 11*. P.F. Collier & Son, New York.

Dayhoff, M. O. (1965). *Atlas of protein sequence and structure. Vol 5*. National Biomedical Research Foundation, Washington.

de Beer, T. A., Berka, K., Thornton, J. M., and Laskowski, R. A. (2014). PDBsum additions. *Nucleic Acids Research*, 42:292–296.

de Farias, S. T., Rêgo, T. G., and José, M. V. (2016). A proposal of the proteome before the last universal common ancestor (LUCA). *International Journal of Astrobiology*, 15:27–31.

de Souza, M. L., Seffernick, J., Martinez, B., Sadowsky, M. J., and Wackett, L. P. (1998a). The atrazine catabolism genes *atz*ABC are widespread and highly conserved. *Journal of Bacteriology*, 180:1951–1954.

de Souza, M. L., Wackett, L. P., and Sadowsky, M. J. (1998b). The *atz*ABC genes encoding atrazine catabolism are located on a self-transmissible plasmid in *Pseudomonas sp.* strain ADP. *Applied and Environmental Microbiology*, 64:2323–2326.

De Voss, J. J., Rutter, K., Schroeder, B. G., Su, H., Zhu, Y., and Barry, C. E. (2000). The salicylate-derived mycobactin siderophores of *Mycobacterium tuberculosis* are essential for growth in macrophages. *Proceedings of the National Academy of Sciences*, 97:1252–1257.

Delmas, J., Chen, Y., Prati, F., Robin, F., Shoichet, B. K., and Bonnet, R. (2008). Structure and dynamics of CTX-M enzymes reveal insights into substrate accommodation by extended-spectrum β-lactamases. *Journal of Molecular Biology*, 375:192–201.

Demain, A. L. and Fang, A. (2000). The natural functions of secondary metabolites. In *History of Modern Biotechnology I*. Springer, Berlin, Heidelberg.

DePristo, M. A., Weinreich, D. M., and Hartl, D. L. (2005). Missense meanderings in sequence space: a biophysical view of protein evolution. *Nature Reviews Genetics*, 6:678–687.

Dhiman, R. K., Mahapatra, S., Slayden, R. A., Boyne, M. E., Lenaerts, A., Hinshaw, J. C., Angala, S. K., Chatterjee, D., Biswas, K., Narayanasamy, P., Kurosu, M., and Crick, D. C. (2009). Menaquinone synthesis is critical for maintaining mycobacterial viability during exponential growth and recovery from non-replicating persistence. *Molecular Microbiology*, 72:85–97.

Di Giulio, M. (2011). The last universal common ancestor (LUCA) and the ancestors of archaea and bacteria were progenotes. *Journal of Molecular Evolution*, 72:119–126.

DiNicolantonio, J. J., Bhutani, J., and O'Keefe, J. H. (2015). The health benefits of vitamin K. *Open Heart*, 2:e000300.

Dobson, R. C., Valegård, K., and Gerrard, J. A. (2004). The crystal structure of three site-directed mutants of *Escherichia coli* dihydrodipicolinate synthase: further evidence for a catalytic triad. *Journal of Molecular Biology*, 338:329–339.

Domagalski, M., Tkaczuk, K., Chruszcz, M., Skarina, T., Onopriyenko, O., Cymborowski, M., Grabowski, M., Savchenko, A., and Minor, W. (2013). Structure of isochorismate synthase DhbC from *Bacillus anthracis*. *Acta Crystallographica Section F: Structural Biology and Crystallization Communications*, 69:956–961.

Dong, Z., Wang, K., Dang, T. K. L., Gültas, M., Welter, M., Wierschin, T., Stanke, M., and Waack, S. (2014). CRF-based models of protein surfaces improve protein-protein interaction site predictions. *BMC Bioinformatics*, 15:277.

Doolittle, W. F. (2000). The nature of the universal ancestor and the evolution of the proteome. *Current Opinion in Structural Biology*, 10:355–358.

Dosselaere, F. and Vanderleyden, J. (2001). A metabolic node in action: chorismate-utilizing enzymes in microorganisms. *Critical Reviews in Microbiology*, 27:75–131.

Douangamath, A., Walker, M., Beismann-Driemeyer, S., Vega-Fernandez, M. C., Sterner, R., and Wilmanns, M. (2002). Structural evidence for ammonia tunneling across the $(\beta\alpha)_8$ barrel of the imidazole glycerol phosphate synthase bienzyme complex. *Structure*, 10:185–193.

Due, A. V., Küper, J., Geerlof, A., von Kries, J. P., and Wilmanns, M. (2011). Bisubstrate specificity in histidine/tryptophan biosynthesis isomerase from *Mycobacterium tuberculosis* by active site metamorphosis. *Proceedings of the National Academy of Sciences*, 108:3554–3559.

Duke, S. O. and Powles, S. B. (2008). Glyphosate: a once-in-a-century herbicide. *Pest Management Science*, 64:319–325.

D'Argenio, V. and Salvatore, F. (2015). The role of the gut microbiome in the healthy adult status. *Clinica Chimica Acta*, 451:97–102.

Eick, G. N., Colucci, J. K., Harms, M. J., Ortlund, E. A., and Thornton, J. W. (2012). Evolution of minimal specificity and promiscuity in steroid hormone receptors. *PLoS Genetics*, 8:e1003072.

Eisenberg, D., Schwarz, E., Komarony, M., and Wall, R. (1984). Amino acid scale: Normalized consensus hydrophobicity scale. *Journal of Molecular Biology*, 179:125–142.

Elcock, A. H. (2010). Models of macromolecular crowding effects and the need for quantitative comparisons with experiment. *Current Opinion in Structural Biology*, 20:196–206.

Engel, S., Jensen, P. R., and Fenical, W. (2002). Chemical ecology of marine microbial defense. *Journal of Chemical Ecology*, 28:1971–1985.

Esmaielbeiki, R., Krawczyk, K., Knapp, B., Nebel, J.-C., and Deane, C. M. (2016). Progress and challenges in predicting protein interfaces. *Briefings in Bioinformatics*, 17:117–131.

Facchini, P. J. (2001). Alkaloid biosynthesis in plants: biochemistry, cell biology, molecular regulation, and metabolic engineering applications. *Annual Review of Plant Biology*, 52:29–66.

Farías-Rico, J. A., Schmidt, S., and Höcker, B. (2014). Evolutionary relationship of two ancient protein superfolds. *Nature Chemical Biology*, 10:710–715.

Fiebig, A., Castro Rojas, C. M., Siegal-Gaskins, D., and Crosson, S. (2010). Interaction specificity, toxicity and regulation of a paralogous set of ParE/RelE-family toxin–antitoxin systems. *Molecular Microbiology*, 77:236–251.

Field, S. F. and Matz, M. V. (2010). Retracing evolution of red fluorescence in GFP-like proteins from Faviina corals. *Molecular Biology and Evolution*, 27:225–233.

Fields, S. and Sternglanz, R. (1994). The two-hybrid system: an assay for protein-protein interactions. *Trends in Genetics*, 10:286–292.

Finkelstein, A. V. and Ptitsyn, O. B. (1987). Why do globular proteins fit the limited set of folding patterns? *Progress in Biophysics and Molecular Biology*, 50:171–190.

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., Salazar, G. A., Tate, J., and Bateman, A. (2016). The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Research*, 44:279–285.

Finnigan, G. C., Hanson-Smith, V., Stevens, T. H., and Thornton, J. W. (2012). Evolution of increased complexity in a molecular machine. *Nature*, 481:360–364.

Firn, R. D. and Jones, C. G. (2000). The evolution of secondary metabolism – a unifying model. *Molecular Microbiology*, 37:989–994.

Firn, R. D. and Jones, C. G. (2009). A darwinian view of metabolism: molecular properties determine fitness. *Journal of Experimental Botany*, 60:719–726.

Fischbach, M. A. and Clardy, J. (2007). One pathway, many products. *Nature Chemical Biology*, 3:353–355.

Fischbach, M. A. and Walsh, C. T. (2006). Assembly-line enzymology for polyketide and nonribosomal peptide antibiotics: logic, machinery, and mechanisms. *Chemical Reviews*, 106:3468–3496.

Fischbach, M. A., Walsh, C. T., and Clardy, J. (2008). The evolution of gene collectives: How natural selection drives chemical innovation. *Proceedings of the National Academy of Sciences*, 105:4601–4608.

Fitch, W. M. (1970). Distinguishing homologous from analogous proteins. *Systematic Biology*, 19:99–113.

Fournier, G. and Alm, E. (2015). Ancestral reconstruction of a pre-LUCA aminoacyl-tRNA synthetase ancestor supports the late addition of Trp to the genetic code. *Journal of Molecular Evolution*, 80:171–185.

Fowler, D. M. and Fields, S. (2014). Deep mutational scanning: a new style of protein science. *Nature Methods*, 11:801–807.

Franke, J., Ishida, K., and Hertweck, C. (2014). Evolution of siderophore pathways in human pathogenic bacteria. *Journal of the American Chemical Society*, 136:5599–5602.

Fraser, H. B., Hirsh, A. E., Steinmetz, L. M., Scharfe, C., and Feldman, M. W. (2002). Evolutionary rate in the protein interaction network. *Science*, 296:750–752.

Freeling, M. and Thomas, B. C. (2006). Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. *Genome Research*, 16:805–814.

Friedberg, E. C. (2003). DNA damage and repair. *Nature*, 421:436–440.

Friedman, M. J. and Trager, W. (1981). The biochemistry of resistance to malaria. *Scientific American*, 244:154–155.

Gallagher, D. T., Mayhew, M., Holden, M. J., Howard, A., Kim, K.-J., and Vilker, V. L. (2001). The crystal structure of chorismate lyase shows a new fold and a tightly retained product. *Proteins: Structure, Function, and Bioinformatics*, 44:304–311.

Gao, M. and Skolnick, J. (2010). Structural space of protein–protein interfaces is degenerate, close to complete, and highly connected. *Proceedings of the National Academy of Sciences*, 107:22517–22522.

Garma, L., Mukherjee, S., Mitra, P., and Zhang, Y. (2012). How many protein-protein interactions types exist in nature? *PLoS One*, 7:e38913.

Gaucher, E. A., Govindarajan, S., and Ganesh, O. K. (2008). Palaeotemperature trend for precambrian life inferred from resurrected proteins. *Nature*, 451:704–707.

Gavin, A.-C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., Rau, C., Jensen, L. J., Bastuck, S., Dümpelfeld, B., et al. (2006). Proteome survey reveals modularity of the yeast cell machinery. *Nature*, 440:631–636.

Gerlt, J. A., Bouvier, J. T., Davidson, D. B., Imker, H. J., Sadkhin, B., Slater, D. R., and Whalen, K. L. (2015). Enzyme function initiative-enzyme similarity tool (EFI-EST): A web tool for generating protein sequence similarity networks. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, 1854:1019–1037.

Gilbert, W. (1986). Origin of life: The RNA world. *Nature*, 319:618.

Giroux, E., Williams, M. K., and Kantrowitz, E. R. (1994). Shared active sites of fructose-1,6-bisphosphatase. Arginine 243 mediates substrate binding and fructose 2,6-bisphosphate inhibition. *Journal of Biological Chemistry*, 269:31404–31409.

Glaser, F., Steinberg, D. M., Vakser, I. A., and Ben-Tal, N. (2001). Residue frequencies and pairing preferences at protein–protein interfaces. *Proteins: Structure, Function, and Bioinformatics*, 43:89–102.

Goerisch, H. and Lingens, F. (1974). Chorismate mutase from *Streptomyces*. Purification, properties, and subunit structure of the enzyme from *Streptomyces aureofaciens* Tü 24. *Biochemistry*, 13:3790–3794.

Goffeau, A., Barrell, B. G., Bussey, H., Davis, R., Dujon, B., Feldmann, H., Galibert, F., Hoheisl, J. D., Jacq, C., Johnston, M., Louis, E. J., Mewes, H. W., Murakami, Y., Philippsen, P., Tettelin, H., and Oliver, S. G. (1996). Life with 6000 genes. *Science*, 274:546.

Golden, E., Paterson, R., Tie, W. J., Anandan, A., Flematti, G., Molla, G., Rosini, E., Pollegioni, L., and Vrielink, A. (2013). Structure of a class III engineered cephalosporin acylase: comparisons with class I acylase and implications for differences in substrate specificity and catalytic activity. *Biochemical Journal*, 451:217–226.

Goto, M., Omi, R., Nakagawa, N., Miyahara, I., and Hirotsu, K. (2004). Crystal structures of CTP synthetase reveal ATP, UTP, and glutamine binding sites. *Structure*, 12:1413–1423.

Goto, Y., Zalkin, H., Keim, P., and Heinrikson, R. (1976). Properties of anthranilate synthetase component II from *Pseudomonas putida*. *Journal of Biological Chemistry*, 251:941–949.

Gould, S. J. and Vrba, E. S. (1982). Exaptation—a missing term in the science of form. *Paleobiology*, 8:4–15.

Gray, J. V., Golinelli-Pimpaneau, B., and Knowles, J. R. (1990). Monofunctional chorismate mutase from *Bacillus subtilis*: purification of the protein, molecular cloning of the gene, and overexpression of the gene product in *Escherichia coli*. *Biochemistry*, 29:376–383.

Grossmann, A., Benlasfer, N., Birth, P., Hegele, A., Wachsmuth, F., Apelt, L., and Stelzl, U. (2015). Phospho-tyrosine dependent protein–protein interaction network. *Molecular Systems Biology*, 11:794.

Guharoy, M. and Chakrabarti, P. (2005). Conservation and relative importance of residues across protein-protein interfaces. *Proceedings of the National Academy of Sciences*, 102:15447–15452.

Guzzetti, D., Lebrun, A., Subileau, M., Grousseau, E., Dubreucq, E., and Drone, J. (2016). A catalytically competent terpene synthase inferred using ancestral sequence reconstruction strategy. *ACS Catalysis*, 6:5345–5349.

Hakem, R. (2008). DNA-damage repair; the good, the bad, and the ugly. *The EMBO Journal*, 27:589–605.

Hall, B. G. (2004). Predicting the evolution of antibiotic resistance genes. *Nature Reviews Microbiology*, 2:430–435.

Hall, D. A., Ptacek, J., and Snyder, M. (2007). Protein microarray technology. *Mechanisms of Ageing and Development*, 128:161–167.

Hallez, R., Geeraerts, D., Sterckx, Y., Mine, N., Loris, R., and Van Melderen, L. (2010). New toxins homologous to ParE belonging to three-component toxin-antitoxin systems in *Escherichia coli* O157:H7. *Molecular Microbiology*, 76:719–732.

Hamer, R., Luo, Q., Armitage, J. P., Reinert, G., and Deane, C. M. (2010). i-patch: Interprotein contact prediction using local network information. *Proteins: Structure, Function, and Bioinformatics*, 78:2781–2797.

Harms, M. J. and Thornton, J. W. (2013). Evolutionary biochemistry: revealing the historical and physical causes of protein properties. *Nature Reviews Genetics*, 14:559–571.

Harris, W. R., Carrano, C. J., Cooper, S. R., Sofen, S. R., Avdeef, A. E., McArdle, J. V., and Raymond, K. N. (1979). Coordination chemistry of microbial iron transport compounds. 19. Stability constants and electrochemical behavior of ferric enterobactin and model complexes. *Journal of the American Chemical Society*, 101:6097–6104.

Hart, G. T., Ramani, A. K., and Marcotte, E. M. (2006). How complete are current yeast and human protein-interaction networks? *Genome Biology*, 7:120.

Hart, K. M., Harms, M. J., Schmidt, B. H., Elya, C., Thornton, J. W., and Marqusee, S. (2014). Thermodynamic system drift in protein evolution. *PLoS Biology*, 12:e1001994.

Hartley, C., Newcomb, R., Russell, R., Yong, C., Stevens, J., Yeates, D., La Salle, J., and Oakeshott, J. (2006). Amplification of dna from preserved specimens shows blowflies were preadapted for the rapid evolution of insecticide resistance. *Proceedings of the National Academy of Sciences*, 103:8757–8762.

Hartmann, T. (2007). From waste products to ecochemicals: fifty years research of plant secondary metabolism. *Phytochemistry*, 68:2831–2846.

Hashimoto, K. and Panchenko, A. R. (2010). Mechanisms of protein oligomerization, the critical role of insertions and deletions in maintaining different oligomeric states. *Proceedings of the National Academy of Sciences*, 107:20352–20357.

Haslam, E. (2014). *The Shikimate Pathway: Biosynthesis of Natural Products Series.* Elsevier, New York.

Hayes, S., Malacrida, B., Kiely, M., and Kiely, P. A. (2016). Studying protein–protein interactions: progress, pitfalls and solutions. *Biochemical Society Transactions*, 44:994–1004.

He, Z., Stigers Lavoie, K. D., Bartlett, P. A., and Toney, M. D. (2004). Conservation of mechanism in three chorismate-utilizing enzymes. *Journal of the American Chemical Society*, 126:2378–2385.

Hedden, P., Phillips, A. L., Rojas, M. C., Carrera, E., and Tudzynski, B. (2001). Gibberellin biosynthesis in plants and fungi: a case of convergent evolution? *Journal of Plant Growth Regulation*, 20:319–331.

Hegyi, H. and Gerstein, M. (2001). Annotation transfer for genomics: measuring functional divergence in multi-domain proteins. *Genome Research*, 11:1632–1640.

Heitman, J., Movva, N. R., Hiestand, P. C., and Hall, M. N. (1991). FK506-binding protein proline rotamase is a target for the immunosuppressive agent FK506 in *Saccharomyces cerevisiae*. *Proceedings of the National Academy of Sciences*, 88:1948–1952.

Hendrickson, L., Davis, C. R., Roach, C., Nguyen, D. K., Aldrich, T., McAda, P. C., and Reeves, C. D. (1999). Lovastatin biosynthesis in *Aspergillus terreus*: characterization of blocked mutants, enzyme activities and a multifunctional polyketide synthase gene. *Chemistry & Biology*, 6:429–439.

Henn-Sax, M., Thoma, R., Schmidt, S., Hennig, M., Kirschner, K., and Sterner, R. (2002). Two $(\beta\alpha)_8$-barrel enzymes of histidine and tryptophan biosynthesis have similar reaction mechanisms and common strategies for protecting their labile substrates. *Biochemistry*, 41:12032–12042.

Higgs, P. G. and Lehman, N. (2015). The RNA world: molecular cooperation at the origins of life. *Nature Reviews Genetics*, 16:7–17.

Hinkley, T., Martins, J., Chappey, C., Haddad, M., Stawiski, E., Whitcomb, J. M., Petropoulos, C. J., and Bonhoeffer, S. (2011). A systems analysis of mutational effects in HIV-1 protease and reverse transcriptase. *Nature Genetics*, 43:487–489.

Hodgson, D. A. (2000). Primary metabolism and its control in streptomycetes: a most unusual group of bacteria. *Advances in Microbial Physiology*, 42:47–238.

Holinski, A., Heyn, K., Merkl, R., and Sterner, R. (2017). Combining ancestral sequence reconstruction with protein design to identify an interface hotspot in a key metabolic enzyme complex. *Proteins: Structure, Function, and Bioinformatics*, in press.

Hommel, U., Eberhard, M., and Kirschner, K. (1995). Phosphoribosyl anthranilate isomerase catalyzes a reversible Amadori reaction. *Biochemistry*, 34:5429–5439.

Hooper, S. D. and Berg, O. G. (2003). On the nature of gene innovation: duplication patterns in microbial genomes. *Molecular Biology and Evolution*, 20:945–954.

Hoyle, F. and Wickramasinghe, N. C. (2000). Biological evolution. In *Astronomical Origins of Life*. Springer, Berlin, Heidelberg.

Huang, H., Jedynak, B. M., and Bader, J. S. (2007). Where have all the interactions gone? Estimating the coverage of two-hybrid protein interaction maps. *PLoS Computational Biology*, 3:e214.

Huang, X., Holden, H. M., and Raushel, F. M. (2001). Channeling of substrates and intermediates in enzyme-catalyzed reactions. *Annual Review of Biochemistry*, 70:149–180.

Hughes, A. L. (1994). The evolution of functionally novel proteins after gene duplication. *Proceedings of the Royal Society of London B: Biological Sciences*, 256:119–124.

Hult, K. and Berglund, P. (2007). Enzyme promiscuity: mechanism and applications. *Trends in Biotechnology*, 25:231–238.

Ingram, V. (1959). Abnormal human haemoglobins. III the chemical difference between normal and sickle cell haemoglobins. *Biochimica et Biophysica Acta*, 36:402–411.

Jackson, R. C. and Handschumacher, R. E. (1970). *Escherichia coli* l-asparaginase. Catalytic activity and subunit nature. *Biochemistry*, 9:3585–3590.

Janin, J., Bahadur, R. P., and Chakrabarti, P. (2008). Protein–protein interaction and quaternary structure. *Quarterly Reviews of Biophysics*, 41:133–180.

Janssen, D. B., Dinkla, I. J., Poelarends, G. J., and Terpstra, P. (2005). Bacterial degradation of xenobiotic compounds: evolution and distribution of novel enzyme activities. *Environmental Microbiology*, 7:1868–1882.

Jenke-Kodama, H. and Dittmann, E. (2009). Evolution of metabolic diversity: insights from microbial polyketide synthases. *Phytochemistry*, 70:1858–1866.

Jenke-Kodama, H., Müller, R., and Dittmann, E. (2008). Evolutionary mechanisms underlying secondary metabolite diversity. In *Natural Compounds as Drugs Volume I*. Springer.

Jensen, R. A. (1976). Enzyme recruitment in evolution of new function. *Annual Reviews in Microbiology*, 30:409–425.

Jones, S. and Thornton, J. M. (1996). Principles of protein-protein interactions. *Proceedings of the National Academy of Sciences*, 93:13–20.

Jones, S. and Thornton, J. M. (1997). Prediction of protein-protein interaction sites using patch analysis. *Journal of Molecular Biology*, 272:133–143.

Jürgens, C., Strom, A., Wegener, D., Hettwer, S., Wilmanns, M., and Sterner, R. (2000). Directed evolution of a $(\beta\alpha)_8$-barrel enzyme to catalyze related reactions in two different metabolic pathways. *Proceedings of the National Academy of Sciences*, 97:9925–9930.

Kaiser, P., Meierhofer, D., Wang, X., and Huang, L. (2008). Tandem affinity purification combined with mass spectrometry to identify components of protein complexes. *Genomics Protocols*, 439:309–326.

Kaltenegger, E., Eich, E., and Ober, D. (2013). Evolution of homospermidine synthase in the convolvulaceae: a story of gene duplication, gene loss, and periods of various selection pressures. *The Plant Cell*, 25:1213–1227.

Kane, J. F. (1977). Regulation of a common amidotransferase subunit. *Journal of Bacteriology*, 132:419–425.

Karp, P. D., Riley, M., Saier, M., Paulsen, I. T., Collado-Vides, J., Paley, S. M., Pellegrini-Toole, A., Bonavides, C., and Gama-Castro, S. (2002). The EcoCyc database. *Nucleic Acids Research*, 30:56–58.

Kerbarh, O., Chirgadze, D. Y., Blundell, T. L., and Abell, C. (2006). Crystal structures of *Yersinia enterocolitica* salicylate synthase and its complex with the reaction products salicylate and pyruvate. *Journal of Molecular Biology*, 357:524–534.

Keskin, O., Tuncbag, N., and Gursoy, A. (2016). Predicting protein–protein interactions from the molecular to the proteome level. *Chemical Reviews*, 116:4884–4909.

Khersonsky, O., Roodveldt, C., and Tawfik, D. S. (2006). Enzyme promiscuity: evolutionary and mechanistic aspects. *Current Opinion in Chemical Biology*, 10:498–508.

Khersonsky, O. and Tawfik, D. S. (2010). Enzyme promiscuity: a mechanistic and evolutionary perspective. *Annual Review of Biochemistry*, 79:471–505.

Kim, D.-W., Kang, S.-M., and Yoon, K.-H. (1999). Isolation of novel *Pseudomonas diminuta* KAC-1 strain producing glutaryl 7-aminocephalosporanic acid acylase. *The Journal of Microbiology*, 37:200–205.

Kim, J., Fuller, J. H., Kuusk, V., Cunane, L., Chen, Z.-w., Mathews, F. S., and McIntire, W. S. (1995). The cytochrome subunit is necessary for covalent FAD attachment to the flavoprotein subunit of p-cresol methylhydroxylase. *Journal of Biological Chemistry*, 270:31202–31209.

Kim, K. M. and Caetano-Anollés, G. (2011). The proteomic complexity and rise of the primordial ancestor of diversified life. *BMC Evolutionary Biology*, 11:140.

Kimura, M. (1983). *The Neutral Theory of Molecular Evolution*. Cambridge University Press, New York.

Kirby, J. and Keasling, J. D. (2009). Biosynthesis of plant isoprenoids: perspectives for microbial engineering. *Annual Review of Plant Biology*, 60:335–355.

Klem, T. J., Chen, Y., and Davisson, V. J. (2001). Subunit interactions and glutamine utilization by *Escherichia coli* imidazole glycerol phosphate synthase. *Journal of Bacteriology*, 183:989–996.

Klem, T. J. and Davisson, V. J. (1993). Imidazole glycerol phosphate synthase: the glutamine amidotransferase in histidine biosynthesis. *Biochemistry*, 32:5177–5186.

Klose, T. and Rossmann, M. G. (2014). Structure of large dsDNA viruses. *Biological Chemistry*, 395:711–719.

Knöchel, T., Ivens, A., Hester, G., Gonzalez, A., Bauerle, R., Wilmanns, M., Kirschner, K., and Jansonius, J. N. (1999). The crystal structure of anthranilate synthase from *Sulfolobus solfataricus*: functional implications. *Proceedings of the National Academy of Sciences*, 96:9479–9484.

Kolappan, S., Zwahlen, J., Zhou, R., Truglio, J. J., Tonge, P. J., and Kisker, C. (2007). Lysine 190 is the catalytic base in MenF, the menaquinone-specific isochorismate synthase from *Escherichia coli*: implications for an enzyme family. *Biochemistry*, 46:946–953.

Koltin, Y., Faucette, L., Bergsma, D., Levy, M., Cafferkey, R., Koser, P., Johnson, R., and Livi, G. (1991). Rapamycin sensitivity in *Saccharomyces cerevisiae* is mediated by a peptidyl-prolyl cis-trans isomerase related to human FK506-binding protein. *Molecular and Cellular Biology*, 11:1718–1723.

Koonin, E. V. (2003). Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nature Reviews Microbiology*, 1:127–136.

Kortemme, T., Kim, D. E., and Baker, D. (2004). Computational alanine scanning of protein-protein interfaces. *Science Signaling*, 2004:pl2.

Kozlowski, M. C., Tom, N. J., Seto, C. T., Sefler, A. M., and Bartlett, P. A. (1995). Chorismate-utilizing enzymes isochorismate synthase, anthranilate synthase, and p-aminobenzoate synthase: mechanistic insight through inhibitor design. *Journal of the American Chemical Society*, 117:2128–2140.

Krissinel, E. and Henrick, K. (2007). Inference of macromolecular assemblies from crystalline state. *Journal of Molecular Biology*, 372:774–797.

Krogan, N. J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A. P., et al. (2006). Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature*, 440:637–643.

Kuang, D., Yao, Y., MacLean, D., Wang, M., Hampson, D. R., and Chang, B. S. (2006). Ancestral reconstruction of the ligand-binding pocket of family CG protein-coupled receptors. *Proceedings of the National Academy of Sciences*, 103:14050–14055.

Kühner, S., van Noort, V., Betts, M. J., Leo-Macias, A., Batisse, C., Rode, M., Yamada, T., Maier, T., Bader, S., Beltran-Alvarez, P., et al. (2009). Proteome organization in a genome-reduced bacterium. *Science*, 326:1235–1240.

Kunkel, T. A. and Bebenek, K. (2000). Dna replication fidelity. *Annual Review of Biochemistry*, 69:497–529.

Kössel, A. (1891). Über die chemische Zusammensetzung der Zelle. *Du Bois-Reymond's Archiv/Arch Anat Physiol Physiol Abt*, pages 181–186.

Küper, J., Doenges, C., and Wilmanns, M. (2005). Two-fold repeated $(\beta\alpha)_4$ half-barrels may provide a molecular tool for dual substrate specificity. *EMBO Reports*, 6:134–139.

Lamb, A. L. (2011). Pericyclic reactions catalyzed by chorismate-utilizing enzymes. *Biochemistry*, 50:7476–7483.

Lamb, A. L. (2015). Breaking a pathogen's iron will: Inhibiting siderophore production as an antimicrobial strategy. *Biochimica et Biophysica Acta - Proteins and Proteomics*, 1854:1054–1070.

Laporte, D. and Koshland, D. (1982). A protein with kinase and phosphatase activities involved in regulation of tricarboxylic acid cycle. *Nature*, 300:458–460.

Laporte, D. C. (1993). The isocitrate dehydrogenase phosphorylation cycle: regulation and enzymology. *Journal of Cellular Biochemistry*, 51:14–18.

Lazcano, A. and Miller, S. L. (1996). The origin and early evolution of life: prebiotic chemistry, the pre-RNA world, and time. *Cell*, 85:793–798.

Leitão, A. L. and Enguita, F. J. (2016). Editorial: Secondary metabolism. An unlimited foundation for synthetic biology. *Frontiers in Microbiology*, 6:1562.

Leonhardt, R. M., Lee, S.-J., Kavathas, P. B., and Cresswell, P. (2007). Severe tryptophan starvation blocks onset of conventional persistence and reduces reactivation of *Chlamydia trachomatis*. *Infection and Immunity*, 75:5105–5117.

Levin, K. B., Dym, O., Albeck, S., Magdassi, S., Keeble, A. H., Kleanthous, C., and Tawfik, D. S. (2009). Following evolutionary paths to protein-protein interactions with high affinity and selectivity. *Nature Structural & Molecular Biology*, 16:1049–1055.

Levitt, M. (2009). Nature of the protein universe. *Proceedings of the National Academy of Sciences*, 106:11079–11084.

Levy, E. D., De, S., and Teichmann, S. A. (2012). Cellular crowding imposes global constraints on the chemistry and evolution of proteomes. *Proceedings of the National Academy of Sciences*, 109:20461–20466.

Levy, E. D., Erba, E. B., Robinson, C. V., and Teichmann, S. A. (2008). Assembly reflects evolution of protein complexes. *Nature*, 453:1262–1265.

Levy, E. D. and Teichmann, S. (2013). Structural, evolutionary, and assembly principles of protein oligomerization. *Progress in Molecular Biology and Translational Science*, 117:25–51.

Li, Q.-A., Mavrodi, D. V., Thomashow, L. S., Roessle, M., and Blankenfeldt, W. (2011). Ligand binding induces an ammonia channel in 2-amino-2-desoxyisochorismate (ADIC) synthase PhzE. *Journal of Biological Chemistry*, 286:18213–18221.

Liberles, D. (2007). *Ancestral Sequence Reconstruction*. Oxford biosciences. Oxford University Press, Oxford.

List, F., Sterner, R., and Wilmanns, M. (2011). Related $(\beta\alpha)_8$-barrel proteins in histidine and tryptophan biosynthesis: A paradigm to study enzyme evolution. *ChemBioChem*, 12:1487–1494.

List, F., Vega, M. C., Razeto, A., Häger, M. C., Sterner, R., and Wilmanns, M. (2012). Catalysis uncoupling in a glutamine amidotransferase bienzyme by unblocking the glutaminase active site. *Chemistry & Biology*, 19:1589–1599.

Liu, B., Liu, B., Liu, F., and Wang, X. (2014). Protein binding site prediction by combining hidden markov support vector machine and profile-based propensities. *The Scientific World Journal*, page 464093.

Liu, J., Quinn, N., Berchtold, G. A., and Walsh, C. T. (1990). Overexpression, purification and characterization of isochorismate synthase (EntC), the first enzyme involved in the biosynthesis of enterobactin from chorismate. *Biochemistry*, 29:1417–1425.

Liu, M., Pazgier, M., Li, C., Yuan, W., Li, C., and Lu, W. (2010). A left-handed solution to peptide inhibition of the p53-MDM2 interaction. *Angewandte Chemie International Edition*, 49:3649–3652.

Loiseau, L., Ollagnier-de Choudens, S., Lascoux, D., Forest, E., Fontecave, M., and Barras, F. (2005). Analysis of the heteromeric CsdA-CsdE cysteine desulfurase, assisting Fe-S cluster biogenesis in *Escherichia coli*. *Journal of Biological Chemistry*, 280:26760–26769.

Longo, L. M. and Blaber, M. (2014). Symmetric protein architecture in protein design: top-down symmetric deconstruction. *Methods in Molecular Biology*, 1216:161–182.

Lopez-Goñi, I., Moriyon, I., and Neilands, J. (1992). Identification of 2,3-dihydroxybenzoic acid as a *Brucella abortus* siderophore. *Infection and Immunity*, 60:4496–4503.

Lynch, M. and Conery, J. S. (2000). The evolutionary fate and consequences of duplicate genes. *Science*, 290:1151–1155.

Ma, B., Wolfson, H. J., and Nussinov, R. (2001). Protein functional epitopes: hot spots, dynamics and combinatorial libraries. *Current Opinion in Structural Biology*, 11:364–369.

Marsh, J. A., Hernández, H., Hall, Z., Ahnert, S. E., Perica, T., Robinson, C. V., and Teichmann, S. A. (2013). Protein complexes are under evolutionary selection to assemble via ordered pathways. *Cell*, 153:461–470.

Marsh, J. A. and Teichmann, S. A. (2014). Parallel dynamics and evolution: protein conformational fluctuations and assembly reflect evolutionary changes in sequence and structure. *BioEssays*, 36:209–218.

Marsh, J. A. and Teichmann, S. A. (2015). Structure, dynamics, assembly, and evolution of protein complexes. *Annual Review of Biochemistry*, 84:551–575.

Martell, J. D., Yamagata, M., Deerinck, T. J., Phan, S., Kwa, C. G., Ellisman, M. H., Sanes, J. R., and Ting, A. Y. (2016). A split horseradish peroxidase for the detection of intercellular protein-protein interactions and sensitive visualization of synapses. *Nature Biotechnology*, 34:774–780.

Mashiyama, S. T., Malabanan, M. M., Akiva, E., Bhosle, R., Branch, M. C., Hillerich, B., Jagessar, K., Kim, J., Patskovsky, Y., Seidel, R. D., and M, S. (2014). Large-scale determination of sequence, structure, and function relationships in cytosolic glutathione transferases across the biosphere. *PLoS Biology*, 12:e1001843.

Massiere, F. and Badet-Denisot, M.-A. (1998). The mechanism of glutamine-dependent amidotransferases. *Cellular and Molecular Life Sciences*, 54:205–222.

Masters, P. A., O'Bryan, T. A., Zurlo, J., Miller, D. Q., and Joshi, N. (2003). Trimethoprim-sulfamethoxazole revisited. *Archives of Internal Medicine*, 163:402–410.

Matthews, T., Salgo, M., Greenberg, M., Chung, J., DeMasi, R., and Bolognesi, D. (2004). Enfuvirtide: the first therapy to inhibit the entry of HIV-1 into host CD4 lymphocytes. *Nature Reviews Drug Discovery*, 3:215–225.

May, J. J., Wendrich, T. M., and Marahiel, M. A. (2001). The *dhb* operon of *Bacillus subtilis* encodes the biosynthetic template for the catecholic siderophore 2,3-dihydroxybenzoate-glycine-threonine trimeric ester bacillibactin. *Journal of Biological Chemistry*, 276:7209–7217.

McAlpine, J. B., Bachmann, B. O., Piraee, M., Tremblay, S., Alarco, A.-M., Zazopoulos, E., and Farnet, C. M. (2005). Microbial genomics as a guide to drug discovery and structural elucidation: Eco-02301, a novel antifungal agent, as an example. *Journal of Natural Products*, 68:493–496.

Medema, M. H., Kottmann, R., Yilmaz, P., Cummings, M., Biggins, J. B., Blin, K., De Bruijn, I., Chooi, Y. H., Claesen, J., Coates, R. C., et al. (2015). Minimum information about a biosynthetic gene cluster. *Nature Chemical Biology*, 11:625–631.

Menche, J., Sharma, A., Kitsak, M., Ghiassian, S. D., Vidal, M., Loscalzo, J., and Barabási, A.-L. (2015). Uncovering disease-disease relationships through the incomplete interactome. *Science*, 347:841–849.

Meneely, K. M., Sundlov, J. A., Gulick, A. M., Moran, G. R., and Lamb, A. L. (2016). An open and shut case: The interaction of magnesium with mst enzymes. *Journal of the American Chemical Society*, 138:9277–9293.

Merino, E., Jensen, R. A., and Yanofsky, C. (2008). Evolution of bacterial *trp* operons and their regulation. *Current Opinion in Microbiology*, 11:78–86.

Merkl, R. and Sterner, R. (2016). Ancestral protein reconstruction: techniques and applications. *Biological Chemistry*, 397:1–21.

Meyer, V., Dinkel, P. D., Luo, Y., Yu, X., Wei, G., Zheng, J., Eaton, G. R., Ma, B., Nussinov, R., Eaton, S. S., and Margittai, M. (2014). Single mutations in tau modulate the populations of fibril conformers through seed selection. *Angewandte Chemie International Edition*, 53:1590–1593.

Mihara, H. and Esaki, N. (2002). Bacterial cysteine desulfurases: their function and mechanisms. *Applied Microbiology and Biotechnology*, 60:12–23.

Miles, B. W., Banzon, J. A., and Raushel, F. M. (1998). Regulatory control of the amidotransferase domain of carbamoyl phosphate synthetase. *Biochemistry*, 37:16773–16779.

Milewski, S. (2002). Glucosamine-6-phosphate synthase—the multi-facets enzyme. *Biochimica et Biophysica Acta - Protein Structure and Molecular Enzymology*, 1597:173–192.

Miller, S. L. (1953). A production of amino acids under possible primitive earth conditions. *Science*, 117:528–529.

Mintseris, J. and Weng, Z. (2005). Structure, function, and evolution of transient and obligate protein–protein interactions. *Proceedings of the National Academy of Sciences*, 102:10930–10935.

Mitchell, A., Chang, H.-Y., Daugherty, L., Fraser, M., Hunter, S., Lopez, R., McAnulla, C., McMenamin, C., Nuka, G., Pesseat, S., et al. (2014). The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Research*, 43:213–221.

Moghe, G. D. and Last, R. L. (2015). Something old, something new: conserved enzymes and the evolution of novelty in plant specialized metabolism. *Plant Physiology*, 169:1512–1523.

Morollo, A. A. and Eck, M. J. (2001). Structure of the cooperative allosteric anthranilate synthase from *Salmonella typhimurium*. *Nature Structural & Molecular Biology*, 8:243–247.

Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, 247:536–540.

Na, Z., Pan, S., Uttamchandani, M., and Yao, S. Q. (2014). Discovery of cell-permeable inhibitors that target the BRCT domain of BRCA1 protein by using a small-molecule microarray. *Angewandte Chemie International Edition*, 53:8421–8426.

Nakagawa, Y. and Bender, M. L. (1969). Modification of α-chymotrypsin by methyl p-nitrobenzenesulfonate. *Journal of the American Chemical Society*, 91:1566–1567.

Nakatsu, T., Kato, H., and Oda, J. (1998). Crystal structure of asparagine synthetase reveals a close evolutionary relationship to class II aminoacyl-tRNA synthetase. *Nature Structural & Molecular Biology*, 5:15–19.

Näsvall, J., Sun, L., Roth, J. R., and Andersson, D. I. (2012). Real-time evolution of new genes by innovation, amplification, and divergence. *Science*, 338:384–387.

Neilands, J. (1995). Siderophores: structure and function of microbial iron transport compounds. *Journal of Biological Chemistry*, 270:26723–26726.

Ngaki, M. N., Louie, G. V., Philippe, R. N., Manning, G., Pojer, F., Bowman, M. E., Li, L., Larsen, E., Wurtele, E. S., and Noel, J. P. (2012). Evolution of the chalcone-isomerase fold from fatty-acid binding to stereospecific catalysis. *Nature*, 485:530–533.

Ning, J., Moghe, G. D., Leong, B., Kim, J., Ofner, I., Wang, Z., Adams, C., Jones, A. D., Zamir, D., and Last, R. L. (2015). A feedback-insensitive isopropylmalate synthase affects acylsugar composition in cultivated and wild tomato. *Plant Physiology*, 169:1821–1835.

Nisbet, E. and Sleep, N. (2001). The habitat and nature of early life. *Nature*, 409:1083–1091.

Nobeli, I., Favia, A. D., and Thornton, J. M. (2009). Protein promiscuity and its implications for biotechnology. *Nature Biotechnology*, 27:157–167.

Noda-García, L., Camacho-Zarco, A. R., Medina-Ruíz, S., Gaytán, P., Carrillo-Tripp, M., Fülöp, V., and Barona-Gómez, F. (2013). Evolution of substrate specificity in a recipient's enzyme following horizontal gene transfer. *Molecular Biology and Evolution*, 30:2024–2034.

Noda-García, L., Camacho-Zarco, A. R., Verdel-Aranda, K., Wright, H., Soberón, X., Fülöp, V., and Barona-Gómez, F. (2010). Identification and analysis of residues contained on $\beta\rightarrow\alpha$-loops of the dual-substrate $(\beta\alpha)_8$ phosphoribosyl isomerase a specific for its phosphoribosyl anthranilate isomerase activity. *Protein Science*, 19:535–543.

Noda-García, L., Juárez-Vázquez, A. L., Ávila-Arcos, M. C., Verduzco-Castro, E. A., Montero-Morán, G., Gaytán, P., Carrillo-Tripp, M., and Barona-Gómez, F. (2015). Insights into the evolution of enzyme substrate promiscuity after the discovery of $(\beta\alpha)_8$ isomerase evolutionary intermediates from a diverse metagenome. *BMC Evolutionary Biology*, 15:107.

Nooren, I. M. and Thornton, J. M. (2003a). Diversity of protein–protein interactions. *The EMBO Journal*, 22:3486–3492.

Nooren, I. M. and Thornton, J. M. (2003b). Structural characterisation and functional significance of transient protein–protein interactions. *Journal of Molecular Biology*, 325:991–1018.

Ober, D., Harms, R., Witte, L., and Hartmann, T. (2003). Molecular evolution by change of function alkaloid-specific homospermidine synthase retained all properties of deoxyhypusine synthase except binding the eIF5A precursor protein. *Journal of Biological Chemistry*, 278:12805–12812.

O'Brien, P. J. and Herschlag, D. (1999). Catalytic promiscuity and the evolution of new enzymatic activities. *Chemistry & Biology*, 6:91–105.

Ofran, Y. and Rost, B. (2003). Analysing six types of protein–protein interfaces. *Journal of Molecular Biology*, 325:377–387.

Ofran, Y. and Rost, B. (2007). Protein–protein interaction hotspots carved into sequences. *PLoS Computational Biology*, 3:e119.

Ohno, S. (1970). *Evolution by gene duplication.* Springer, Berlin, Heidelberg.

Ollis, D. L., Cheah, E., Cygler, M., Dijkstra, B., Frolow, F., Franken, S. M., Harel, M., Remington, S. J., Silman, I., and Schrag, J. (1992). The α/β hydrolase fold. *Protein Engineering*, 5:197–211.

Olson, R. E. (1984). The function and metabolism of vitamin K. *Annual Review of Nutrition*, 4:281–337.

Oltersdorf, T., Elmore, S. W., Shoemaker, A. R., Armstrong, R. C., Augeri, D. J., Belli, B. A., Bruncko, M., Deckwerth, T. L., Dinges, J., Hajduk, P. J., et al. (2005). An inhibitor of Bcl-2 family proteins induces regression of solid tumours. *Nature*, 435:677–681.

Oparin, A. and Morgulis, S. (2003). *The Origin of Life.* Dover phoenix editions. Dover Publications, Mineola.

Orengo, C. A. and Thornton, J. M. (2005). Protein families and their evolution – a structural perspective. *Annual Reviews of Biochemistry*, 74:867–900.

Ortlund, E. A., Bridgham, J. T., Redinbo, M. R., and Thornton, J. W. (2007). Crystal structure of an ancient protein: evolution by conformational epistasis. *Science*, 317:1544–1548.

Ouzounis, C. A., Kunin, V., Darzentas, N., and Goldovsky, L. (2006). A minimal estimate for the gene content of the last universal common ancestor—exobiology from a terrestrial perspective. *Research in Microbiology*, 157:57–68.

Ovchinnikov, S., Kamisetty, H., and Baker, D. (2014). Robust and accurate prediction of residue–residue interactions across protein interfaces using evolutionary information. *Elife*, 3:e02030.

Palmer, D. R., Garrett, J. B., Sharma, V., Meganathan, R., Babbitt, P. C., and Gerlt, J. A. (1999). Unexpected divergence of enzyme function and sequence: "N-acylamino acid racemase" is a *o*-succinylbenzoate synthase. *Biochemistry*, 38:4252–4258.

Papp, B., Pal, C., and Hurst, L. D. (2003). Dosage sensitivity and the evolution of gene families in yeast. *Nature*, 424:194–197.

Park, K., Shen, B. W., Parmeggiani, F., Huang, P.-S., Stoddard, B. L., and Baker, D. (2015). Control of repeat-protein curvature by computational protein design. *Nature Structural & Molecular Biology*, 22:167–174.

Parrish, J. R., Gulyas, K. D., and Finley, R. L. (2006). Yeast two-hybrid contributions to interactome mapping. *Current Opinion in Biotechnology*, 17:387–393.

Parsons, J. F., Jensen, P. Y., Pachikara, A. S., Howard, A. J., Eisenstein, E., and Ladner, J. E. (2002). Structure of *Escherichia coli* aminodeoxychorismate synthase: architectural conservation and diversity in chorismate-utilizing enzymes. *Biochemistry*, 41:2198–2208.

Parsons, J. F., Shi, K. M., and Ladner, J. E. (2008). Structure of isochorismate synthase in complex with magnesium. *Acta Crystallographica Section D: Biological Crystallography*, 64:607–610.

Patrick, W. M. and Matsumura, I. (2008). A study in molecular contingency: glutamine phosphoribosylpyrophosphate amidotransferase is a promiscuous and evolvable phosphoribosylanthranilate isomerase. *Journal of Molecular Biology*, 377:323–336.

Pauling, L. and Zuckerkandl, E. (1963). Chemical paleogenetics. *Acta Chemica Scandinavica*, 17:9–16.

Perez-Jimenez, R., Inglés-Prieto, A., Zhao, Z.-M., Sanchez-Romero, I., Alegre-Cebollada, J., Kosuri, P., Garcia-Manyes, S., Kappock, T. J., Tanokura, M., Holmgren, A., Sanchez-Ruiz, J. M., Gaucher, E. A., and Fernandez, J. M. (2011). Single-molecule paleoenzymology probes the chemistry of resurrected enzymes. *Nature Structural & Molecular Biology*, 18:592–596.

Perica, T., Kondo, Y., Tiwari, S. P., McLaughlin, S. H., Kemplen, K. R., Zhang, X., Steward, A., Reuter, N., Clarke, J., and Teichmann, S. A. (2014). Evolution of oligomeric state through allosteric pathways that mimic ligand binding. *Science*, 346:1254346.

Petta, I., Lievens, S., Libert, C., Tavernier, J., and De Bosscher, K. (2015). Modulation of protein–protein interactions for the development of novel therapeutics. *Molecular Therapy*, 24:707–718.

Pires, D. E., Ascher, D. B., and Blundell, T. L. (2014). mCSM: predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics*, 30:335–342.

Plowman, S. and Hancock, J. (2005). Ras signaling from plasma membrane and endomembrane microdomains. *Biochimica et Biophysica Acta - Molecular Cell Research*, 1746:274–283.

Pocker, Y. and Stone, J. (1965). The catalytic versatility of erythrocyte carbonic anhydrase. The enzyme-catalyzed hydrolysis of p-nitrophenyl acetate. *Journal of the American Chemical Society*, 87:5497–5498.

Punta, M., Coggill, P. C., Eberhardt, R. Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., Heger, A., Holm, L., Sonnhammer, E. L., Eddy, S. R., Bateman, A., and Finn, R. D. (2012). The Pfam protein families database. *Nucleic Acids Research*, 40:290–301.

Rascher, A., Hu, Z., Viswanathan, N., Schirmer, A., Reid, R., Nierman, W. C., Lewis, M., and Hutchinson, C. R. (2003). Cloning and characterization of a gene cluster for geldanamycin production in *Streptomyces hygroscopicus* NRRL3602. *FEMS Microbiology Letters*, 218:223–230.

Raschle, T., Amrhein, N., and Fitzpatrick, T. B. (2005). On the two components of pyridoxal 5'-phosphate synthase from *Bacillus subtilis*. *Journal of Biological Chemistry*, 280:32291–32300.

Ratledge, C. (2004). Iron, mycobacteria and tuberculosis. *Tuberculosis*, 84:110–130.

Ratledge, C. and Dover, L. G. (2000). Iron metabolism in pathogenic bacteria. *Annual Reviews in Microbiology*, 54:881–941.

Raushel, F. M., Thoden, J. B., and Holden, H. M. (2003). Enzymes with molecular tunnels. *Accounts of Chemical Research*, 36:539–548.

Reichmann, D., Rahat, O., Cohen, M., Neuvirth, H., and Schreiber, G. (2007). The molecular architecture of protein–protein binding sites. *Current Opinion in Structural Biology*, 17:67–76.

Reid, T. W. and Fahrney, D. (1967). The pepsin-catalyzed hydrolysis of sulfite esters. *Journal of the American Chemical Society*, 89:3941–3943.

Reisinger, B., Sperl, J., Holinski, A., Schmid, V., Rajendran, C., Carstensen, L., Schlee, S., Blanquart, S., Merkl, R., and Sterner, R. (2013). Evidence for the existence of elaborate enzyme complexes in the paleoarchean era. *Journal of the American Chemical Society*, 136:122–129.

Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: the european molecular biology open software suite. *Trends in Genetics*, 16:276–277.

Richter, M., Bosnali, M., Carstensen, L., Seitz, T., Durchschlag, H., Blanquart, S., Merkl, R., and Sterner, R. (2010). Computational and experimental evidence for the evolution of a $(\beta\alpha)_8$-barrel protein from an ancestral quarter-barrel stabilised by disulfide bonds. *Journal of Molecular Biology*, 398:763–773.

Risso, V. A., Gavira, J. A., Mejia-Carmona, D. F., Gaucher, E. A., and Sanchez-Ruiz, J. M. (2013). Hyperstability and substrate promiscuity in laboratory resurrections of precambrian β-lactamases. *Journal of the American Chemical Society*, 135:2899–2902.

Rivalta, I., Sultan, M. M., Lee, N.-S., Manley, G. A., Loria, J. P., and Batista, V. S. (2012). Allosteric pathways in imidazole glycerol phosphate synthase. *Proceedings of the National Academy of Sciences*, 109:1428–1436.

Robert, F. and Chaussidon, M. (2006). A palaeotemperature curve for the precambrian oceans based on silicon isotopes in cherts. *Nature*, 443:969–972.

Robin, G., Sato, Y., Desplancq, D., Rochel, N., Weiss, E., and Martineau, P. (2014). Restricted diversity of antigen binding residues of antibodies revealed by computational alanine scanning of 227 antibody–antigen complexes. *Journal of Molecular Biology*, 426:3729–3743.

Rolland, T., Taşan, M., Charloteaux, B., Pevzner, S. J., Zhong, Q., Sahni, N., Yi, S., Lemmens, I., Fontanillo, C., Mosca, R., et al. (2014). A proteome-scale map of the human interactome network. *Cell*, 159:1212–1226.

Roodveldt, C. and Tawfik, D. S. (2005). Shared promiscuous activities and evolutionary features in various members of the amidohydrolase superfamily. *Biochemistry*, 44:12728–12736.

Safran, R. and Nosil, P. (2012). Speciation: The origin of new species. *Nature Education Knowledge*, 3:17.

Salwinski, L., Miller, C. S., Smith, A. J., Pettit, F. K., Bowie, J. U., and Eisenberg, D. (2004). The database of interacting proteins: 2004 update. *Nucleic Acids Research*, 32:449–451.

Sarkar, M., Li, C., and Pielak, G. J. (2013). Soft interactions and crowding. *Biophysical Reviews*, 5:187–194.

Scannell, D. R. and Wolfe, K. H. (2008). A burst of protein sequence evolution and a prolonged period of asymmetric evolution follow gene duplication in yeast. *Genome Research*, 18:137–147.

Schadt, H. S., Schadt, S., Oldach, F., and Süssmuth, R. D. (2009). 2-amino-2-deoxyisochorismate is a key intermediate in *Bacillus subtilis* p-aminobenzoic acid biosynthesis. *Journal of the American Chemical Society*, 131:3481–3483.

Schreiber, G. and Keating, A. E. (2011). Protein binding specificity versus promiscuity. *Current Opinion in Structural Biology*, 21:50–61.

Schuster-Böckler, B. and Bateman, A. (2008). Protein interactions in human genetic diseases. *Genome Biology*, 9:R9.

Schwecke, T., Aparicio, J. F., Molnar, I., König, A., Khaw, L. E., Haydock, S. F., Oliynyk, M., Caffrey, P., Cortes, J., and Lester, J. B. (1995). The biosynthetic gene cluster for the polyketide immunosuppressant rapamycin. *Proceedings of the National Academy of Sciences*, 92:7839–7843.

Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., and Serrano, L. (2005). The FoldX web server: an online force field. *Nucleic Acids Research*, 33:382–388.

Seffernick, J. L., de Souza, M. L., Sadowsky, M. J., and Wackett, L. P. (2001). Melamine deaminase and atrazine chlorohydrolase: 98 percent identical but functionally different. *Journal of Bacteriology*, 183:2405–2410.

Seffernick, J. L. and Wackett, L. P. (2016). Ancient evolution and recent evolution converge for the biodegradation of cyanuric acid and related triazines. *Applied and Environmental Microbiology*, 82:1638–1645.

Sehgal, S. (2003). Sirolimus: its discovery, biological properties, and mechanism of action. In *Transplantation Proceedings*, volume 35, pages 7–14. Elsevier.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13:2498–2504.

Shapir, N., Mongodin, E., Sadowsky, M., Daugherty, S., Nelson, K., and Wackett, L. (2007). Evolution of catabolic pathways: genomic insights into microbial S-triazine metabolism. *Journal of Bacteriology*, 189:674–682.

Sheinerman, F. B., Norel, R., and Honig, B. (2000). Electrostatic aspects of protein–protein interactions. *Current Opinion in Structural Biology*, 10:153–159.

Shi, Y. and Yokoyama, S. (2003). Molecular analysis of the evolutionary significance of ultraviolet vision in vertebrates. *Proceedings of the National Academy of Sciences*, 100:8308–8313.

Shoemaker, B. A. and Panchenko, A. R. (2007). Deciphering protein-protein interactions. part II. computational methods to predict protein and domain interaction partners. *PLoS Computational Biology*, 3:e43.

Sikosek, T. and Chan, H. S. (2014). Biophysics of protein evolution and evolutionary protein biophysics. *Journal of The Royal Society Interface*, 11:20140419.

Skolnick, J., Arakaki, A. K., Lee, S. Y., and Brylinski, M. (2009). The continuity of protein structure space is an intrinsic property of proteins. *Proceedings of the National Academy of Sciences*, 106:15690–15695.

Smith, J. L. (1998). Glutamine PRPP amidotransferase: snapshots of an enzyme in action. *Current Opinion in Structural Biology*, 8:686–694.

Smock, R. G., Yadid, I., Dym, O., Clarke, J., and Tawfik, D. S. (2016). De novo evolutionary emergence of a symmetrical protein is shaped by folding constraints. *Cell*, 164:476–486.

Sosio, M., Kloosterman, H., Bianchi, A., de Vreugd, P., Dijkhuizen, L., and Donadio, S. (2004). Organization of the teicoplanin gene cluster in *Actinoplanes teichomyceticus*. *Microbiology*, 150:95–102.

Spraggon, G., Kim, C., Nguyen-Huu, X., Yee, M.-C., Yanofsky, C., and Mills, S. E. (2001). The structures of anthranilate synthase of *Serratia marcescens* crystallized in the presence of (i) its substrates, chorismate and glutamine, and a product, glutamate, and (ii) its end-product inhibitor, L-tryptophan. *Proceedings of the National Academy of Sciences*, 98:6021–6026.

Sridharan, S., Howard, N., Kerbarh, O., Błaszczyk, M., Abell, C., and Blundell, T. L. (2010). Crystal structure of *Escherichia coli* enterobactin-specific isochorismate synthase (EntC) bound to its reaction product isochorismate: implications for the enzyme mechanism and differential activity of chorismate-utilizing enzymes. *Journal of Molecular Biology*, 397:290–300.

Stone, M. and Williams, D. (1992). On the evolution of functional secondary metabolites (natural products). *Molecular Microbiology*, 6:29–34.

Straight, P. D., Fischbach, M. A., Walsh, C. T., Rudner, D. Z., and Kolter, R. (2007). A singular enzymatic megacomplex from *Bacillus subtilis*. *Proceedings of the National Academy of Sciences*, 104:305–310.

Strange, R. C., Spiteri, M. A., Ramachandran, S., and Fryer, A. A. (2001). Glutathione-*S*-transferase family of enzymes. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 482:21–26.

Strobel, G. and Daisy, B. (2003). Bioprospecting for microbial endophytes and their natural products. *Microbiology and Molecular Biology Reviews*, 67:491–502.

Strohmeier, M., Raschle, T., Mazurkiewicz, J., Rippe, K., Sinning, I., Fitzpatrick, T. B., and Tews, I. (2006). Structure of a bacterial pyridoxal 5'-phosphate synthase complex. *Proceedings of the National Academy of Sciences*, 103:19284–19289.

Stumpf, M. P., Thorne, T., de Silva, E., Stewart, R., An, H. J., Lappe, M., and Wiuf, C. (2008). Estimating the size of the human interactome. *Proceedings of the National Academy of Sciences*, 105:6959–6964.

Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguez, P., Doerks, T., Stark, M., Muller, J., Bork, P., Jensen, L. J., and Mering, C. (2011). The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Research*, 39:561–568.

Tang, L., Shah, S., Chung, L., Carney, J., Katz, L., Khosla, C., and Julien, B. (2000). Cloning and heterologous expression of the epothilone gene cluster. *Science*, 287:640–642.

Tesmer, J. J., Klem, T. J., Deras, M. L., and Davisson, V. J. (1996). The crystal structure of GMP synthetase reveals a novel catalytic triad and is a structural paradigm for two enzyme families. *Nature Structural Biology*, 3:74–86.

Thoden, J. B., Holden, H. M., Wesenberg, G., Raushel, F. M., and Rayment, I. (1997). Structure of carbamoyl phosphate synthetase: a journey of 96 Å from substrate to product. *Biochemistry*, 36:6305–6316.

Thoden, J. B., Miran, S. G., Phillips, J. C., Howard, A. J., Raushel, F. M., and Holden, H. M. (1998). Carbamoyl phosphate synthetase: caught in the act of glutamine hydrolysis. *Biochemistry*, 37:8825–8831.

Thomson, J. M., Gaucher, E. A., Burgan, M. F., De Kee, D. W., Li, T., Aris, J. P., and Benner, S. A. (2005). Resurrecting ancestral alcohol dehydrogenases from yeast. *Nature Genetics*, 37:630–635.

Thorn, K. S. and Bogan, A. A. (2001). Asedb: a database of alanine mutations and their effects on the free energy of binding in protein interactions. *Bioinformatics*, 17:284–285.

Tohge, T., Watanabe, M., Hoefgen, R., and Fernie, A. R. (2013). Shikimate and phenylalanine biosynthesis in the green lineage. *Frontiers in Plant Science*, 4:62.

Tomarev, S. I. and Piatigorsky, J. (1996). Lens crystallins of invertebrates. *European Journal of Biochemistry*, 235:449–465.

Tomatis, P. E., Fabiane, S. M., Simona, F., Carloni, P., Sutton, B. J., and Vila, A. J. (2008). Adaptive protein evolution grants organismal fitness by improving catalysis and flexibility. *Proceedings of the National Academy of Sciences*, 105:20605–20610.

Traitcheva, N., Jenke-Kodama, H., He, J., Dittmann, E., and Hertweck, C. (2007). Non-colinear polyketide biosynthesis in the aureothin and neoaureothin pathways: An evolutionary perspective. *ChemBioChem*, 8:1841–1849.

Trefzer, A., Pelzer, S., Schimana, J., Stockert, S., Bihlmaier, C., Fiedler, H.-P., Welzel, K., Vente, A., and Bechthold, A. (2002). Biosynthetic gene cluster of simocyclinone, a natural multihybrid antibiotic. *Antimicrobial Agents and Chemotherapy*, 46:1174–1182.

Tudzynski, B. (2005). Gibberellin biosynthesis in fungi: genes, enzymes, evolution, and impact on biotechnology. *Applied Microbiology and Biotechnology*, 66:597–611.

Tuller, T., Birin, H., Gophna, U., Kupiec, M., and Ruppin, E. (2010). Reconstructing ancestral gene content by coevolution. *Genome Research*, 20:122–132.

Ugalde, J. A., Chang, B. S., and Matz, M. V. (2004). Evolution of coral pigments recreated. *Science*, 305:1433–1433.

Valdar, W. S. and Thornton, J. M. (2001). Conservation helps to identify biologically relevant crystal contacts. *Journal of Molecular Biology*, 313:399–416.

Van Petegem, F., Duderstadt, K. E., Clark, K. A., Wang, M., and Minor, D. L. (2008). Alanine-scanning mutagenesis defines a conserved energetic hotspot in the $\mathrm{Ca}_V\alpha_1$ AID-$\mathrm{Ca}_V\beta$ interaction site that is critical for channel modulation. *Structure*, 16:280–294.

Van Regenmortel, M. H. (2014). Specificity, polyspecificity, and heterospecificity of antibody-antigen recognition. *Journal of Molecular Recognition*, 27:627–639.

Veitia, R. A. (2004). Gene dosage balance in cellular pathways. *Genetics*, 168:569–574.

Venkatesan, K., Rual, J.-F., Vazquez, A., Stelzl, U., Lemmens, I., Hirozane-Kishikawa, T., Hao, T., Zenkner, M., Xin, X., Goh, K.-I., et al. (2009). An empirical framework for binary interactome mapping. *Nature Methods*, 6:83–90.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., et al. (2001). The sequence of the human genome. *Science*, 291:1304–1351.

Verduzco-Castro, E. A., Michalska, K., Endres, M., Juárez-Vazquez, A. L., Noda-García, L., Chang, C., Henry, C. S., Babnigg, G., Joachimiak, A., and Barona-Gómez, F. (2016). Co-occurrence of analogous enzymes determines evolution of a novel $(\beta\alpha)_8$-isomerase subfamily after non-conserved mutations in flexible loop. *Biochemical Journal*, 473:1141–1152.

Verma, I. M., Stevenson, J. K., Schwarz, E. M., Van Antwerp, D., and Miyamoto, S. (1995). Rel/NFκB/IκB family: intimate tales of association and dissociation. *Genes & Development*, 9:2723–2735.

Villiers, B. R. and Hollfelder, F. (2009). Mapping the limits of substrate specificity of the adenylation domain of tyca. *ChemBioChem*, 10:671–682.

Vincent, J. P. and Lazdunski, M. (1972). Trypsin-pancreatic trypsin inhibitor association. dynamics of the interaction and role of disulfide bridges. *Biochemistry*, 11:2967–2977.

Vining, L. (1992). Secondary metabolism, inventive evolution and biochemical diversity—a review. *Gene*, 115:135–140.

Völkel, P., Le Faou, P., and Angrand, P.-O. (2010). Interaction proteomics: characterization of protein complexes using tandem affinity purification–mass spectrometry. *Biochemical Society Transactions*, 38:883–887.

Voordeckers, K., Brown, C. A., Vanneste, K., van der Zande, E., Voet, A., Maere, S., and Verstrepen, K. J. (2012). Reconstruction of ancestral metabolic enzymes reveals molecular mechanisms underlying evolutionary innovation through gene duplication. *PLoS Biology*, 10:e1001446.

Walsh, C. T., Erion, M. D., Walts, A. E., Delany III, J. J., and Berchtold, G. A. (1987). Chorismate aminations: partial purification of *Escherichia coli* PABA synthase and mechanistic comparison with anthranilate synthase. *Biochemistry*, 26:4734–4745.

Walsh, C. T., Liu, J., Rusnak, F., and Sakaitani, M. (1990). Molecular studies on enzymes in chorismate metabolism and the enterobactin biosynthetic pathway. *Chemical Reviews*, 90:1105–1129.

Wang, N., Ozer, E. A., Mandel, M. J., and Hauser, A. R. (2014). Genome-wide identification of *Acinetobacter baumannii* genes necessary for persistence in the lung. *mBio*, 5:e01163–14.

Wang, Y., Li, C., and Pielak, G. J. (2010). Effects of proteins on protein diffusion. *Journal of the American Chemical Society*, 132:9392–9397.

Warbrick, E. (1997). Two's company, three's a crowd: the yeast two hybrid system for mapping molecular interactions. *Structure*, 5:13–17.

Weiss, M. C., Sousa, F. L., Mrnjavac, N., Neukirchen, S., Roettger, M., Nelson-Sathi, S., and Martin, W. F. (2016). The physiology and habitat of the last universal common ancestor. *Nature Microbiology*, 1:16116.

Weiss, U. (1986). Early research on the shikimate pathway: some personal remarks and reminiscences. In *The Shikimic Acid Pathway*. Springer, Berlin, Heidelberg.

Weng, J.-K. (2014). The evolutionary paths towards complexity: a metabolic perspective. *New Phytologist*, 201:1141–1149.

Weng, J.-K. and Noel, J. (2012). The remarkable pliability and promiscuity of specialized metabolism. In *Cold Spring Harbor symposia on quantitative biology*. Cold Spring Harbor Laboratory Press, New York.

Weng, J.-K., Philippe, R. N., and Noel, J. P. (2012). The rise of chemodiversity in plants. *Science*, 336:1667–1670.

Wente, S. R. and Schachman, H. (1987). Shared active sites in oligomeric enzymes: model studies with defective mutants of aspartate transcarbamoylase produced by site-directed mutagenesis. *Proceedings of the National Academy of Sciences*, 84:31–35.

Wetie, A. G. N., Sokolowska, I., Woods, A. G., Roy, U., Deinhardt, K., and Darie, C. C. (2014). Protein-protein interactions: switch from classical methods to proteomics and bioinformatics-based approaches. *Cellular and Molecular Life Sciences*, 71:205–228.

Williams, D. H., Stone, M. J., Hauck, P. R., and Rahman, S. K. (1989). Why are secondary metabolites (natural products) biosynthesized? *Journal of Natural Products*, 52:1189–1208.

Williamson, M. P. and Sutcliffe, M. J. (2010). Protein–protein interactions. *Biochemical Society Transactions*, 38:875–878.

Wilson, C., Agafonov, R., Hoemberger, M., Kutter, S., Zorba, A., Halpin, J., Buosi, V., Otten, R., Waterman, D., Theobald, D., and Kern, D. (2015). Using ancient protein kinases to unravel a modern cancer drug's mechanism. *Science*, 347:882–886.

Wink, M. (2011). *Annual plant reviews, biochemistry of plant secondary metabolism*, volume 40. John Wiley & Sons, New York.

Woese, C. (1998). The universal ancestor. *Proceedings of the National Academy of Sciences*, 95:6854–6859.

Wright, H., Barona-Gomez, F., Hodgson, D. A., and FülöP, V. (2004). Expression, purification and preliminary crystallographic analysis of phosphoribosyl isomerase (PriA) from *Streptomyces coelicolor*. *Acta Crystallographica Section D: Biological Crystallography*, 60:534–536.

Wright, H., Noda-García, L., Ochoa-Leyva, A., Hodgson, D. A., Fülöp, V., and Barona-Gomez, F. (2008). The structure/function relationship of a dual-substrate $(\beta\alpha)_8$-isomerase. *Biochemical and Biophysical Research Communications*, 365:16–21.

Xu, Z., Horwich, A. L., and Sigler, P. B. (1997). The crystal structure of the asymmetric GroEL-GroES-(ADP)$_7$ chaperonin complex. *Nature*, 388:741–750.

Yakhnin, H., Yakhnin, A. V., and Babitzke, P. (2007). Translation control of *trp*G from transcripts originating from the folate operon promoter of *Bacillus subtilis* is influenced by translation-mediated displacement of bound TRAP, while translation control of transcripts originating from a newly identified *trp*G promoter is not. *Journal of Bacteriology*, 189:872–879.

Yamaguchi, Y., Park, J.-H., and Inouye, M. (2011). Toxin-antitoxin systems in bacteria and archaea. *Annual Review of Genetics*, 45:61–79.

Yanofsky, C. (2007). Rna-based regulation of genes of tryptophan synthesis and degradation, in bacteria. *RNA*, 13:1141–1154.

Yanofsky, C. and Crawford, I. (1987). The tryptophan operon. *Escherichia coli and Salmonella typhimurium: Cellular and Molecular Biology*, 2:1453–1472.

Yanofsky, C., Kelley, R. L., and Horn, V. (1984). Repression is relieved before attenuation in the *trp* operon of *Escherichia coli* as tryptophan starvation becomes increasingly severe. *Journal of Bacteriology*, 158:1018–1024.

Yokoyama, S., Xing, J., Liu, Y., Faggionato, D., Altun, A., and Starmer, W. T. (2014). Epistatic adaptive evolution of human color vision. *PLoS Genetics*, 10:e1004884.

Yokoyama, S., Yang, H., and Starmer, W. T. (2008). Molecular basis of spectral tuning in the red-and green-sensitive (M/LWS) pigments in vertebrates. *Genetics*, 179:2037–2043.

Yu, C.-Y., Chou, L.-C., and Chang, D. T.-H. (2010a). Predicting protein-protein interactions in unbalanced data using the primary structure of proteins. *BMC Bioinformatics*, 11:167.

Yu, H., Braun, P., Yıldırım, M. A., Lemmens, I., Venkatesan, K., Sahalie, J., Hirozane-Kishikawa, T., Gebreab, F., Li, N., Simonis, N., et al. (2008). High-quality binary protein interaction map of the yeast interactome network. *Science*, 322:104–110.

Yu, J., Guo, M., Needham, C. J., Huang, Y., Cai, L., and Westhead, D. R. (2010b). Simple sequence-based kernels do not predict protein–protein interactions. *Bioinformatics*, 26:2610–2614.

Yu, T.-W., Müller, R., Müller, M., Zhang, X., Draeger, G., Kim, C.-G., Leistner, E., and Floss, H. G. (2001). Mutational analysis and reconstituted expression of the biosynthetic genes involved in the formation of 3-amino-5-hydroxybenzoic acid, the starter unit of rifamycin biosynthesis in *Amycolatopsis mediterranei* S699. *Journal of Biological Chemistry*, 276:12546–12555.

Yusupov, M. M., Yusupova, G. Z., Baucom, A., Lieberman, K., Earnest, T. N., Cate, J., and Noller, H. F. (2001). Crystal structure of the ribosome at 5.5 Å resolution. *Science*, 292:883–896.

Zalkin, H. and Smith, J. (2009). Enzymes utilizing glutamine as an amide donor. *Advances in Enzymology and Related Areas of Molecular Biology*, 72:87–144.

Zein, F., Zhang, Y., Kang, Y.-N., Burns, K., Begley, T. P., and Ealick, S. E. (2006). Structural insights into the mechanism of the PLP synthase holoenzyme from *Thermotoga maritima. Biochemistry*, 45:14609–14620.

Zellner, H., Staudigel, M., Trenner, T., Bittkowski, M., Wolowski, V., Icking, C., and Merkl, R. (2012). Prescont: Predicting protein-protein interfaces utilizing four residue properties. *Proteins: Structure, Function, and Bioinformatics*, 80:154–168.

Zhang, Q. C., Petrey, D., Deng, L., Qiang, L., Shi, Y., Thu, C. A., Bisikirska, B., Lefebvre, C., Accili, D., Hunter, T., Maniatis, T., Califano, A., and Honig, B. (2012a). Structure-based prediction of protein-protein interactions on a genome-wide scale. *Nature*, 490:556–560.

Zhang, W., Dourado, D. F., Fernandes, P. A., Ramos, M. J., and Mannervik, B. (2012b). Multidimensional epistasis and fitness landscapes in enzyme evolution. *Biochemical Journal*, 445:39–46.

Zhang, W., Fisher, J. F., and Mobashery, S. (2009). The bifunctional enzymes of antibiotic resistance. *Current Opinion in Microbiology*, 12:505–511.

Zhang, Y., Hubner, I. A., Arakaki, A. K., Shakhnovich, E., and Skolnick, J. (2006). On the origin and highly likely completeness of single-domain protein structures. *Proceedings of the National Academy of Sciences*, 103:2605–2610.

Zhang, Y. and Skolnick, J. (2005). The protein structure prediction problem could be solved using the current PDB library. *Proceedings of the National Academy of Sciences*, 102:1029–1034.

Zhang, Y. J., Ioerger, T. R., Huttenhower, C., Long, J. E., Sassetti, C. M., Sacchettini, J. C., and Rubin, E. J. (2012c). Global assessment of genomic regions required for growth in *Mycobacterium tuberculosis. PLoS Pathogens*, 8:e1002946.

Zhao, S., Sakai, A., Zhang, X., Vetting, M. W., Kumar, R., Hillerich, B., San Francisco, B., Solbiati, J., Steves, A., Brown, S., and E, A. (2014). Prediction and characterization of enzymatic activities guided by sequence similarity and genome neighborhood networks. *Elife*, 3:e03275.

Ziebart, K. T. and Toney, M. D. (2010). Nucleophile specificity in anthranilate synthase, aminodeoxychorismate synthase, isochorismate synthase, and salicylate synthase. *Biochemistry*, 49:2851–2859.

Ziemert, N., Lechner, A., Wietz, M., Millán-Aguiñaga, N., Chavarria, K. L., and Jensen, P. R. (2014). Diversity and evolution of secondary metabolism in the marine actinomycete genus *Salinispora. Proceedings of the National Academy of Sciences*, 111:1130–1139.

Zinzalla, G. and Thurston, D. (2009). Targeting protein-protein interactions for therapeutic intervention: a challenge for the future. *Future Medicinal Chemistry*, 1:65–93.

Zou, T., Risso, V. A., Gavira, J. A., Sanchez-Ruiz, J. M., and Ozkan, S. B. (2014). Evolution of conformational dynamics determines the conversion of a promiscuous generalist into a specialist enzyme. *Molecular Biology and Evolution*, 32:132–143.

Zwahlen, J., Kolappan, S., Zhou, R., Kisker, C., and Tonge, P. J. (2007). Structure and mechanism of MbtI, the salicylate synthase from *Mycobacterium tuberculosis*. *Biochemistry*, 46:954–964.

7 Publications

7.1 Publication A

Long-term persistence of bi-functionality contributes to the robustness of microbial life through exaptation

Maximilian G. Plach*, Bernd Reisinger*, Reinhard Sterner, and Rainer Merkl (2016).

* Equal contributions

PLoS Genetics 12:e1005836

PLOS | GENETICS

RESEARCH ARTICLE

Long-Term Persistence of Bi-functionality Contributes to the Robustness of Microbial Life through Exaptation

Maximilian G. Plach☯, Bernd Reisinger☯, Reinhard Sterner*, Rainer Merkl*

Institute of Biophysics and Physical Biochemistry, University of Regensburg, Regensburg, Germany

☯ These authors contributed equally to this work.
* Reinhard.Sterner@ur.de (RS); Rainer.Merkl@ur.de (RM)

OPEN ACCESS

Citation: Plach MG, Reisinger B, Sterner R, Merkl R (2016) Long-Term Persistence of Bi-functionality Contributes to the Robustness of Microbial Life through Exaptation. PLoS Genet 12(1): e1005836. doi:10.1371/journal.pgen.1005836

Editor: Jianzhi Zhang, University of Michigan, UNITED STATES

Received: November 10, 2015

Accepted: January 9, 2016

Published: January 29, 2016

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This work was supported by Fonds der Chemischen Industrie, PhD Fellowship DO 192/63 to MGP, https://www.vci.de; Cusanuswerk, PhD Fellowship to BR, https://www.cusanuswerk.de; the Deutsche Forschungsgemeinschaft Award Number: STE891/9-1 to RS, http://www.dfg.de; and by the Deutsche Forschungsgemeinschaft Award Number: ME2259/2-1 to RM, http://www.dfg.de. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

Modern enzymes are highly optimized biocatalysts that process their substrates with extreme efficiency. Many enzymes catalyze more than one reaction; however, the persistence of such ambiguities, their consequences and evolutionary causes are largely unknown. As a paradigmatic case, we study the history of bi-functionality for a time span of approximately two billion years for the sugar isomerase HisA from histidine biosynthesis. To look back in time, we computationally reconstructed and experimentally characterized three HisA predecessors. We show that these ancient enzymes catalyze not only the HisA reaction but also the isomerization of a similar substrate, which is commonly processed by the isomerase TrpF in tryptophan biosynthesis. Moreover, we found that three modern-day HisA enzymes from Proteobacteria and Thermotogae also possess low TrpF activity. We conclude that this bi-functionality was conserved for at least two billion years, most likely without any evolutionary pressure. Although not actively selected for, this trait can become advantageous in the case of a gene loss. Such exaptation is exemplified by the Actinobacteria that have lost the *trp*F gene but possess the bi-functional HisA homolog PriA, which adopts the roles of both HisA and TrpF. Our findings demonstrate that bi-functionality can perpetuate in the absence of selection for very long time-spans.

Author Summary

The term exaptation describes the process by which a trait that is initially just a by-product of another function may become important in a later evolutionary phase. For example, feathers served to insulate dinosaurs before helping birds fly. On the level of enzymes, bi-functionality can contribute to microbial evolution through exaptation. However, bi-functional enzymes may cause metabolic conflicts, if they are involved in different metabolic pathways. By characterizing properties of modern and computationally reconstructed ancestral variants of the sugar isomerase HisA, we demonstrate that it has been a bi-functional enzyme for the last two billion years. Most likely, bi-functionality persisted because the remaining TrpF activity is not harmful or its elimination would concurrently

Competing Interests: The authors have declared that no competing interests exist.

compromise HisA activity. Moreover, this substrate ambiguity is advantageous, as it allows compensating a gene loss as exemplified by the Actinobacteria. These microbes have lost the isomerase TrpF but possess the bi-functional HisA homolog PriA, which takes over the roles of both HisA and TrpF. Our results argue to view bi-functionality not as an evolutionary disadvantage but rather as a contribution to the evolvability of novel functions via exaptation.

Introduction

Enzymes are remarkably specific catalysts and this characteristic led to the traditional view of "one enzyme, one substrate, one reaction", which assumes an evolutionary preference for mono-functionality. However, it is clear now that enzymes can catalyze reactions other than those for which they evolved; see [1] and references therein. Prominent examples of multi-functional enzymes are glutathione S-transferases and cytochrome P450s, which can process several different substrates [1]. However, multi-functional enzymes may cause metabolic conflicts if they are involved in different, possibly independent, metabolic pathways [2]. Along these lines, multi-functionality seems to be of no immediate advantage for an organism, which argues against a positive selection of this trait. Presumably, neutral drift causes the broadening or narrowing of reaction specificity, see [1] and references therein; however it is unclear, whether multi-functionality is a short-term or a long-term trait.

Some evolutionary innovations originate non-adaptively as exaptations [3], *i. e.* as by-products of other, positively selected traits. These features were not built by natural selection for their current role. For example, feathers evolved for temperature regulation prior to their function in flight [3] and the light-refracting lens crystallins stem from enzymes [4]. *In silico* analyses suggest that exaptation is an important means of evolutionary innovation for metabolic systems [5]. The contribution of exaptation to evolutionary processes would be of even greater importance, if such traits existed over a long evolutionary time-span. In order to address this issue, we traced bi-functionality of a key metabolic enzyme over two billion years.

Most microbial genomes harbor a *his*A and a *trp*F gene, which are located within the histidine and tryptophan operons, respectively. The gene products HisA and TrpF catalyze analogous isomerizations of the aminoaldoses ProFAR (N′-[(5′-phosphoribosyl)-formimino]-5-aminoimidazole-4-carboxamide-ribonucleotide) and PRA (N-(5′-phosphoribosyl)anthranilate) into the aminoketoses PRFAR (N′-[(5′-phosphoribulosyl)-formimino]-5-aminoimidazole-4-carboxamide-ribonucleotide) and CdRP, respectively [6] (Fig 1). Most likely, genes for HisA and TrpF were present in the genome of the last universal common ancestor (LUCA) [7]; thus it can be expected that their modern successors process their specific substrates with high efficiency. The situation is different, however, in the Actinobacteria. Within this phylum, the *trp*F gene is missing in many genomes. As a substitute, the bi-functional isomerase PriA catalyzes both the HisA and the TrpF reactions [8] (Fig 1). PriA is a HisA homolog; the two enzymes are highly similar to each other with respect to sequence and structure [9, 10].

A detailed tracing of HisA bi-functionality required an analysis in two dimensions: A survey of PriA-like characteristics in modern HisA homologs and a retrospect of ancestors related to bacterial speciation. To begin with, we used *in silico* analyses and *in vitro* characterization of extant HisA enzymes and found that PriA-like bi-functionality is not strictly limited to Actinobacteria. Furthermore, we reconstructed *in silico* the sequences of the HisA/PriA ancestors of all Actinobacteria, all Proteobacteria, and all Bacteria, and tested the resulting precursor proteins for their ProFAR and PRA isomerase activities. Our results show that all three

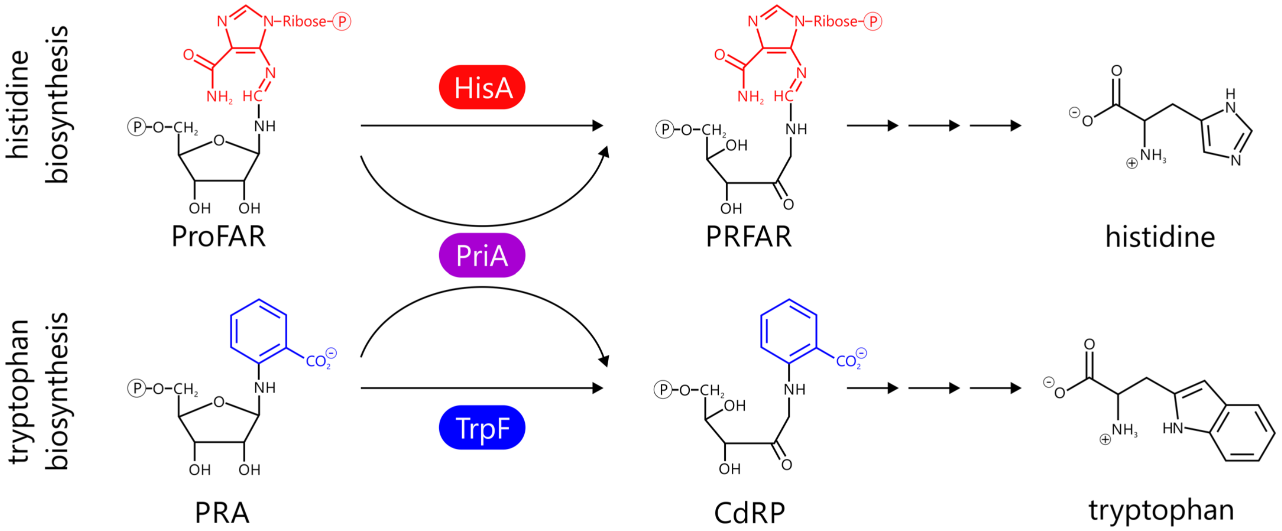


Fig 1. Isomerization of the aminoaldoses ProFAR and PRA to the aminoketoses PRFAR and CdRP. In most prokaryotes the two reactions are catalyzed by the enzymes HisA and TrpF, respectively. In Actinobacteria, however, the bi-functional PriA catalyzes both isomerizations.

doi:10.1371/journal.pgen.1005836.g001

reconstructed ancestral enzymes are bi-functional *in vitro* and *in vivo*. Thus, our findings provide an example for an enzyme, whose bi-functionality pertained for two billion years of evolution, most likely without obvious, immediate benefit, except for exaptation.

Results

Occurrence and functional characterization of extant HisA and PriA enzymes

The existence of the bi-functional PriA enzyme has originally been described for two actinobacterial species, namely *Streptomyces coelicolor* and *Mycobacterium tuberculosis* [8]. In order to determine the distribution of PriA-like enzymes within all bacterial phyla, we computed a sequence similarity network (SSN) of the HisA/PriA superfamily (Fig 2). In an SSN, nodes represent individual sequences and edges correspond to statistically significant similarities deduced from pairwise alignments, calculated by BLAST [11]. Our analysis showed that *his*A genes are present in all major phylogenetic groups (Fig 2A) and that the occurrence of annotated *pri*A genes is indeed restricted to the Actinobacteria (Fig 2B, top right cluster). The mean sequence identity in the Actinobacteria cluster is 52±9%; it can thus be assumed that all these sequences correspond to PriA enzymes.

The ability of PriA to catalyze both the HisA and the TrpF reaction requires that its active site can bind the two respective substrates in a productive conformation. As it is evident from the crystal structure of PriA from *M. tuberculosis* (mtPriA) [9], both substrates are bound in the same active site pocket (Fig 3). The most notable difference between the HisA state (Fig 3A) and the TrpF state (Fig 3B) is a twist of loop 5 and a concomitant swap of the localization of R143 and W145. This goes along with rearranged hydrogen bond networks at positions 19 and 109. Despite that, however, the same eight residues are involved in forming the active site in both states. We thus analyzed and compared their conservation in HisA and PriA sequences from the major SSN clusters. The actinobacterial PriA active site is characterized by a strong residue conservation resulting in the motif D-R-E-D-R-G-W-D (Fig 3C, Actinobacteria

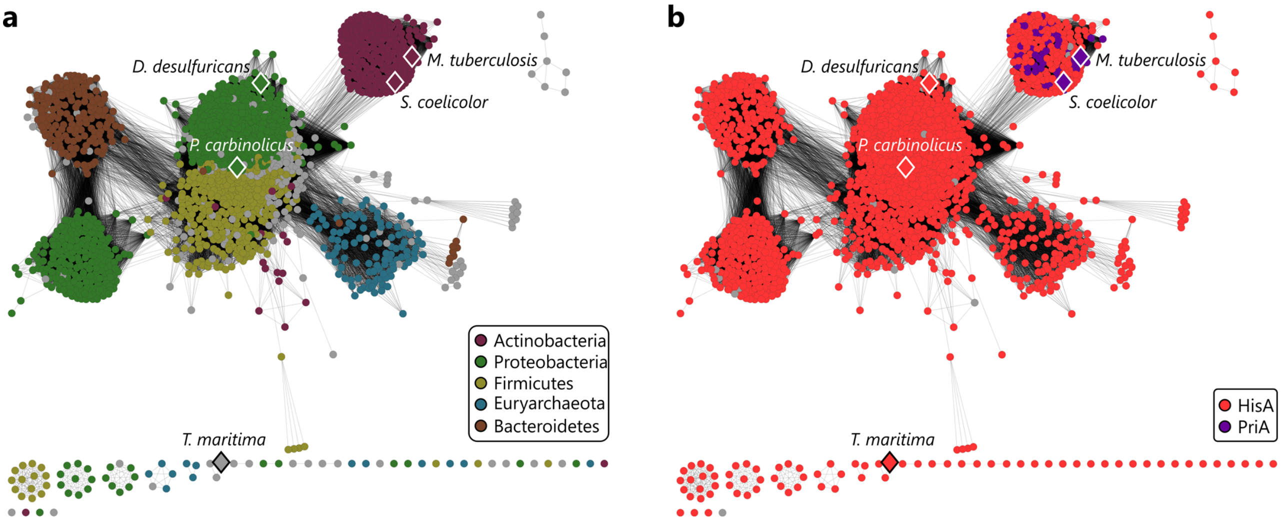


Fig 2. Sequence similarity network of the HisA/PriA superfamily. Nodes are colored by either the main five phyla contributing to this superfamily (A) or by annotation as HisA or PriA (B). The network was generated from all-by-all BLAST comparisons of 7428 HisA and PriA sequences. Each node represents a single sequence or a group of sequences with more than 95% identical residues; experimentally characterized HisA or PriA proteins are marked by diamonds. Each edge in the network represents a bi-directional BLAST hit with an E-value $\leq$ 1E−54 (corresponding to a median sequence identity of 44%). At this cutoff, the PriA cluster is clearly separated from, but still connected to the central HisA cluster. Lengths of edges are not meaningful except that sequences in tightly clustered groups are relatively more similar to each other than sequences with few connections.

doi:10.1371/journal.pgen.1005836.g002

sequence logo). In contrast, the majority of HisA sequences deviate from the PriA-typical motif in 2-3 residues, mainly at positions 109 and 143 (Fig 3C, remaining sequence logos). Surprisingly, however, the PriA-typical motif is present in some HisA enzymes from Bacteroidetes (1 representative corresponding to 0.4% of all Bacteroidetes sequences), Euryarchaeota (6 / 5.1%), Firmicutes (25 / 8.9%), and Proteobacteria (43 / 4.9%). Moreover, the PriA-typical motif is also found in HisA from *Thermotoga maritima* (tmHisA), except that Lys is present at position 143 instead of the PriA-typical Arg.

In order to test if the presence of the PriA-typical active site sequence motif in HisA enzymes leads to TrpF activity, tmHisA and two HisA enzymes from Proteobacteria (*Pelobacter carbinolicus*, pcHisA; *Desulfovibrio desulfuricans*, ddHisA) were produced by heterologous gene expression in *Escherichia coli*. The recombinant proteins were purified and characterized by steady-state kinetics with respect to their ProFAR and PRA isomerization activities. Compared to PriA from *S. coelicolor* (scPriA) and *M. tuberculosis* (mtPriA), the catalytic efficiencies k_{cat}/K_M^{ProFAR} of tmHisA, ddHisA, and pcHisA are about tenfold higher (Table 1, HisA reaction). They are comparable to the catalytic efficiency k_{cat}/K_M^{ProFAR} of HisA from *Salmonella enterica* (seHisA), which is considered to be an archetypical representative of the HisA family [12]. Strikingly, tmHisA, ddHisA, and pcHisA also displayed TrpF-activity, something that has not been shown before. However, their catalytic efficiencies k_{cat}/K_M^{PRA} are lower by about 10^5–10^6-fold compared to scPriA and mtPriA (Table 1, TrpF reaction).

In vivo complementation experiments showed that tmHisA, ddHisA, and pcHisA were able to rescue the growth deficiency of an *E. coli* Δ*his*A strain. Moreover, despite their weak *in vitro* TrpF activity, they were also able to complement a Δ*trp*F strain (Table 2). The enzymes were further able to complement a Δ*his*AΔ*trp*F double deletion strain (Table 2), whereby the time required for complementation is clearly limited by their TrpF activity.

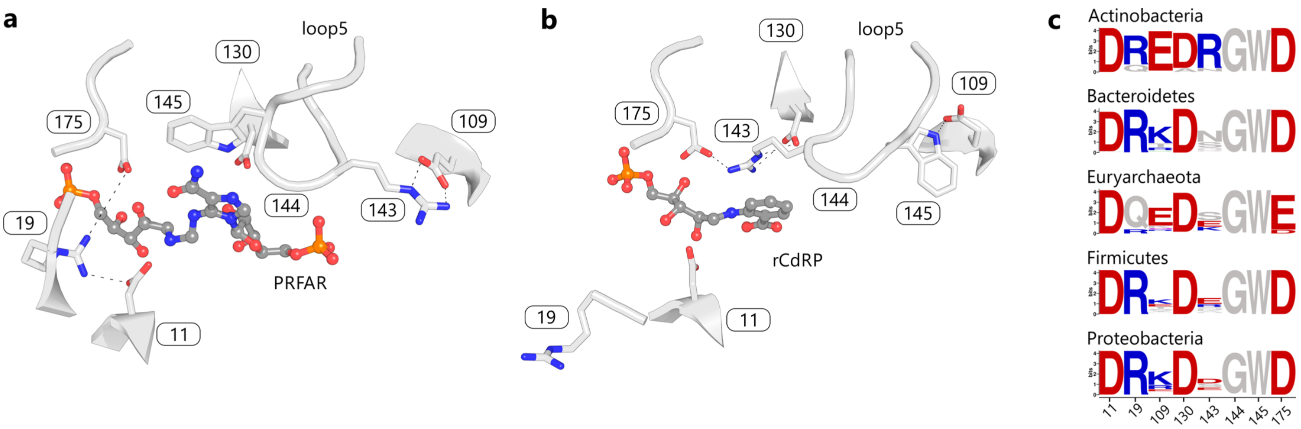


Fig 3. Two states of the PriA active site from *M. tuberculosis*. (a) Schematic view of the site in the HisA-state (bound product PRFAR, PDB ID 3zs4). (b) The same active site in the TrpF-state (bound product analogue reduced-CdRP, PDB ID 2y85). Residues of the active site are shown as stick models. Residue numbering is based on PDB ID 3zs4. (c) Sequence logos showing the conservation of the motif as deduced from SSN clusters of the HisA/PriA superfamily. Basic and acidic residues are colored blue and red, respectively.

doi:10.1371/journal.pgen.1005836.g003

Reconstruction of ancient sequences

We next asked whether the bi-functionality of HisA is an ancient feature that has been conserved in certain extant enzymes. To this end, we computationally reconstructed three HisA precursors as described in the following. It has been shown that concatenating related sequences increases the strength of the phylogenetic signal available for tree construction [14]. Thus, we concatenated species-wise HisA with HisH and HisF sequences. The respective genes were most likely part of the LUCA genome [7] and have remained elements of the histidine

Table 1. Steady-state kinetic parameters of extant PriA and HisA enzymes, and reconstructed HisA ancestors.

	HisA reaction			TrpF reaction		
Enzyme	k_{cat} [s^{-1}]	K_M^{ProFar} [μM]	k_{cat}/K_M^{ProFAR} [M^{-1}s^{-1}]	k_{cat} [s^{-1}]	K_M^{PRA} [μM]	k_{cat}/K_M^{PRA} [M^{-1}s^{-1}]
scPriA [1]	0.9	28	$3.2 \cdot 10^4$	12	4	$3.0 \cdot 10^6$
mtPriA [2]	0.23	19	$1.2 \cdot 10^4$	3.6	21	$1.7 \cdot 10^5$
tmHisA [3]	1.0	5.6	$1.8 \cdot 10^5$	$6.7 \cdot 10^{-3}$	60	$1.1 \cdot 10^1$
ddHisA [3]	1.3	2.8	$4.6 \cdot 10^5$	$2.3 \cdot 10^{-3}$	161	$1.4 \cdot 10^1$
pcHisA [3]	0.4	1.8	$2.2 \cdot 10^5$	$1.0 \cdot 10^{-3}$	303	$0.3 \cdot 10^1$
seHisA [4]	7.8	17.0	$4.5 \cdot 10^5$	n. d.	n. d.	n. d.
CA-Act-HisA [5]	n. d.	n. d.	$3.0 \cdot 10^2$	$1.0 \cdot 10^{-2}$	3	$3.3 \cdot 10^3$
CA-Prot-HisA [6]	0.05	0.3	$1.8 \cdot 10^5$	$5.3 \cdot 10^{-4}$	2.7	$2.0 \cdot 10^2$
CA-Bact-HisA [6]	0.05	0.5	$1.0 \cdot 10^5$	$2.3 \cdot 10^{-4}$	3.2	$0.7 \cdot 10^2$

[1] Data taken from [10].

[2] Data taken from [9].

[3] Unlike in previous work [13], tmHisA (as well as ddHisA and pcHisA) showed measurable albeit very low TrpF activity. Although the exact reasons for this discrepancy are unknown, it may be due to differences in enzyme purification and handling.

[4] Data taken from [12]; n. d.: values were not determined.

[5] n. d.: k_{cat} and K_M^{ProFAR} could not be determined individually; k_{cat}/K_M^{ProFAR} was deduced from the linear part of the saturation curve.

[6] The k_{cat} and K_M^{ProFAR} values were determined by analyzing entire transition curves with the integrated Michaelis-Menten equation.

[3,5,6] The standard errors for k_{cat} and K_M were between 5% and 40%.

doi:10.1371/journal.pgen.1005836.t001

Table 2. *In vivo* complementation of auxotrophic *E. coli* strains by PriA, HisA, HisA ancestors, and TrpF.

	complementation time in hours		
	Δ*his*A strain	Δ*trp*F strain	Δ*his*AΔ*trp*F strain
scPriA	22	22	23
tmHisA	16	114	144
ddHisA	16	153	181
pcHisA	15	70	63
CA-Act-HisA	48	23	47
CA-Prot-HisA	16	33	28
CA-Bact-HisA	16	45	39
tmTrpF	no growth	24	no growth

For all experiments, the mean time is given after which visible colonies appeared on minimal medium agar plates. All experiments were repeated independently at least three times. A growth time of 16 hours indicates that colonies appeared over night. Growth times below 120 hours could be reproduced with a maximum error of 25%, growth times above 120 hours with a maximum error of 40%. “No growth” indicates that no colonies were observed after 14 days. A negative control with empty pTNA plasmid did not lead to growth within 14 days, either.

doi:10.1371/journal.pgen.1005836.t002

operon since then. Bacterial and archaeal genomes were scanned for the occurrence of *his*A genes, and species were selected for which *his*A, *his*F, and *his*H were gene neighbors. We picked sequences from Euryarchaeota (5 species), Crenarchaeota (20), Bacteroidetes (8), Firmicutes (11), Spirochaetes (5), and the α-, β-, γ-, and δ-Proteobacteria (21, 5, 1, 5). Moreover, we added 22 actinobacterial sequence sets, by selecting genes whose products contain the above mentioned PriA active site sequence motif.

The resulting MSA_{HisFAH} comprised 103 concatenations (species names listed in S1 Table). After preprocessing this input, a phylogenetic tree was determined and assessed by means of PhyloBayes v3.3 [15]. Four independent MCMC samplings of length 50,000 were computed using pb and compared to ensure convergence. Several parameters confirmed the validity of our approach: Convergence and mixing were checked by means of the discrepancy index maxdiff; for the pairwise comparison of all chains, the maxdiff value was at most 0.06. The effective size was at least 100, as determined by means of tracecomp. A consensus tree was deduced from the concatenation of these four chains (S1 Fig). The posterior probability of edges interlinking ancestors of phyla or classes was at least 0.87, which testifies to the high quality of the tree.

This tree and the corresponding MSA_{HisFAH} were used to deduce a predecessor of the actinobacterial enzymes (CA-Act-HisA) by means of FASTML [16]. In order to exclude any effect of the 22 actinobacterial sequences (and especially their active site motif) on the reconstruction of more ancient predecessors of HisA, these sequences were removed from MSA_{HisFAH}. The resulting $MSA_{HisFAH\text{-}Act}$, which contained the remaining 81 non-actinobacterial sequences, was used to calculate a second tree (S2 Fig). Applying FASTML, the sequences of the common ancestors of Proteobacteria (CA-Prot-HisA) and of Bacteria (CA-Bact-HisA) were determined. A schematic representation of the two trees is given in Fig 4. The archaeal sequences served as an outgroup in both reconstructions.

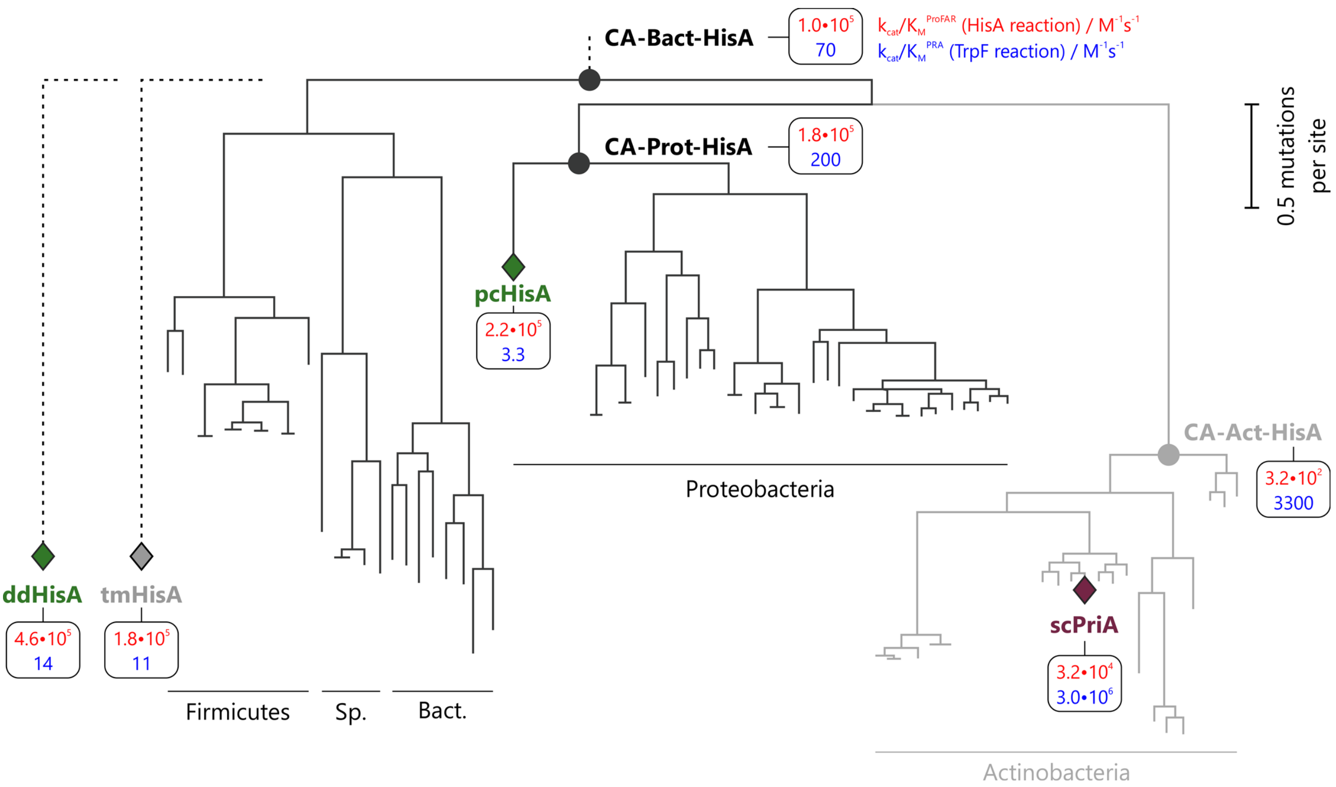


Fig 4. Phylogenetic tree depicting the position of extant HisA and PriA enzymes (diamonds) and their relationship to the reconstructed ancestral HisA enzymes (circles). The topology of the tree was inferred from the phylogenetic trees used for sequence reconstruction (S1 and S2 Figs). CA-Act-HisA, CA-Prot-HisA, and CA-Bact-HisA are the predecessor of HisA enzymes from Actinobacteria, Proteobacteria and Bacteria, respectively. Note that actinobacterial sequences were omitted for reconstruction of CA-Prot-HisA and CA-Bact-HisA (indicated by grey shading of the Actinobacteria branch). ddHisA and tmHisA were not used for sequence reconstruction and are only listed because they were characterized experimentally. The vertical bar indicates the branch length that corresponds to 0.5 mutations per site. The catalytic efficiencies k_{cat}/K_M of the enzymes for processing ProFAR and PRA are given in red and blue, respectively. Abbreviations: sc, *S. coelicolor*; dd, *D. desulfuricans*; pp, *P. carbinolicus*; tm, *T. maritima*; Sp., Spirochaetes; Bact., Bacteroidetes.

doi:10.1371/journal.pgen.1005836.g004

Experimental assessment of HisA precursors

The genes coding for the three precursors were synthesized and heterologously expressed in *E. coli*. The recombinant proteins were soluble and stable, and could be purified. Steady-state kinetic analysis yielded k_{cat}/K_M^{ProFAR} values in the order of 10^2–10^5 $M^{-1}s^{-1}$ for the HisA reaction, and k_{cat}/K_M^{PRA} values in the order of 10^2–10^3 $M^{-1}s^{-1}$ for the TrpF reaction (Table 1). Compared to scPriA and mtPriA, the catalytic efficiency of the ancestral proteins for the TrpF reaction is therefore only two to three orders of magnitude lower. For all three proteins this is the result of a lower k_{cat} value; the K_M^{PRA} is practically identical to that of scPriA. Furthermore, all three precursors were able to complement the growth deficiencies of Δ*his*A and Δ*trp*F strains (Table 2). The time required for *in vivo* complementation agrees well with k_{cat}/K_M values determined from *in vitro* measurements. For example, CA-Bact-HisA and CA-Prot-HisA have the highest k_{cat}/K_M^{ProFAR} values and required the least time to complement the Δ*his*A strain. CA-Act-HisA has the highest k_{cat}/K_M^{PRA} value and required the least time to complement the Δ*trp*F strain. All three HisA-ancestors were further able to complement the growth deficiency of a Δ*his*AΔ*trp*F double deletion strain (Table 2). The observed complementation times agree well with those determined from the single deletion strains. The complementation by CA-Act-HisA is limited by its ability to compensate for the missing HisA reaction, whereas complementation by CA-Prot-HisA and CA-Bact-HisA is limited by their ability to catalyze the missing TrpF reaction.

The active site sequence motif of CA-Act-HisA is identical to that of modern PriA enzymes. The motifs of CA-Prot-HisA and CA-Bact-HisA match in six of the eight residues. Non-matching is position 109, which contains a Lys instead of a Glu. At the second non-matching position 143, both precursors contain a Lys instead of an Arg. It is therefore plausible to assume that a basic residue at position 143 is crucial for bi-functionality. In contrast, the recently published SGG sequence motif of PriA [17] seems not to be required for bi-functionality. Only the immediate actinobacterial precursor CA-Act-HisA contains the SGG-motif whereas both other precursors displayed significant bi-functionality albeit containing a GGG-motif.

Discussion

In contrast to previous results [18], the reconstructed CA-Prot-HisA and CA-Bact-HisA are to our knowledge the first examples of ancestral metabolic enzymes from approximately 2.5 to 2.0 billion years ago [19] that were shown to be bi-functional. This trait is even more interesting when one considers that only extant HisA sequences but no extant PriA sequences were selected to reconstruct the CA-Prot-HisA and CA-Bact-HisA predecessors.

Strikingly, we also detected bi-functionality in the modern tmHisA, pcHisA, and ddHisA and thus provide the first examples of HisA/TrpF bi-functionality in extant HisA enzymes. It is worth noting that these three species all contain a *trp*F gene, which suggests that no selective pressure exists for these species to maintain the bi-functionality in HisA. Moreover, the *in vivo* complementation experiments show that tmTrpF is functional and is able to rescue an *E. coli* Δ*trp*F strain (Table 2). Also, the bi-functionality of these modern HisA enzymes does not force their hosts to face functional trade-offs because K_M^{PRA} values are 10- to 170-fold higher than K_M^{ProFAR} values. Thus the obligate HisA activity of these enzymes is most likely not impaired by the binding of PRA or CdRP. Moreover, the catalytic efficiencies k_{cat}/K_M^{PRA} are in a physiologically irrelevant range below 14 $M^{-1}s^{-1}$ thus making TrpF side-activity tolerable. Along these lines, the CA-Bact-HisA predecessor evolved most likely in a similar way such that the remaining TrpF side-activity was physiologically not harmful.

Our results do not allow us to decide whether all modern HisA enzymes are bi-functional: We have performed *in vivo* complementation experiments with four additional HisA enzymes from Bacteroidetes, Firmicutes, Proteobacteria, and Euryarchaeota lacking the PriA-typical sequence motif. These enzymes were unable to rescue Δ*trp*F or Δ*his*AΔ*trp*F deletion strains within eight days. Nevertheless, extremely slow growing colonies were observed occasionally. This growth may be due to residual TrpF activity of inherent *E. coli* enzymes like PurF [20] and may therefore indicate the existence of additional routes of exaptation. The active site motifs (Fig 3) suggest that bi-functionality is determined by Glu 109 and Arg 143. HisA homologs that retained bi-functionality have conserved the PriA typical residues at these two positions, despite a relatively low overall sequence identity. As this bi-functionality seems to be neither beneficial nor harmful for an organism, we assume that its presence is simply a matter of historical contingency. This conclusion is in agreement with the finding that a few mutations acquired in not more than several thousand generations were sufficient to transform a bi-functional HisA variant from *S. enterica* into a specialized HisA enzyme lacking TrpF activity or *vice versa* [21]. Along these lines, the bi-functional PriA became a mono-functional HisA enzyme in the Corynebacteria, a distinct genus within the Actinobacteria. This re-narrowing of substrate specificity in the so-called subHisA occurred after the horizontal acquisition of a whole pathway tryptophan operon (including a *trp*F gene) from a member of the γ-Proteobacteria [22]. Again, this transition from a bi-functional PriA to a mono-functional HisA enzyme required only subtle sequence alterations [17]. Noteworthy is a change from Arg 143 to an Asn, which supports the important role of Arg 143 for bi-functionality. Again, mono-functionality of HisA is easily accessible, if under evolutionary constraints. For Corynebacteria, this evolutionary pressure is most likely due to a metabolic conflict between histidine and tryptophan biosynthesis.

This bi-functionality provided a means for compensating the loss of the *trp*F gene within the Actinobacteria. Importantly, such exaptations are not rare: A screening of 104 single-gene knockout strains made clear that approximately 20% of these auxotrophs were rescued by the overexpression of at least one noncognate *E. coli* gene [23]. Thus, the functional diversity of gene products contributes to metabolic robustness and evolvability. These evolutionary advantages are further increased, if a bi-functionality that confers no cost or benefit to organismal fitness, can be conserved throughout long evolutionary time-spans. The characteristics of ancient and extant HisA and PriA enzymes confirm that this is feasible, even for enzymes of the primary metabolism.

Materials and Methods

Generation of sequence similarity networks

The SSN of the HisA/PriA-superfamily (7824 sequences, IPR023016 from InterPro release 47.0 [24]) was created using standard methods [25] provided by the Enzyme Function Initiative [26]. In order to eliminate sequence fragments, the length of the sequences that were included in the all-by-all BLAST comparison was restricted to 230–260 amino acids. From the remaining 7428 sequences, a representative network with an E-value cut-off of 1E-54 was generated in which sequences that share >95% identity were grouped into single nodes by CD-HIT [27]. Detailed phylogenetic information (superkingdom, phylum, class, order, family, genus) was added for each node using a modified version of Key2Ann [28]. Networks were visualized with the organic y-files layout in Cytoscape 3.2.0 [29, 30]. Phylum-specific sequence sets were compiled from the SSN and used to compute sequence logos of the active site residues, essentially as described [31].

Reconstruction of ancestral sequences

BLAST [11] and the nr database of the NCBI were used to search for the sequences of HisA homologs in completely sequenced genomes. Species where chosen, where *his*A and the *his*F and *his*H genes were neighbors in the genome; the respective sequences were concatenated. We selected species from the archeal phyla Euryarchaeota and Crenarchaeota, and from the bacterial phyla Bacteroidetes, Firmicutes, Spirochaetes, Actinobacteria, and Proteobacteria. A multiple sequence alignment (MSA) was deduced by means of MAFFT [32]. Positions containing more than 50% gaps were removed by using GBlocks [33]. The resulting MSA contained 430 meaningful positions. The program `pb` (version 3.3 of PhyloBayes, [15]) with options`-cat-gtr` was used to compute in four independent Monte Carlo Markov Chains (MCMC) 50 000 samples each. The options`-cat-gtr` induce an infinite mixture model, whose components differ by their equilibrium frequencies. The quality of mixing was assessed by computing the discrepancy index (`maxdiff`) by means of `bpcomp` and the minimum effective size with `tracecomp`. A consensus tree was determined by means of `readpb`, the burnin was 5000.

An MSA and a rooted tree determined as described were the input for FASTML [16]. The `JTT` substitution model and the `maximum likelihood` method were used for indel reconstruction. As a representative predecessor, we chose the most likely sequence related to the respective node of the phylogenetic tree. Nucleotide and amino acid sequences of synthesized genes for ancestral proteins are given in S2 Table.

Site directed mutagenesis and cloning

A list of all oligonucleotides used for cloning and site-directed mutagenesis is provided in S3 Table. The sc*pri*A gene from *S. coelicolor*, which served as a positive control in the *in vivo* complementation assays, was amplified from scPriA-pTYB4 (a gift of Dr. Matthias Wilmanns) by standard PCR, using the oligonucleotides 5′sc*pri*A_*Sph*I/3′sc*pri*A_ Stop_*Hind*III, and cloned into the pTNA vector [6] via the introduced restriction sites for *Sph*I and *Hind*III. The tm*trpF* gene from *T. maritima*, which served as a negative control in the *in vivo* complementation assays, was available in a pTNA vector from previous work [34].

The *his*A gene from *T. maritima* (tm*his*A) was amplified using the template pDS56/RBSII_*his*A [35] with the oligonucleotides 5′tm*his*A_*Nde*I/3′tm*his*A_*Not*I (pET21a) and 5′tm*his*A_*Sph*I/3′tm*his*A_Stopp_*Hind*III (pTNA) and subsequently cloned into pET21a (Stratagene) and pTNA vectors using the respective terminal restriction sites. The genomic DNA of *D. desulfuricans ssp. Desulfuricans* and *P. carbinolicus* were ordered from DSMZ (DSM2380 and DSM6949, respectively). The respective *his*A genes (dd*his*A and pc*his*A) were amplified in a standard PCR using the oligonucleotides 5′dd*his*A_*Nde*I/3′dd*his*A_*Xho*I and 5′pc*his*A_*Nde*I/3′pc*his*A_*Xho*I, respectively, and subsequently cloned into the pET24a vector (Stratagene) via the introduced restriction sites for *Nde*I and *Xho*I. For *in vivo* complementation assays both *his*A genes were cloned into the pTNA vector via the restriction sites for *Sph*I and *Hind*III. To this end, pc*his*A was amplified with the oligonucleotides 5′pc*his*A_*Sph*I and 3′pc*his*A_Stopp_*Hind*III, whereas in the case of dd*his*A an overlap extension PCR [36] was necessary to remove an intrinsic *Sph*I restriction site. This reaction was performed with the oligonucleotides 5′dd*his*A_*Sph*I, 3′dd*his*A_C516T, 5′dd*his*A_C516T, and 3′dd*his*A_Stopp_*Hind*III.

The genes coding for the reconstructed ancestors were optimized for their expression in *E. coli*, synthesized (LifeTechnologies), and cloned into the pTNA and pET24a vectors using the terminal restriction sites for *Sph*I and *Hind*III. In order to render pET24a compatible for cloning with *Sph*I, two QuikChange mutagenesis steps were performed: the *Nde*I restriction site of pET24a was replaced by a *Sph*I restriction site using the oligonucleotides 5′pET24a_*Nde*I_to_*Sph*I and 3′pET24a_*Nde*I_to_*Sph*I, whereas a *Sph*I restriction site remote from the multiple cloning site was removed using the oligonucleotides 5′pET24a_A536T and 3′pET24a_A536T. All gene constructs were entirely sequenced to exclude inadvertent mutations.

Heterologous expression and purification of recombinant proteins

Gene expression, harvesting of cells, and cell lysis were performed essentially as described [18]. The genes pc*his*A and dd*his*A were expressed in *E. coli* T7 Express cells (New England Biolabs) containing the pRARE helper plasmid [34]. The gene tm*his*A was expressed in *E. coli* BL21-CodonPlus-(DE3)-RIPL cells (Agilent Technologies). The genes for the reconstructed proteins were expressed in *E. coli* BL21-Gold (DE3) cells (Agilent Technologies). For purification of tmHisA, heat denaturation (70°C, 15 min) was performed to remove most of the host proteins. Soluble cell extracts were loaded onto a HisTrapFF crude column (5 mL; GE Healthcare), which had been equilibrated with 50 mM potassium phosphate, pH 7.5, 300 mM sodium chloride, and 10 mM imidazole. After washing with equilibration buffer, the bound protein was eluted by applying a linear gradient of 10–375 mM imidazole. Subsequently, fractions with pure protein were pooled and dialyzed twice against 50 mM Tris·HCl, pH 7.5. Before dialyzing the reconstructed proteins CA-Bact-HisA, CA-Prot-HisA, and CA-Act-HisA in the same manner, fractions containing the respective protein were loaded onto a Superdex75 column

(HiLoad 26/60, 320 mL, GE Healthcare) operated with 50 mM Tris·HCl, pH 7.5, and 50 mM sodium chloride at 4°C. In all cases, at least 1 mg protein was obtained per liter of culture. All proteins were more than 95% pure, as judged by SDS-PAGE.

Steady-state enzyme kinetics

The HisA reaction was measured spectrophotometrically at 300 nm and 25°C as described [6]. The TrpF reaction was followed at 25°C by a fluorimetric assay (excitation at 350 nm, emission at 400 nm) [37]. The substrate PRA was generated *in situ* from anthranilate and phosphoribosylpyrophosphate (PRPP) using 1 μM yeast anthranilate phosphoribosyl transferase. To assure a constant concentration of the unstable PRA during the individual TrpF activity measurements, a 30-fold molar excess of PRPP over anthranilate was used. The k_{cat} and K_M values for both reactions were determined by fitting the hyperbolic saturation curves with the Michaelis-Menten equation. For unknown reasons, the CA-Prot-HisA and CA-Bact-HisA proteins exhibited a strong hysteresis, both in the HisA and TrpF reaction. Therefore, entire progress curves were recorded starting with as many as five different initial substrate concentrations. The curves were analyzed with COSY [38] using the integrated Michaelis-Menten equation for progress curves of the HisA reaction and a Michaelis-Menten equation that includes product inhibition for progress curves of the TrpF reaction.

E. *coli* knockout strains

The *E*. *coli* Δ*his*A strain was generated according to a classical protocol [39]. In brief, an ampicillin resistance gene was integrated into an *E*. *coli* DY329 helper strain to replace the genomic *his*A gene with the aid of this strain's genetically encoded bacteriophage λ Red recombination system [40]. The resistance gene was then transferred to *E*. *coli* BW25113 via P1 phage transduction and replaced the genomic *his*A gene. The complete deletion of the *his*A gene was verified by sequencing. The *E*. *coli* Δ*his*AΔ*trp*F double deletion strain was generated from the Δ*his*A strain in the same manner, with the genomic *trp*F gene being replaced by a chloramphenicol resistance gene. The *E*. *coli* Δ*trp*F single deletion strain (*E*. *coli* JMB9r-m+Δ*trp*F) was available from previous work [41].

In vivo complementation assays

Complementation assays with pTNA_sc*pri*A, pTNA_tm*his*A, pTNA_tm*trp*F, pTNA_dd*his*A, and pTNA_pc*his*A, as well as with the pTNA constructs of the reconstructed ancestors CA-Act-HisA, CA-Prot-HisA, and CA-Bact-HisA were performed on M9 minimal medium agar plates. An identical experimental procedure was followed in all cases: First, the respective plasmid was used to transform either chemical competent Δ*his*A, Δ*trp*F, or Δ*his*AΔ*trp*F *E*. *coli* cells. Next, single colonies were picked in order to inoculate 5 mL of LB medium supplemented with 150 μg/mL ampicillin only (Δ*his*A cells) or with 150 μg/mL ampicillin and 30 μg/mL chloramphenicol (Δ*trp*F and Δ*his*AΔ*trp*F cells). After incubation at 37°C overnight, 5 mL of LB medium containing the respective resistance markers were inoculated (optical density of 0.1 at 600 nm) and incubated at 37°C until an optical density of about 1 at 600 nm was reached (corresponding to approximately 10^8 cells). Subsequently, the cells in 1 mL suspension were collected by centrifugation (4°C, 4000 *g*, 10 min) and washed three times with 1% NaCl. Finally, 1:10^5 and 1:10^4 dilutions were streaked out on M9 minimal medium agar plates containing 150 μg/mL ampicillin and incubated at 37°C.

Supporting Information

S1 Fig. Phylogenetic tree based on 103 HisA sequences. Each sequence consists of the concatenated sequences of a HisA, a HisF, and a HisH protein. The tree was determined using pb, which is part of the PhyloBayes package. Posteriori probabilities are given for the splits; the length of the bar at the top corresponds to 0.5 mutations per site. Names encode the phylogenetic lineage of the species, see S1 Table. The node that corresponds to the reconstructed common ancestor of Actinobacteria (CA-Act-HisA) is marked with a filled circle.
(PDF)

S2 Fig. Phylogenetic tree based on 81 HisA sequences. Each sequence set consists of the concatenated sequences of a HisA, a HisF, and a HisH protein. The tree was determined using pb, which is part of the PhyloBayes package. Posteriori probabilities are given for the splits; the length of the bar at the top corresponds to 0.5 mutations per site. Names encode the phylogenetic lineage of the species, see S1 Table. The nodes that correspond to the reconstructed common ancestor of Proteobacteria (CA-Prot-HisA) and Bacteria (CA-Bact-HisA) are marked with a filled circle.
(PDF)

S1 Table. Species names and their abbreviations. The data set for the determination of a phylogenetic tree consisted of concatenated sequences of one HisA, one HisF, and one HisH, originating from the species listed below. For each phylum, the number of sequences is given in brackets. In the list, each species name is followed by the abbreviation (in brackets) used to label leaves of phylogenetic trees. The first symbol of the abbreviation indicates the superkingdom, the next four groups of two characters each give phylum, class, order, family, and the last three characters indicate the species name. Additional numbers were added by the algorithm used to create the abbreviations [28] but have no meaning in this context.
(PDF)

S2 Table. Nucleotide and amino acid sequences of synthesized genes for ancestral proteins.
(PDF)

S3 Table. List of oligonucleotides used for cloning and site-directed mutagenesis.
(PDF)

Acknowledgments

We thank Dr. Josef Sperl for his help with cloning, Sonja Fuchs and Jeanette Ueckert for expert technical assistance, and two anonymous referees for their constructive comments.

Author Contributions

Conceived and designed the experiments: MGP BR RS RM. Performed the experiments: MGP BR. Analyzed the data: MGP RM RS RM. Contributed reagents/materials/analysis tools: MGP BR. Wrote the paper: MGP BR RS RM. Reconstructed sequences: RM.

References

1. Khersonsky O, Tawfik DS. Enzyme promiscuity: a mechanistic and evolutionary perspective. Annu Rev Biochem. 2010; 79:471–505. doi: 10.1146/annurev-biochem-030409-143718 PMID: 20235827

2. Voordeckers K, Brown CA, Vanneste K, van der Zande E, Voet A, Maere S, et al. Reconstruction of ancestral metabolic enzymes reveals molecular mechanisms underlying evolutionary innovation through gene duplication. PLoS Biol. 2012; 10(12):e1001446. doi: 10.1371/journal.pbio.1001446 PMID: 23239941

3. Gould SJ, Vrba ES. Exaptation-a missing term in the science of form. Paleobiology. 1982; 8(1):4–15.

4. Tomarev SI, Piatigorsky J. Lens crystallins of invertebrates-diversity and recruitment from detoxification enzymes and novel proteins. Eur J Biochem. 1996; 235(3):449–65. PMID: 8654388

5. Barve A, Wagner A. A latent capacity for evolutionary innovation through exaptation in metabolic systems. Nature. 2013; 500(7461):203–6. doi: 10.1038/nature12301 PMID: 23851393

6. Henn-Sax M, Thoma R, Schmidt S, Hennig M, Kirschner K, Sterner R. Two $(\beta\alpha)_8$-barrel enzymes of histidine and tryptophan biosynthesis have similar reaction mechanisms and common strategies for protecting their labile substrates. Biochemistry. 2002; 41(40):12032–42. PMID: 12356303

7. Mirkin BG, Fenner TI, Galperin MY, Koonin EV. Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. BMC Evol Biol. 2003; 3(1):2.

8. Barona-Gómez F, Hodgson DA. Occurrence of a putative ancient-like isomerase involved in histidine and tryptophan biosynthesis. EMBO Rep. 2003; 4(3):296–300. PMID: 12634849

9. Due AV, Kuper J, Geerlof A, von Kries JP, Wilmanns M. Bisubstrate specificity in histidine/tryptophan biosynthesis isomerase from *Mycobacterium tuberculosis* by active site metamorphosis. Proc Natl Acad Sci U S A. 2011; 108(9):3554–9. doi: 10.1073/pnas.1015996108 PMID: 21321225

10. Kuper J, Dönges C, Wilmanns M. Two-fold repeated $(\beta\alpha)_4$ half-barrels may provide a molecular tool for dual substrate specificity. EMBO Rep. 2005; 6:134–9. PMID: 15654319

11. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997; 25 (17):3389–402. PMID: 9254694

12. Söderholm A, Guo X, Newton MS, Evans GB, Näsvall J, Patrick WM, et al. Two-step ligand binding in a $(\beta\alpha)_8$ barrel enzyme: SUBSTRATE-BOUND STRUCTURES SHED NEW LIGHT ON THE CATALYTIC CYCLE OF HisA. J Biol Chem. 2015; 290(41):24657–68. doi: 10.1074/jbc.M115.678086 PMID: 26294764

13. Jürgens C, Strom A, Wegener D, Hettwer S, Wilmanns M, Sterner R. Directed evolution of a $(\beta\alpha)_8$-barrel enzyme to catalyze related reactions in two different metabolic pathways. Proc Natl Acad Sci U S A. 2000; 97(18):9925–30. PMID: 10944186

14. Boussau B, Blanquart S, Necsulea A, Lartillot N, Gouy M. Parallel adaptations to high temperatures in the Archaean eon. Nature. 2008; 456(7224):942–5. doi: 10.1038/nature07393 PMID: 19037246

15. Lartillot N, Lepage T, Blanquart S. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics. 2009; 25(17):2286–8. doi: 10.1093/bioinformatics/btp368 PMID: 19535536

16. Ashkenazy H, Penn O, Doron-Faigenboim A, Cohen O, Cannarozzi G, Zomer O, et al. FastML: a web server for probabilistic reconstruction of ancestral sequences. Nucleic Acids Res. 2012; 40(Web Server issue):W580–4. doi: 10.1093/nar/gks498 PMID: 22661579

17. Noda-García L, Juárez-Vázquez AL, Ávila-Arcos MC, Verduzco-Castro EA, Montero-Morán G, Gaytán P, et al. Insights into the evolution of enzyme substrate promiscuity after the discovery of $(\beta\alpha)_8$ isomerase evolutionary intermediates from a diverse metagenome. BMC Evol Biol. 2015; 15:107. doi: 10.1186/s12862-015-0378-1 PMID: 26058375

18. Reisinger B, Sperl J, Holinski A, Schmid V, Rajendran C, Carstensen L, et al. Evidence for the existence of elaborate enzyme complexes in the Paleoarchean era. J Am Chem Soc. 2014; 136(1):122–9. doi: 10.1021/ja4115677 PMID: 24364418

19. Perez-Jimenez R, Inglés-Prieto A, Zhao ZM, Sanchez-Romero I, Alegre-Cebollada J, Kosuri P, et al. Single-molecule paleoenzymology probes the chemistry of resurrected enzymes. Nat Struct Mol Biol. 2011; 18(5):592–6. doi: 10.1038/nsmb.2020 PMID: 21460845

20. Patrick WM, Matsumura I. A study in molecular contingency: glutamine phosphoribosylpyrophosphate amidotransferase is a promiscuous and evolvable phosphoribosylanthranilate isomerase. J Mol Biol. 2008; 377(2):323–36. doi: 10.1016/j.jmb.2008.01.043 PMID: 18272177

21. Näsvall J, Sun L, Roth JR, Andersson DI. Real-time evolution of new genes by innovation, amplification, and divergence. Science. 2012; 338(6105):384–7. doi: 10.1126/science.1226521 PMID: 23087246

22. Noda-García L, Camacho-Zarco AR, Medina-Ruíz S, Gaytán P, Carrillo-Tripp M, Fülöp V, et al. Evolution of substrate specificity in a recipient's enzyme following horizontal gene transfer. Mol Biol Evol. 2013; 30(9):2024–34. doi: 10.1093/molbev/mst115 PMID: 23800623

23. Patrick WM, Quandt EM, Swartzlander DB, Matsumura I. Multicopy suppression underpins metabolic evolvability. Mol Biol Evol. 2007; 24(12):2716–22. PMID: 17884825

24. Hunter S, Jones P, Mitchell A, Apweiler R, Attwood TK, Bateman A, et al. InterPro in 2011: new developments in the family and domain prediction database. Nucleic Acids Res. 2012; 40(Database issue): D306–12. doi: 10.1093/nar/gkr948 PMID: 22096229

25. Atkinson HJ, Morris JH, Ferrin TE, Babbitt PC. Using sequence similarity networks for visualization of relationships across diverse protein superfamilies. PLoS One. 2009; 4(2):e4345. doi: 10.1371/journal.pone.0004345 PMID: 19190775

26. Gerlt JA, Bouvier JT, Davidson DB, Imker HJ, Sadkhin B, Slater DR, et al. Enzyme Function Initiative-Enzyme Similarity Tool (EFI-EST): A web tool for generating protein sequence similarity networks. Biochim Biophys Acta. 2015; 1854(8):1019–37. doi: 10.1016/j.bbapap.2015.04.015 PMID: 25900361

27. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006; 22(13):1658–9. PMID: 16731699

28. Pürzer A, Grassmann F, Birzer D, Merkl R. Key2Ann: a tool to process sequence sets by replacing database identifiers with a human-readable annotation. J Integr Bioinform. 2011; 8(1):153.

29. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003; 13(11):2498–504. PMID: 14597658

30. Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T. Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics. 2011; 27(3):431–2. doi: 10.1093/bioinformatics/btq675 PMID: 21149340

31. Plach MG, Löffler P, Merkl R, Sterner R. Conversion of anthranilate synthase into isochorismate synthase: implications for the evolution of chorismate-utilizing enzymes. Angewandte Chemie. 2015; 54(38):11270–4. doi: 10.1002/anie.201505063 PMID: 26352034

32. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013; 30(4):772–80. doi: 10.1093/molbev/mst010 PMID: 23329690

33. Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. 2000; 17(4):540–52. PMID: 10742046

34. Claren J, Malisi C, Höcker B, Sterner R. Establishing wild-type levels of catalytic activity on natural and artificial $(\beta\alpha)_8$-barrel protein scaffolds. Proc Natl Acad Sci U S A. 2009; 106(10):3704–9. doi: 10.1073/pnas.0810342106 PMID: 19237570

35. Thoma R, Obmolova G, Lang DA, Schwander M, Jeno P, Sterner R, et al. Efficient expression, purification and crystallisation of two hyperthermostable enzymes of histidine biosynthesis. FEBS Lett. 1999; 454(1–2):1–6. PMID: 10413084

36. Ho SN, Hunt HD, Horton RM, Pullen JK, Pease LR. Site-directed mutagenesis by overlap extension using the polymerase chain reaction. Gene. 1989; 77(1):51–9. PMID: 2744487

37. Hommel U, Eberhard M, Kirschner K. Phosphoribosyl anthranilate isomerase catalyzes a reversible Amadori reaction. Biochemistry. 1995; 34(16):5429–39. PMID: 7727401

38. Eberhard M. A set of programs for analysis of kinetic and equilibrium data. Comput Appl Biosci. 1990; 6 (3):213–21. PMID: 2207745

39. Datsenko KA, Wanner BL. One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. Proc Natl Acad Sci U S A. 2000; 97(12):6640–5. PMID: 10829079

40. Yu D, Ellis HM, Lee EC, Jenkins NA, Copeland NG, Court DL. An efficient recombination system for chromosome engineering in *Escherichia coli*. Proc Natl Acad Sci U S A. 2000; 97(11):5978–83. PMID: 10811905

41. Sterner R, Dahm A, Darimont B, Ivens A, Liebl W, Kirschner K. $(\beta\alpha)_8$-barrel proteins of tryptophan biosynthesis in the hyperthermophile *Thermotoga maritima*. EMBO J. 1995; 14(18):4395–402. PMID: 7556082

Supporting Information for Publication A

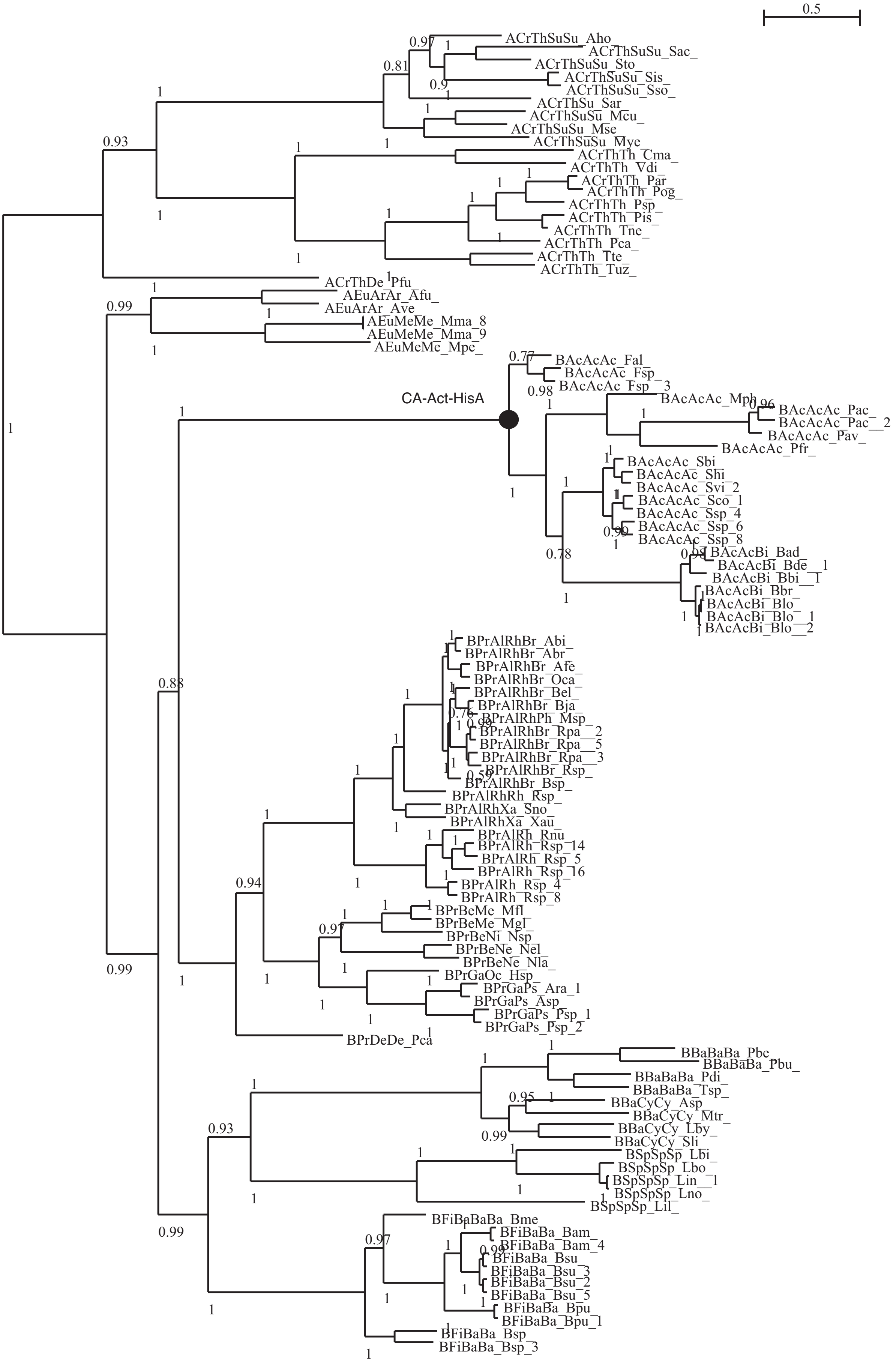
Long-term persistence of bi-functionality contributes to the robustness of microbial life through exaptation

Maximilian G. Plach*, Bernd Reisinger*, Reinhard Sterner, and Rainer Merkl (2016).

* Equal contributions

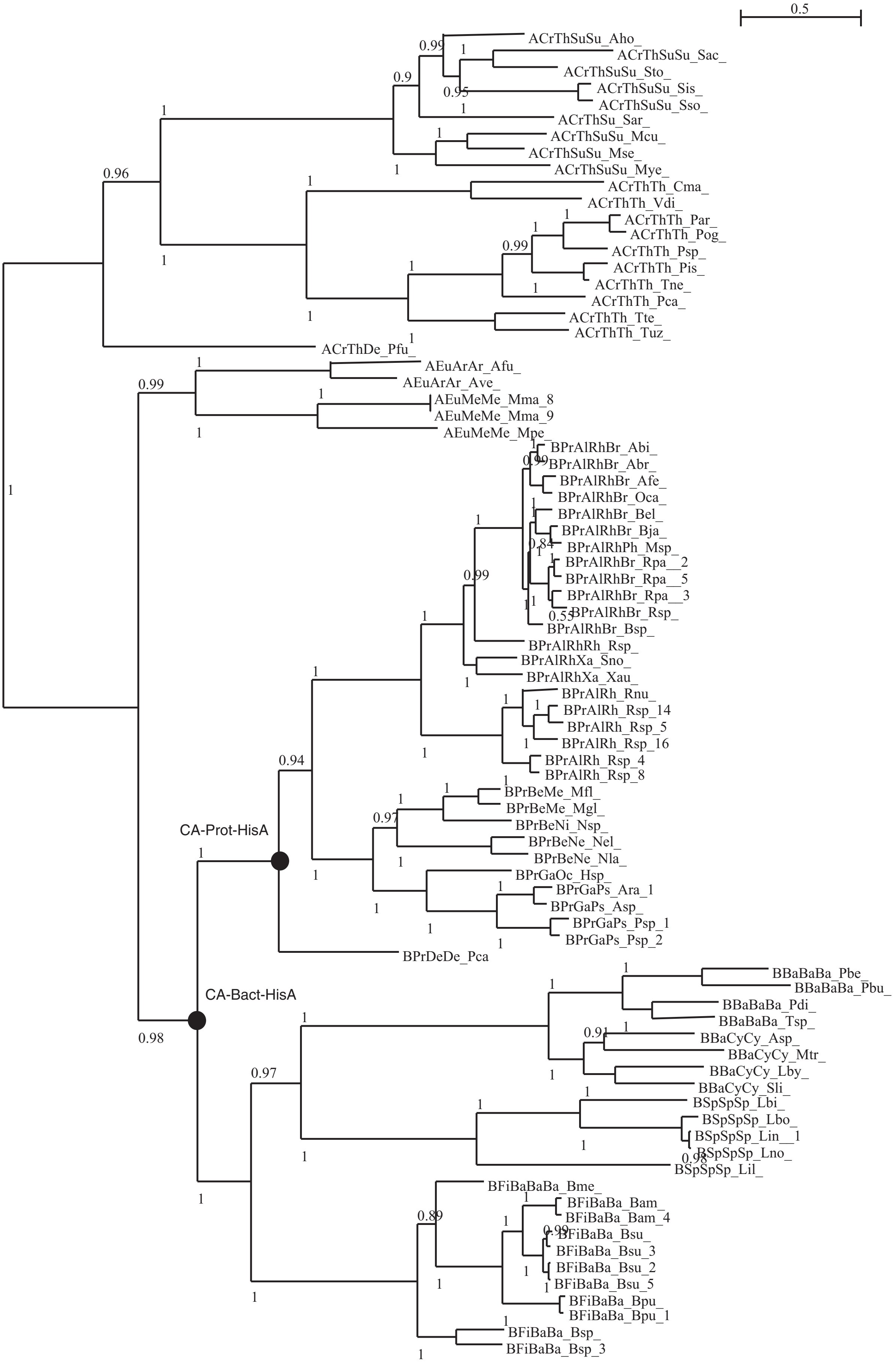
PLoS Genetics 12:e1005836

Supplementary Figure 1



.

Supplementary Figure 2



S1 Table: Species names and their abbreviations

The data set for the determination of a phylogenetic tree consisted of concatenated sequences of one HisA, one HisF, and one HisH, originating from the species listed below. For each phylum, the number of sequences is given in brackets. In the list, each species name is followed by the abbreviation (in brackets) used to label leaves of phylogenetic trees. The first symbol of the abbreviation indicates the superkingdom, the next four groups of two characters each give phylum, class, order, family, and the last three characters indicate the species name. Additional numbers were added by the algorithm used to create the abbreviations [1] but have no meaning in this context.

Crenarchaeota (20): *Vulcanisaeta distributa* (ACrThTh_Vdi), *Pyrolobus fumarii* (ACrThDe_Pfu), *Acidianus hospitalis* (ACrThSuSu_Aho), *Metallosphaera cuprina* (ACrThSuSu_Mcu), *Metallosphaera sedula* (ACrThSuSu_Mse), *Metallosphaera yellowstonensis* (ACrThSuSu_Mye), *Sulfolobus acidocaldarius* (ACrThSuSu_Sac), *Sulfolobus islandicus* (ACrThSuSu_Sis), *Sulfolobus solfataricus* (ACrThSuSu_Sso), *Sulfolobus tokodaii* (ACrThSuSu_Sto), *Sulfolobales archaeon Acd1* (ACrThSu_Sar), *Caldivirga maquilingensis* (ACrThTh_Cma), *Pyrobaculum arsenaticum* (ACrThTh_Par), *Pyrobaculum calidifontis* (ACrThTh_Pca), *Pyrobaculum islandicum* (ACrThTh_Pis), *Pyrobaculum oguniense* (ACrThTh_Pog), *Pyrobaculum sp. 1860* (ACrThTh_Psp), *Thermoproteus neutrophilum* (ACrThTh_Tne), *Thermoproteus tenax* (ACrThTh_Tte), *Thermoproteus uzoniensis* (ACrThTh_Tuz)

Euryarchaeota (5): *Archaeoglobus veneficus* (AEuArAr_Ave), *Methanoculleus marisnigri* (AEuMeMe_Mma_8), *Methanoculleus marisnigri* (AEuMeMe_Mma_9), *Methanoplanus petrolearius* (AEuMeMe_Mpe), *Archaeoglobus fulgidus* (AEuArAr_Afu)

Bacterioidetes (8): *Prevotella buccae* (BBaBaBa_Pbu), Parabacteroides distasonis (BBaBaBa_Pdi), *Tannerella sp. 6_1_58FAA_CT1* (BBaBaBa_Tsp), *Algoriphagus machipongonensis* (BBaCyCy_Asp), *Leadbetterella byssophila* (BBaCyCy_Lby), *Marivirga tractuosa* (BBaCyCy_Mtr), *Spirosoma linguale* (BBaCyCy_Sli), *Prevotella bergensis* (BBaBaBa_Pbe)

Firmicutes (11): *Bacillus amyloliquefaciens* (BFiBaBa_Bam), *Bacillus amyloliquefaciens* (BFiBaBa_Bam_4), *Bacillus pumilus* (BFiBaBa_Bpu), *Bacillus pumilus* (BFiBaBa_Bpu_1), *Bacillus* sp. NRRL B-14911 (BFiBaBa_Bsp), *Bacillus sp. 1NLA3E* (BFiBaBa_Bsp_3), *Bacillus subtilis* (BFiBaBa_Bsu), *Bacillus subtilis* (BFiBaBa_Bsu_2), *Bacillus subtilis* (BFiBaBa_Bsu_3), *Bacillus subtilis* (BFiBaBa_Bsu_5), *Bacillus* megaterium (BFiBaBaBa_Bme)

Spirochaetes (5): *Leptospira biflexa* (BSpSpSp_Lbi), *Leptospira* borgpetersenii (BSpSpSp_Lbo), *Leptonema illini* (BSpSpSp_Lil), *Leptospira* (BSpSpSp_Lin__1), *Leptospira noguchii* (BSpSpSp_Lno)

Proteobacteria (32): *Afipia birgiae* (BPrAlRhBr_Abi), *Afipia bromae* (BPrAlRhBr_Abr), *Afipia felis* (BPrAlRhBr_Afe), *Bradyrhizobium elkanii* (BPrAlRhBr_Bel), *Bradyrhizobium japonicum* (BPrAlRhBr_Bja), *Bradyrhizobium sp. BTAi1* (BPrAlRhBr_Bsp), *Oligotropha carboxidovorans* (BPrAlRhBr_Oca), *Rhodopseudomonas palustris* (BPrAlRhBr_Rpa__*2), Rhodopseudomonas palustris* (BPrAlRhBr_Rpa__*3), Rhodopseudomonas palustris* (BPrAlRhBr_Rpa__*5), Rhodopseudomonas sp. B29* (BPrAlRhBr_Rsp), *Bradyrhizobium sp. WSM4349* (BPrAlRhPh_Msp), *Rhodovulum sp. PH10* (BPrAlRhRh_Rsp), *Starkeya novella* (BPrAlRhXa_Sno), *Xanthobacter autotrophicus* (BPrAlRhXa_Xau), *Roseovarius nubinhibens* (BPrAlRh_Rnu), *Roseovarius sp. TM1035* (BPrAlRh_Rsp_14), *Roseobacter sp. AzwK-3b* (BPrAlRh_Rsp_16), *Rhodobacter sphaeroides* (BPrAlRh_Rsp_4), *Roseovarius sp. 217* (BPrAlRh_Rsp_5), *Rhodobacter sphaeroides* (BPrAlRh_Rsp_8), *Methylovorus* (BPrBeMe_Mgl), *Neisseria elongata* (BPrBeNe_Nel), *Neisseria lactamica* (BPrBeNe_Nla), *Nitrosomonas sp. AL212* (BPrBeNi_Nsp), *Methylobacillus flagellatus* (BPrBeMe_Mfl), *Pelobacter carbinolicus* (BPrDeDe_Pca), *Psychrobacter sp. PRwf-1* (BPrGaPs_Psp_2), *Halomonas* (BPrGaOc_Hsp), *Acinetobacter* (BPrGaPs_Ara_1), *Acinetobacter* (BPrGaPs_Asp), *Psychrobacter sp. 1501* (BPrGaPs_Psp_1)

Actinobacteria (22): *Propionibacterium avidum* (BAcAcAc_Pav), *Propionibacterium acnes* (BAcAcAc_Pac__2), *Propionibacterium acnes* (BAcAcAc_Pac), *Propionibacterium freudenreichii* (BAcAcAc_Pfr), *Microlunatus phosphovorus* (BAcAcAc_Mph), *Bifidobacterium longum* (BAcAcBi_Blo__2), *Bifidobacterium longum* (BAcAcBi_Blo__1), *Bifidobacterium longum* (BAcAcBi_Blo), *Bifidobacterium breve* (BAcAcBi_Bbr), *Bifidobacterium bifidum* (BAcAcBi_Bbi__1), *Bifidobacterium dentium* (BAcAcBi_Bde__1), *Bifidobacterium adolescentis* (BAcAcBi_Bad), *Streptomyces sp. C* (BAcAcAc_Ssp_8), *Streptomyces sp. Mg1* (BAcAcAc_Ssp_6), *Streptomyces sp. e14* (BAcAcAc_Ssp_4), *Streptomyces coelicolor* (BAcAcAc_Sco_1), *Streptomyces violaceusniger* (BAcAcAc_Svi_2), *Streptomyces himastatinicus* (BAcAcAc_Shi), *Streptomyces bingchenggensis* (BAcAcAc_Sbi), *Frankia sp. EUN1f* (BAcAcAc_Fsp__3), Frankia sp. EAN1pec (BAcAcAc_Fsp), *Frankia alni* (BAcAcAc_Fal)

S2 Table. Nucleotide and amino acid sequences of synthesized genes for ancestral proteins

Nucleotide sequence for CA-Act-HisA (restriction sites for *Sph*I and *Hind*III are underlined):

GCATGCTCACCCTGCTGCCTGCAGTTGATGTTCGTGATGGTCAGGCAGTTCGTCTGGTTCAGGGTGAA
GCAGGTAGCGAAACCAGCTATGGTGATCCGCTGGAAGCAGCACGTACCTGGCAAGAGGATGGTGCAGA
ATGGATTCATCTGGTTGATCTGGATGCAGCATTTGGTCGTGGTAGCAATCGTGAACTGATTGCCGAAG
TTGTTCGTGCCGTTGATGTTAATGTTGAACTGAGCGGTGGTATTCGTGATGATGATAGCCTGGATGCC
GCACTGGCAACCGGTGCAGCACGTGTTAATATTGGCACCGCAGCACTGGAAAATCCGGAATGGGTTCG
TAAAGTTATTGATCGTTATGGCGATCGTATTGCAGTTGGTCTGGATGTGCGTGGCACCACCCTGGCAG
CCCGTGGTTGGACCCGTGATGGTGGTGAACTGTTTGAAGTTCTGGCACGTCTGGATGCGGCAGGTTGT
GCACGTTATGTTGTTACCGATGTTGCACGTGATGGTATGCTGACCGGTCCGAATGTGGAACTGCTGCG
TGAAGTTACCGCAGCAACCGATCGTCCGGTTGTTGCAAGCGGTGGTGTTAGCAGTCTGGATGATCTGC
GTGCACTGGCAGCGCTGGTTCCGGAAGGTGTTGAAGGTGCAATTGTTGGTAAAGCACTGTATGCCGGT
GCATTTACCCTGCCGGAAGCACTGGCCGTTGCACGTAAGCTT

Nucleotide sequence for CA-Prot-HisA (restriction sites for *Sph*I and *Hind*III are underlined):

GCATGCTGATTATTCCGGCAATCGATCTGAAAGATGGTCGTTGTGTTCGTCTGGAACAGGGTGATATG
GAAAAAGCAACCGTGTATAATGATGATCCGGCAGCAATGGCACGTCAGTGGGTTGAGCAGGGTGCAGA
ATGGCTGCATCTGGTTGATCTGGATGGTGCATTTGCAGGTAAACCGGTTAATGAAGATGCCATTAAAG
CAATTGCAGAAGCAGTTAGCATTCCGGTTCAGCTGGGTGGTGGTATTCGTGATCTGGAAACCATTGAA
GCATATCTGGAAGCAGGTATTGATCGTGTTATTATTGGCACCGTTGCAGTGAAAAATCCGGAACTGGT
TCGTGAAGCATGTCGTGCATTTCCGGGTCGTATTGTTGTTGGTATTGATGCACGTGATGGTATGGTTG
CAGTTAAAGGTTGGGCAGAAGTTACCGAAGTTAAAGCCACCGATCTGGCCAAACGTTTTGAAGATGCG
GGTGTTGAAGCAATCATTTATACCGATATTGCCCGTGATGGCATGATGCAGGGTCCGAATATTGAAGC
AACCCGTGCACTGGCAAAAGCAGTTTCAATTCCGGTTATTGCAAGCGGTGGTGTTAGCAGCCTGGAAG
ATATCGAAGCACTGCTGGCAATTGAAGATAGCGGTGTGACCGGTGTTATTACCGGCAAAGCACTGTAT
GAAGGTAGCCTGGATCTGCGTGAAGCACTGGCACTGGCCAAAAAGCTT

Nucleotide sequence for CA-Bact-HisA (restriction sites for *Sph*I and *Hind*III are underlined):

GCATGCGCATTATTCCGGCAATCGATCTGAAAGATGGTCGTTGTGTTCGTCTGGTTCAGGGTGATATG
GAAAAAGCAACCGTGTATAATGATGATCCGCTGGAAATGGCAAAACAGTGGGTTGAACAGGGTGCAGA
ATGGCTGCATGTTGTTGATCTGGATGGTGCATTTGCAGGTAAACCGGTTAATGAAGATGTGATCAAAG
AAATCGCACAGAAAGTTAGCGTTAAAGTTCAGCTGGGTGGTGGTATTCGTGATCTGGAAGATATTGAA
GCATATCTGGATGCCGGTGTTGATCGTGTTATTATTGGCACCGTTGCAGTTAAAAATCCGGAACTGGT
TCGTGAAATGGTGGAAAAATATGGTGAACGTATTGTGGTTGGTATTGATGCACGTGATGGTATGGTTG

CCGTTAAAGGTTGGAAAGAAACCACCGAAGTTAAAGCCACCGATCTGGCCAAACGTTTTGAAGATGCA
GGCGTTGAAGCAATCATTTATACCGATATTGCCCGTGATGGCATGATGCAGGGTCCGAACATTGAAGC
CATTCGTGAACTGGCAAAAGCAGTTAGCCTGCCGGTTATTGCAAGCGGTGGTGTTAGCAGCCTGGAAG
ATATCGAGGCACTGCTGGCAATTGAAGAAAGCGGTGTTGCGGGTGTTATTGTTGGTAAAGCACTGTAT
GAAGGTCGTCTGGATCTGCGTGAAGCACTGGCACTGGCCAAAAAGCTT

Amino acid sequence of CA-Act-HisA:

MLTLLPAVDVRDGQAVRLVQGEAGSETSYGDPLEAARTWQEDGAEWIHLVDLDAAFGRGSNRELIAEV
VRAVDVNVELSGGIRDDDSLDAALATGAARVNIGTAALENPEWVRKVIDRYGDRIAVGLDVRGTTLAA
RGWTRDGGELFEVLARLDAAGCARYVVTDVARDGMLTGPNVELLREVTAATDRPVVASGGVSSLDDLR
ALAALVPEGVEGAIVGKALYAGAFTLPEALAVAR

Amino acid sequence of CA-Prot-HisA:

MLIIPAIDLKDGRCVRLEQGDMEKATVYNDDPAAMARQWVEQGAEWLHLVDLDGAFAGKPVNEDAIKA
IAEAVSIPVQLGGGIRDLETIEAYLEAGIDRVIIGTVAVKNPELVREACRAFPGRIVVGIDARDGMVA
VKGWAEVTEVKATDLAKRFEDAGVEAIIYTDIARDGMMQGPNIEATRALAKAVSIPVIASGGVSSLED
IEALLAIEDSGVTGVITGKALYEGSLDLREALALAK

Amino acid sequence of CA-Bact-HisA:

MRIIPAIDLKDGRCVRLVQGDMEKATVYNDDPLEMAKQWVEQGAEWLHVVDLDGAFAGKPVNEDVIKE
IAQKVSVKVQLGGGIRDLEDIEAYLDAGVDRVIIGTVAVKNPELVREMVEKYGERIVVGIDARDGMVA
VKGWKETTEVKATDLAKRFEDAGVEAIIYTDIARDGMMQGPNIEAIRELAKAVSLPVIASGGVSSLED
IEALLAIEESGVAGVIVGKALYEGRLDLREALALAK

S3 Table. List of oligonucleotides used for cloning and site-directed mutagenesis

Name	**Sequence**
5`dd*his*A_*Nde*I	ccgtatcatatgattatttttcccgctgttg
3`dd*his*A_*Xho*I	agccccctcgaggcgctttcgggcgtgtttttc
5`dd*his*A_*Sph*I	ggagatggcatgctgatttttcccgctg
3`dd*his*A_Stopp_*Hind*III	tgccaagcttcaggcgctttcgggcgtg
5`dd*his*A_C516T	attgaacgcgacgg**t**atgcagtgcggc
3`dd*his*A_C516T	gccgcactgcat**a**ccgtcgcgttcaat
5`pc*his*A_*Nde*I	ccgtatcatatgatagttattcccgccatcg
3`pc*his*A_*Xho*I	agccccctcgaggctttgcccttggtcagagcc
5`pc*his*A_*Sph*I	ggagatggcatgctggttattcccgcca
3`pc*his*A_Stopp_*Hind*III	tgccaagcttcaggctttgcccttggt
5`tm*his*A_*Nde*I	ccgtatcatatgctcgttgtcccggcgat
3`tm*his*A_*Not*I	atagcggccgcgcgagcatatctcttcatcac
5`tm*his*A_*Sph*I	atagcatgctcgttgtcccggcg
3`tm*his*A_Stopp_*Hin*dIII	tgccaagctttagcgagcatatct
5`sc*pri*A_*Sph*I	ggagatggcatgcgcaagctcgaactc
3`sc*pri*A_ Stopp_*Hind*III	tgccaagcttcacgacgtagcctccaa
5`pET24a_*Nde*I_to_*Sph*I	ccaccagtcatgctagcc**gc**atgcatatctccttcttaaag
3`pET24a_*Nde*I_to_*Sph*I	ctttaagaaggagatatgcat**gc**ggctagcatgactggtgg
5`pET24a_A536T	cgccatctccttgc**t**tgcaccattccttg
3`pET24a_A536T	caaggaatggtgca**a**gcaaggagatggcg

Restriction sites are underlined; modified bases are in bold.

7.2 Publication B

Conversion of anthranilate synthase into isochorismate synthase: Implications for the evolution of chorismate-utilizing enzymes

Maximilian G. Plach, Patrick Löffler, Rainer Merkl, and Reinhard Sterner (2015).

Angewandte Chemie International Edition 54:11270-11274

Enzymes

International Edition: DOI: 10.1002/anie.201505063
German Edition: DOI: 10.1002/ange.201505063

Conversion of Anthranilate Synthase into Isochorismate Synthase: Implications for the Evolution of Chorismate-Utilizing Enzymes

*Maximilian G. Plach, Patrick Löffler, Rainer Merkl, and Reinhard Sterner**

Abstract: *Chorismate-utilizing enzymes play a vital role in the biosynthesis of metabolites in plants as well as free-living and infectious microorganisms. Among these enzymes are the homologous primary metabolic anthranilate synthase (AS) and secondary metabolic isochorismate synthase (ICS). Both catalyze mechanistically related reactions by using ammonia and water as nucleophiles, respectively. We report that the nucleophile specificity of AS can be extended from ammonia to water by just two amino acid exchanges in a channel leading to the active site. The observed ICS/AS bifunctionality demonstrates that a secondary metabolic enzyme can readily evolve from a primary metabolic enzyme without requiring an initial gene duplication event. In a general sense, these findings add to our understanding how nature has used the structurally predetermined features of enzyme superfamilies to evolve new reactions.*

Chorismate (CH) is a central metabolic branch-point molecule and the common precursor of essential primary (folate, tryptophan) and important secondary (menaquinones, siderophores, antibiotics) metabolites that are vital for plants as well as free living and infectious microorganisms[1] (Figure 1). The CH-related pathways are therefore notable targets for antimicrobials and herbicides. The enzymes catalyzing the committed steps of these pathways share a common fold and use similar reaction mechanisms. Presumably, they originated from a common ancestor and have therefore been grouped together as the MST (*m*enaquinone, *s*iderophores, *t*ryptophan) superfamily.[2] Within this superfamily, the primary metabolic anthranilate and aminodeoxychorismate synthases (AS, ADCS) employ ammonia as a nucleophile to form aminated chorismate derivatives, whereas the secondary metabolic isochorismate and salicylate synthases (ICS, SS) use water as a nucleophile to hydroxylate chorismate (Figure 1). These two subfamilies are hereafter termed ammonia-utilizing and water-utilizing MST enzymes, respectively (AMEs, WMEs). Based on the assumption that each enzyme of secondary metabolism stems from an enzyme of primary metabolism,[3] the subdivision of the MST superfamily suggests that a transition of nucleophile specificity from ammonia to water underlay the evolution of WMEs (ICS, SS) from AME (AS, ADCS) ancestors. We retraced this putative evolutionary path by identifying the residues that contribute to nucleophile specificity in MST enzymes and by subsequently establishing ICS activity on an AS scaffold.

The AS from *Salmonella typhimurium* (stAS) is a heterotetramer comprising two synthase and two glutaminase subunits (stTrpE and stTrpG, respectively).[4] Glutamine is hydrolyzed in the active site of stTrpG to yield nascent ammonia, which is subsequently channeled to the active site of stTrpE.[4,5] To identify the residues of stTrpE that are involved in this channeling and that may therefore come into contact with the ammonia nucleophile, we applied MOLE 2.0, a program for analyzing macromolecular channels.[6] We identified a 30 Å-long channel connecting the active sites in stAS (Figure 2a). This channel is similar to the channel observed in the crystal structure of the homologous aminodeoxyisochorismate (ADIC) synthase PhzE.[7] As the channel approaches the CH ligand, it is predominantly shaped by three residues: Gln263 in β-strand 11 as well as Met364 and Leu365 in α-helix 12. To estimate whether the nature of these residues correlates with nucleophile specificity in AMEs and WMEs, we computed individual multiple sequence alignments (MSAs) of ADCS, AS, ICS, and SS. Notably, these sequences formed four distinct subtrees in a phylogenetic analysis (Figure S1 in the Supporting Information), which makes them representative for their MST groups. Furthermore, all of the MSAs contained sequences of at least four major phyla. The resulting sequence logos of β-strand 11 and α-helix 12 (Figure 2b) confirmed the strict conservation of Gln263 in AS and of Lys at the corresponding position in ICS and SS, which has been noted previously.[2] It was further shown that this Lys acts as a catalytic base for water activation in ICS and SS.[2,8] In a number of ADCS sequences, Gln263 is substituted by Glu. Residues 364 and 365 are conserved to a large extent within AMEs and WMEs, but clearly differ between the two groups. In AMEs, Met364 is strictly conserved, as are Leu365 in AS and Ile365 in ADCS. In WMEs however, several hydrophobic residues (Leu, Ile, Phe, Val) are abundant at position 364, as are Val in ICS and Ser in SS at position 365. Other positions were not considered since they are either conserved throughout all four MST groups, not conserved within AMEs or WMEs, or not involved in forming the nucleophile channel in stAS.

Based on this analysis, we attempted to shift the nucleophile specificity of stAS from ammonia to water by mutating residues 263, 364, and 365 of the stTrpE subunit. For this purpose, 16 variants were generated and assayed by HPLC for the formation of reaction products starting from CH. The WME-typical catalytic Lys replaced the wild-type Gln263 in all of the variants and was combined with the different residues 364/365 found in WMEs (Table S1). The variants are hereafter denoted by their residues 263–364–365 combination

[*] M. G. Plach, P. Löffler, Prof. Dr. R. Merkl, Prof. Dr. R. Sterner
Institut für Biophysik und physikalische Biochemie
Universität Regensburg, 93040 Regensburg (Germany)
E-mail: Reinhard.Sterner@ur.de

Supporting information for this article is available on the WWW under http://dx.doi.org/10.1002/anie.201505063.

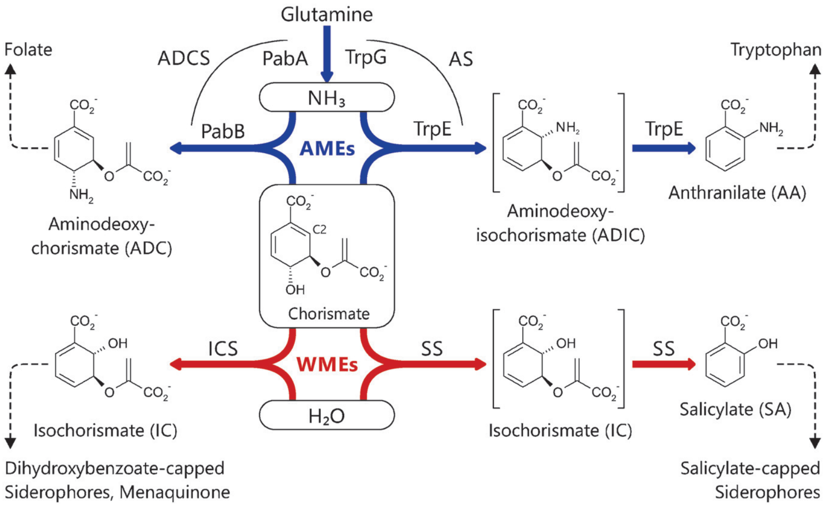


Figure 1. Chorismate is a central branch point between primary and secondary metabolism. It is converted into aminodeoxychorismate (ADC) and anthranilate (AA) by the primary metabolic ammonia-utilizing MST enzymes (AMEs) ADCS and AS. Both enzymes are heteromeric complexes comprising glutaminase (PabA, TrpG) and synthase (PabB, TrpE) subunits. ADC and AA are part of folate and tryptophan biosynthesis, respectively. The water-utilizing MST-enzymes (WMEs) ICS and SS convert chorismate into isochorismate (IC) and salicylate (SA), respectively. Both products are precursors of important secondary metabolites like iron-chelating siderophores and respiratory-chain components (menaquinone). A special characteristic of AS and SS is that pyruvate is eliminated from the reaction intermediates aminodeoxyisochorismate (ADIC) and IC to yield aromatic products.

(e.g. KML for the variant with Lys263, Met364, and Leu365). All variants expressed as soluble proteins and no adverse effects on protein stability were observed compared to wild-type stTrpE.

Twelve of these variants formed IC in the absence of an ammonia source (Figure 3a and Figure S2a). The product was unambiguously identified by ESI mass spectrometry and enzymatic conversion to SA (Figure S3). All twelve variants feature the Gln263Lys substitution. In three of these variants, the Gln263Lys substitution is combined with a mutation at position 365 (KMV, KMS, KMA). Therefore, just two mutations are sufficient for the establishment of IC specificity on the stTrpE scaffold. In nine of the variants, an additional mutation is present at position 364. Averaged over the 12 variants, 20% of supplied CH was converted to IC; the best variant, KIA, converted 37.1%. In comparison, the ICS EntC from *Escherichia coli* formed 31.5% IC, with the observed incomplete conversion being caused by equilibrium between CH and IC.[9] Upon replacing Lys263 by Ala in the IC-forming KLS variant, the resulting ALS variant became inactive, pointing to a catalytic role for Lys263 in our stAS variants similar to that in native ICS and SS.[2,8] Furthermore, no IC was formed in control reactions with wild-type stTrpE or in the absence of any enzyme.

For the KML variant, which carries only the single Gln263Lys replacement, neither IC nor SA was detected, thus indicating that Lys263 alone is not sufficient for the use of water as a nucleophile. Accordingly, the Abell group could also not detect IC or SA when characterizing the equivalent single Gln263Lys replacement in the AS from *Serratia marcescens*.[10] Ziebart and Toney reported in a comprehensive study on nucleophile specificity in MST enzymes that the Gln263Lys variant of stAS (equivalent to our KML variant) formed traces of IC and SA.[11] However, only 0.008% and 0.03% of CH were converted to IC and SA, respectively, under conditions comparable to our experimental setup. In any case, the central finding of our work, namely inversion of the nucleophile specificity of stAS through a few amino acid replacements, as exemplified by the KAS and KAA variants, is independent of whether the single Gln263Lys replacement leads to no or extremely low amounts of IC or SA.

It is worth noting that only four of the ICS-active stAS variants also formed SA, which involves the elimination of pyruvate from IC. This finding seems counterintuitive at first because the generic product of stAS is AA, which has, like SA, undergone elimination of pyruvate. However, it was recently shown that the elimination of pyruvate from chorismate-derived reaction intermediates in MST enzymes is controlled by the conformation of the chorismate ring structure.[12] Therefore, the most plausible explanation for the lack of pyruvate-elimination activity in most ICS-active stAS variants is perturbations in the IC ring conformation caused by the introduced mutations.

To examine the effects of the mutations on the use of the generic nucleophile ammonia, we investigated product formation by the stAS variants in the presence of glutamine (Figure 3b and Figure S2b). Under these conditions, all twelve ICS-active variants formed not only IC but also AA and its precursor ADIC, thus implying a broadened nucleophile specificity. Interestingly, while forming substantial amounts of AA, some variants showed only a modest reduction in IC formation compared to the absence of glutamine (KMV: 2.5-fold, KIV: 1.4-fold). Other variants, while also forming considerable amounts of AA, even showed an increase in IC formation upon the addition of glutamine (e.g., KAV: 2.2-fold).

Independent of the presence or absence of an ammonia source, a striking increase in IC formation was observed for the residue sequence Leu→Val→Ser/Ala365. Variants with Leu365 are catalytically inactive, both for IC and AA formation. The presence of Val, Ser, and Ala instead leads to average CH to IC conversions of 8.2%, 28.1%, and 24.6%, respectively.

As pointed out before, residues 263, 364, and 365 line the nucleophile channel in stTrpE. We therefore assumed that differently sized residues at these positions might reshape the channel and thus lead to the different reactivities observed. To allow a statistically sound prediction of the general

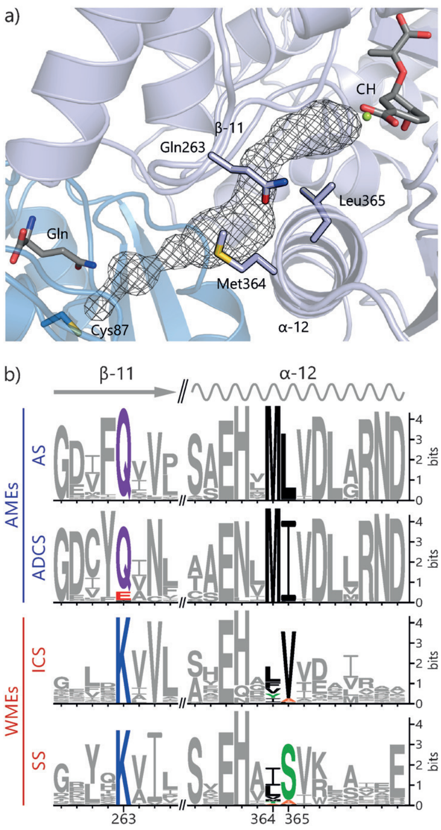


Figure 2. Location and conservation of residues characteristic for AMEs and WMEs. a) Nucleophile channel connecting the active sites of stTrpG (blue; represented by Cys87 and a glutamine ligand) and stTrpE (pale blue; represented by CH and a Mg^{2+} ion) in stAS (model based on PDB ID 1i7q). The channel-lining residues of stTrpE mutated in this work are shown as stick models. b) Sequence logos of β-strand 11 and α-helix 12. The β-strand contains a conserved Gln/Glu in AMEs and a strictly conserved Lys in WMEs. The α-helix contains the conserved Met–Leu and Met–Ile pairs in AS and ADCS, respectively. In ICS, this pair is mainly conserved as [Leu,Phe,Val]–[Val,Ala] and in SS as [Ile,Leu]–[Ser,Ala]. Residue numbering is based on stTrpE. Residues 263, 364, and 365 are colored by their chemical properties (purple: amido functionality, red: acidic, blue: basic, black: hydrophobic, green: hydroxyl functionality, orange: small).

nucleophile path in the stTrpE variants, we performed molecular dynamics (MD) simulations and applied MOLE to generate 600 putative nucleophile trajectories (PNTs) for each variant. A PNT is defined by the centerline of the corresponding MOLE channel.

We found that a reduction in the size of residue 365 leads to a shift in PNT localization in a 10 Å shell around CH (Figure 4a). In wild-type stTrpE, PNTs pass Met364 and Gln263, bend around Leu365, and end at C2 of CH, which is where the nucleophilic attack by ammonia takes place. The single Gln263Lys mutation in the KML variant does not alter this course. Since both KML and wild-type stTrpE feature Leu365, this course of PNTs was termed the L-path (Figure 4b). In the KML variant, however, the PNT-associated channels are clearly constricted between Lys263, Leu365, and Val265 owing to a change in the van der Waals volume of residue 263 from 114 $Å^3$ (Gln) to 135 $Å^3$ (Lys).[13] Ammonia or water therefore cannot access CH via these PNTs, thus resulting in the observed loss of activity in the Leu365 variants. Replacing Leu365 by Val (volume reduction from 124 $Å^3$ to 105 $Å^3$) has two consequences for the PNTs: The associated channels of L-path PNTs are enlarged and 31 % of the PNTs shift away from the L-path and approach CH from Val365 and Thr425. This effect becomes even more pronounced for the KMS variant with Ser365, the low volume (73 $Å^3$) of which allows 79 % of the PNTs to approach CH via this alternative path. Finally, Ala365 (67 $Å^3$) leads to a complete shift of PNTs to this path, which was therefore termed the A-path (Figure 4b).

For the other twelve stAS variants, MD simulations and computations of PNTs were performed in a similar manner (Figure S4, Table S2). The PNT distributions in these variants are consistent with those described above for the KML, KMV, KMS, and KMA variants. Taken together, it is evident that high CH to IC conversion rates correlate with high fractions of PNTs proceeding along the A-path (Figure 4c).

These findings provide a conclusive view on water utilization by WMEs. Ziebart and Toney suggested that two properties of WMEs promote the water-activating capacity of Lys263.[11] First, two hydrogen-bond acceptors were thought to assist in deprotonating the ε-amino group. However, one of these acceptors, a glutamic acid residue, is conserved not only in WMEs but in all but one of the 1222 MST enzymes in our dataset. The other, a serine residue corresponding to Ser365, is not conserved in SS (7 % Ala) and is not present at all in ICS. Furthermore, we could show that Ser365 is not required for IC formation (Figure 4c). The second element proposed was the residue corresponding to Met364 in stTrpE. It is conserved as a hydrophobic residue (Leu, Ile, Val, Phe) in WMEs and was thought to assist in the proper orientation of Lys263 by making van der Waals contact with the Lys methylene groups. However, variants with wild-type Met364 or Ala364 display IC specificity as well (Figure 4c), which argues against a role for residue 364 in positioning the catalytic Lys.

In summary, we have demonstrated that nucleophile specificity in MST enzymes is controlled by two factors: 1) the presence or absence of Lys263 as a catalytic base and 2) sufficient space for the nucleophile to reach CH. These findings agree well with the amino acid distribution in naturally occurring ICS and SS. Only five ICS and no SS in our MST dataset contain a residue larger than Val at position 365. Moreover, the assignment of Leu365 in those five ICS sequences is most likely due to a misalignment, since in all cases, Val directly precedes this Leu.

The straightforward establishment of ICS activity on the evolutionarily related AS scaffold indicates that the MST-superfamily represents a flexible and adaptive link between primary and secondary metabolism. It is generally accepted that each secondary metabolic pathway has its origin in mutations that accumulated in a primary metabolic gene. Consequently, these mutations should 1) enhance overall metabolic chemical diversity by yielding molecules with novel biological activities, 2) allow the retention of existing chemical variety, and 3) do so with minimum fitness costs.[3a] Our data strongly support this hypothesis. Only two mutations are necessary to establish ICS activity and all these stTrpE variants were still able to employ ammonia as a nucleophile to form AA for tryptophan biosynthesis. Following this reasoning, the evolution of IC formation does not necessarily require gene duplication but could have proceeded via a bifunctional intermediate similar to the stTrpE variants of this work.

Substrate or nucleophile channeling, which is crucial for enzymatic function in MST enzymes, is no rare phenomenon. More than 64 % of all known enzyme structures contain at least two channels reaching to their active sites.[14] Our approach may thus prove valuable for modifying the reactivities of such enzymes, which could ultimately yield novel enzymatic functions.

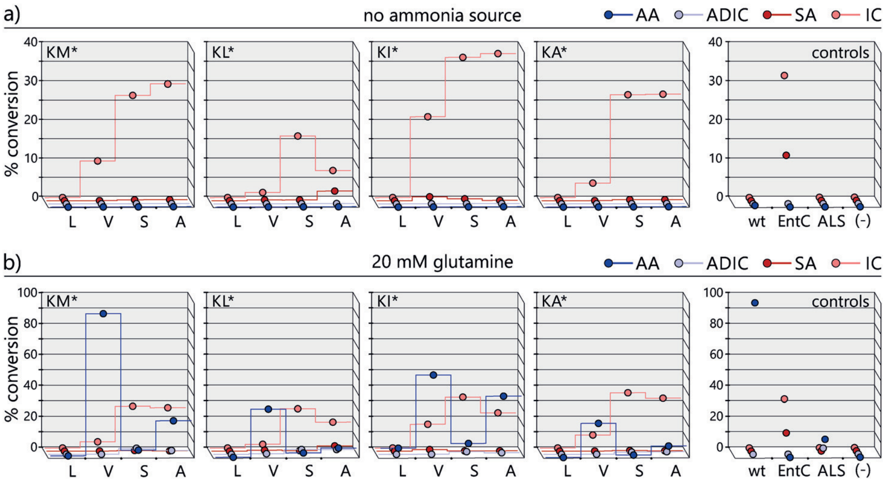


Figure 3. Product distributions of reactions of stTrpE variants with chorismate in the absence (a) and presence (b) of glutamine as an ammonia source. The variants share Lys263 and are grouped by residue 364 (e.g., KM* describes the four variants with Lys263, Met364, and either Leu, Val, Ser, or Ala at position 365). All stTrpE variants were assayed in the presence of stTrpG. The product distributions of control reactions are shown on the right: wild-type stTrpE (wt), ecEntC (EntC), Lys263Ala knockout of the KLS variant (ALS), and enzyme-free control (−). For each product, its fraction out of all of the products formed is shown (% conversion). Each data point represents the mean of at least three independent experiments. Error bars are omitted for clarity because the average and maximum absolute errors were only 0.3 % and 3.7 %, respectively. Step lines connecting data points were added to aid visual tracking of product-specific values.

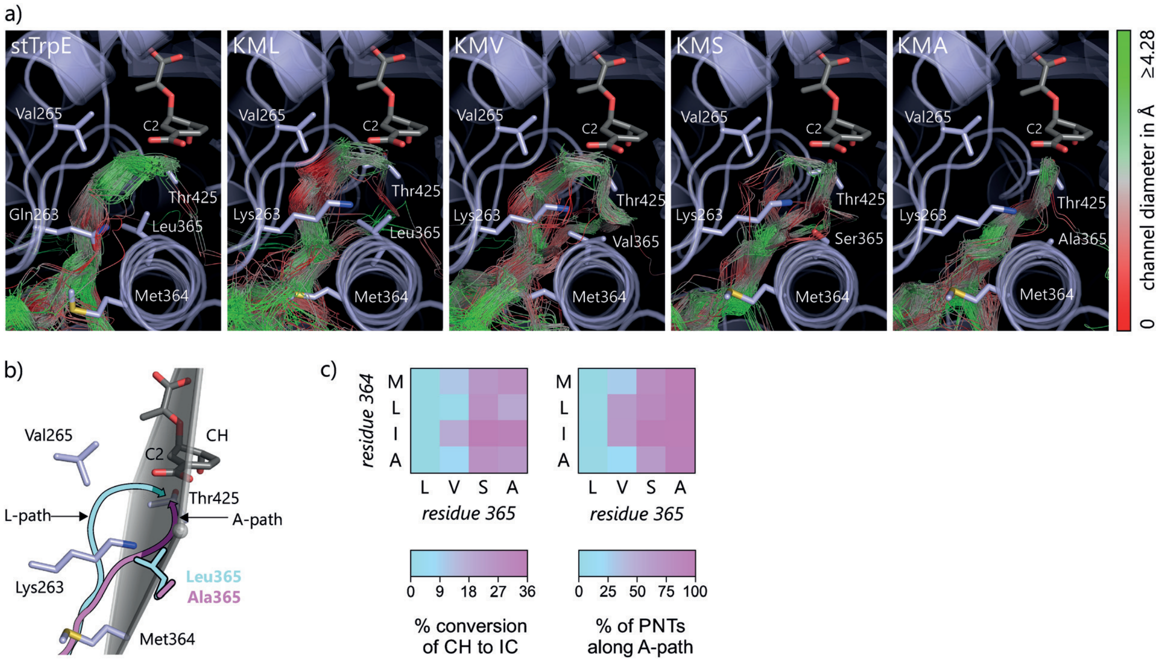


Figure 4. Visualization and quantitative analysis of PNTs in wild-type and mutant stTrpE variants. a) Comparison of PNTs in wild-type stTrpE and the KML, KMV, KMS, and KMA variants. The diameters of the PNT-associated channels are color-coded and important channel-lining residues are shown as stick models. The key observation is the correlation between decreasing size of residue 365 (Leu > Val > Ser > Ala) and a shift in PNT localization from the L-path to the A-path [see (b)]. b) The L-path (cyan) and A-path (magenta) reflect the majority of PNTs in variants with Leu365 and Ala365, respectively, and thus show the boundaries of the PNT shift. The directions from which the two paths approach CH can be separated by a plane specified by CH-C2 and the Cα-atoms of Met364 and Thr425 (for details see the methods section in the Supporting Information). c) Comparison of the average CH-to-IC conversion by all stTrpE variants with Lys263 and the fraction of PNTs proceeding along the A-path in these variants.

Acknowledgements

M.G.P. was supported by a PhD fellowship from the Fonds der Chemischen Industrie. We thank Kristina Heyn for help with phylogenetic analyses and Christiane Endres, Sonja Fuchs, and Jeannette Ueckert for expert technical assistance.

Keywords: chorismate · enzyme catalysis · enzyme evolution · nucleophile specificity · protein design

How to cite: *Angew. Chem. Int. Ed.* **2015**, *54*, 11270–11274
Angew. Chem. **2015**, *127*, 11423–11427

[1] F. Dosselaere, J. Vanderleyden, *Crit. Rev. Microbiol.* **2001**, *27*, 75–131.

[2] S. Kolappan, J. Zwahlen, R. Zhou, J. J. Truglio, P. J. Tonge, C. Kisker, *Biochemistry* **2007**, *46*, 946–953.

[3] a) R. D. Firn, C. G. Jones, *Mol. Microbiol.* **2000**, *37*, 989–994; b) L. C. Vining, *Gene* **1992**, *115*, 135–140.

[4] A. A. Morollo, M. J. Eck, *Nat. Struct. Biol.* **2001**, *8*, 243–247.

[5] A. A. Morollo, R. Bauerle, *Proc. Natl. Acad. Sci. USA* **1993**, *90*, 9983–9987.

[6] D. Sehnal, R. Svobodova Varekova, K. Berka, L. Pravda, V. Navratilova, P. Banas, C. M. Ionescu, M. Otyepka, J. Koca, *J. Cheminf.* **2013**, *5*, 39.

[7] Q. A. Li, D. V. Mavrodi, L. S. Thomashow, M. Roessle, W. Blankenfeldt, *J. Biol. Chem.* **2011**, *286*, 18213–18221.

[8] J. Zwahlen, S. Kolappan, R. Zhou, C. Kisker, P. J. Tonge, *Biochemistry* **2007**, *46*, 954–964.

[9] J. Liu, N. Quinn, G. A. Berchtold, C. T. Walsh, *Biochemistry* **1990**, *29*, 1417–1425.

[10] O. Kerbarh, D. Y. Chirgadze, T. L. Blundell, C. Abell, *J. Mol. Biol.* **2006**, *357*, 524–534.

[11] K. T. Ziebart, M. D. Toney, *Biochemistry* **2010**, *49*, 2851–2859.

[12] J. E. Culbertson, D. H. Chung, K. T. Ziebart, E. Espiritu, M. D. Toney, *Biochemistry* **2015**, *54*, 2372–2384.

[13] N. J. Darby, T. E. Creighton, *Protein structure*, Oxford University Press, Oxford, UK, **1993**.

[14] L. Pravda, K. Berka, R. Svobodova Varekova, D. Sehnal, P. Banas, R. A. Laskowski, J. Koca, M. Otyepka, *BMC Bioinf.* **2014**, *15*, 379.

Received: June 3, 2015
Revised: June 23, 2015
Published online: August 6, 2015

Supporting Information for Publication B

Conversion of anthranilate synthase into isochorismate synthase: Implications for the evolution of chorismate-utilizing enzymes

Maximilian G. Plach, Patrick Löffler, Rainer Merkl, and Reinhard Sterner (2015).

Angewandte Chemie International Edition 54:11270-11274

Angewandte Chemie

Eine Zeitschrift der Gesellschaft Deutscher Chemiker

Supporting Information

Conversion of Anthranilate Synthase into Isochorismate Synthase: Implications for the Evolution of Chorismate-Utilizing Enzymes

*Maximilian G. Plach, Patrick Löffler, Rainer Merkl, and Reinhard Sterner**

anie_201505063_sm_miscellaneous_information.pdf

Supporting Information

Table of Contents

Page Contents

Experimental Materials and Methods

Materials.

Chorismate was purchased as the barium salt from Sigma Aldrich. Barium was removed by the addition of a slight excess of sodium sulfate. All other chemical reagents were purchased from Sigma Aldrich in analytical grade and used without further purification.

Computation of AS, ADCS, ICS, and SS sequence logos.

Sequences of ADCS (6189), AS (7330), ICS (7933), and SS (986) were taken from the InterPro entry IPR015890 (InterPro release 47.0 [1]) to create multiple sequence alignments (MSAs) by means of MAFFT in FFT-NS1 mode [2]. The resulting MSAs were filtered to represent pairwise sequence identity values ranging from 25 to 90%. Finally, the remaining sequences (410 ADCS, 428 AS, 313 ICS, and 71 SS) were combined in a single dataset and aligned using MAFFT in L-INS-I mode; the final MSA was named MSA_MST.

To verify that MSA_MST contains only correctly annotated ADCS, AS, ICS, and SS sequences a phylogenetic tree of MSA_MST was constructed as follows: In order to eliminate poorly aligned regions, Gblocks was utilized to remove columns that contain more than 50% gaps [3]. PhyloBayes [4] was used to generate a total of 100,000 tree topologies, based on a Markov Chain Monte Carlo approach with five independent Markov chains. The first 10,000 burn-in topologies of each chain were discarded and an unrooted consensus tree was derived from the remaining topologies using *readpdb* in PhyloBayes.

This tree (Figure S1) confirmed the correct annotation and grouping of the four different types of MST sequences in MSA_MST. Therefore, these AS, ADCS, ICS, and SS sequences were used to calculate sequence logos of β-strand 11 (Gly259 - Pro266 in the TrpE subunit of stAS) and α-helix 12 (Ser359 - Leu372). The logos were created with WebLogo, version 3.3 [5].

Cloning and site-directed mutagenesis.

The *Salmonella typhimurium trpE* and *trpG* genes were amplified from *S. enterica subsp. enterica serovar Typhimurium str. LT2* genomic DNA using the oligonucleotides 5'st*trpE_NdeI* / 3'st*trpE_XhoI* and 5'st*trpG_NdeI* / 3'st*trpG_XhoI*, respectively. The *E. coli entC* gene was cloned from *E. coli* MG1655 whole cell lysate using the oligonucleotides 5'ec*entC_NdeI* / 3'ec*entC_XhoI*. *Pseudomonas aeruginosa pchB* was amplified from *P. aeruginosa* PAO1 genomic DNA using the oligonucleotides 5'pa*pchB_NdeI* / 3'pa*pchB_XhoI*. The sequences of the oligonucleotides used for gene amplification are listed in Table S3. All amplified genes were subsequently cloned into the pET21a vector (Stratagene, providing a C-terminal hexahistidine-tag) via the introduced restriction sites for *NdeI* and *XhoI*. Variants of st*trpE* were generated by a modified QuickChange mutagenesis protocol [6] using the combinations of plasmids and oligonucleotides specified in Table S1. All gene constructs were entirely sequenced to exclude inadvertent mutations.

Heterologous expression and purification of recombinant proteins.

The stTrpE, stTrpG, ecEntC and paPchB wild-type proteins as well as all stTrpE variants were produced by expressing the respective genes in *E. coli* BL21-Gold (DE3) cells (Agilent Technologies). To this end, 4 L of Luria broth (LB) medium supplemented with 150 µg/mL ampicillin were inoculated with pre-cultures of individual clones and incubated at 37 °C. After an OD_{600} of 0.6 was reached, the temperature was lowered to 20 °C. Expression was induced by adding 0.5 mM isopropyl β-D-1-thiogalactopyranoside (IPTG) and growth was continued overnight. Cells were harvested by centrifugation (2700 g, 4 °C), suspended in 50 mM Tris-HCl, pH 7.5, 300 mM potassium chloride, 10 mM imidazole, and lysed by sonification. The soluble fraction of the cell extract was separated from the insoluble fraction by centrifugation (23000 g, 4 °C). Supernatants were loaded onto a HisTrapFF crude column (5 mL, GE Healthcare), which had been equilibrated with 50 mM Tris-HCl, pH 7.5, 300 mM potassium chloride, 10 mM imidazole. After washing with equilibration buffer, the bound protein was eluted by applying a linear gradient of 10-750 mM imidazole. Subsequently, fractions containing the protein of interest, as judged by SDS-PAGE, were pooled. For further purification, the pool was loaded onto a Superdex 75 column (HiLoad 26/60, 320 mL, GE Healthcare), equilibrated with 50 mM Tris-HCl, pH 7.5, 50 mM KCl, 5 mM $MgCl_2$, 2 mM DTT at 4 °C. In all cases but one, at least 15 mg pure protein was obtained per liter of culture. Only for the KAA variant a maximum of 1 mg protein per liter of culture could be obtained. EcEntC and paPchB were additionally purified by anion-exchange chromatography. To this end, both proteins were dialyzed twice against 50 mM Tris-HCl, pH 7.5, 2 mM $MgCl_2$, 2 mM DTT and subsequently loaded onto a MonoQ column (HR 16/10, 20 mL, GE Healthcare), which had been equilibrated with the same buffer. The column was washed and bound protein was eluted by applying a linear gradient of 0-3 M potassium chloride. Fractions containing ecEntC and paPchB, respectively, as judged by SDS-PAGE, were pooled and dialyzed against 50 mM Tris-HCl, pH 7.5, 50 mM KCl, 5 mM MgCl2, 2 mM DTT.

HPLC analysis of reaction products.

The standard reaction for the enzymatic generation of various products from chorismate was carried out in 50 mM potassium phosphate buffer, pH 7.0, and contained 5 mM $MgCl_2$, 1 mM DTT, 10 µM stTrpE variant or ecEntC, 10 µM stTrpG, and 500 µM chorismate. If needed, 20 mM glutamine were added as an ammonia source. Conditions for control reactions (no enzymes) were identical. After 3 h at 25 °C the samples were ultrafiltrated to remove proteins (Amicon Ultra 0.5 mL centrifugal unit, 10 kDa molecular weight cut-off, 4 °C). Protein-free filtrates were analyzed on an Agilent 1200 HPLC system with a 5 µm Agilent Zorbax Eclipse XDB C18 column (150 mm × 4.6 mm, Figure S2). Mobile phase A consisted of 0.1% formic acid in H_2O, mobile phase B of 0.1% formic acid in acetonitrile. Reaction mixtures were separated at 10 °C with a flow rate of 1 mL/min in the following manner: isocratic elution at 5% B from 0-5 minutes, linear gradient from 5-100% B from 6-20 minutes. Elution was monitored by using absorbance at 280 nm and 320 nm and fluorescence emission at 400 nm following excitation at 310 nm. To report each product formed by the different stTrpE variants as a fraction of all products formed, peak areas were converted into molar amounts either by comparison with standards of known concentration (for AA and SA) or by using the molar extinction coefficients of 11500 $M^{-1}cm^{-1}$ at 280 nm for ADIC and 8300 $M^{-1}cm^{-1}$ at 278 nm for IC [7]. AA and SA peaks were assigned based on a comparison of retention times with those of authentic standards. ADIC was identified based on previously reported HPLC analyses of chorismate derivatives [8]. IC was unambiguously identified by HPLC-ESI mass spectrometry (Figures S3a and S3b) and enzymatic conversion to SA by the isochorismate-pyruvate-lyase PchB from *P. aeruginosa* (Figure S3c).

Mass spectrometry.

The bona fide IC peak detected in the HPLC experiments, was confirmed by HPLC-ESI mass spectrometry. To this end, the stTrpE_KLS variant was incubated with 500 µM chorismate under the following conditions: 10 mM bicine buffer, pH 8.5, 5 mM $MgCl_2$ and 1 mM DTT. After 3 h at 25 °C the sample was ultrafiltrated to remove the protein (Amicon Ultra 0.5 mL centrifugal unit, 10 kDa molecular weight cut-off, 4 °C). The protein-free filtrate was analyzed on an Agilent 1290 HPLC system with a 5 µm Agilent Zorbax Eclipse XDB C18 column (150 mm × 4.6 mm) coupled to an Agilent 6540 Ultra High Definition Accurate Mass Q-TOF LC/MS System. Mobile phases and elution profile were identical to the standard HPLC setup described above. Elution was monitored by using absorbance at 280 nm and 320 nm. ESI mass spectrometry was performed in negative ion mode.

Enzymatic conversion of isochorismate to salicylate.

IC was additionally identified in the product mixtures of the stTrpE variants by enzymatic conversion to SA. The isochorismate-pyruvate-lyase PchB from *Pseudomonas aeruginosa* converts IC to SA and pyruvate. It is also reported to directly convert CH to prephenic acid by its promiscuous chorismate mutase activity [9]. A direct conversion of CH to SA, however, has not been described for PchB. StTrpE variants KMA, KLA, and KLS, as well as ecEntC, were incubated with chorismate under reaction conditions similar to those described in the mass spectrometry section. After 3 h at 25 °C the samples were ultrafiltrated to remove proteins (Amicon Ultra 0.5 mL centrifugal unit, 10 kDa molecular weight cut-off, 4 °C). One third of the protein-free filtrate was directly subjected to standard HPLC analysis. The second third of the filtrate was incubated at 25 °C for 30 minutes without any further modification before standard HPLC analysis and the last third was supplemented with PchB (5 µM), incubated at 25 °C for 30 minutes, ultrafiltrated, and subjected to standard HPLC analysis.

Computation of the structures of wild-type stTrpE and stTrpE variants.

The available crystal structure of stTrpE in complex with stTrpG (PDB ID 1i1q) represents an unliganded, open, inactive T-state form [10]. Therefore a stTrpE:stTrpG homology model was generated based on the crystal structure of the TrpE:TrpG complex from *Serratia marcescens*

(PDB ID 1i7q), which resembles a ligand-bound form with a closed active site [11]. Modeling was performed with YASARA Structure Version 14.7.17 employing the YASARA2 force field [12]. The high similarity of target and template sequences argues in favor of a good 3D-model. Sequence identity values determined by EMBOSS Needle were 71.3% (TrpE) and 79.8% (TrpG). Moreover, YASARA's Z-scores were -0.462 (TrpE) and -0.352 (TrpG), indicating good model quality. CH was placed in the active site of the stTrpE model, substituting the benzoate and pyruvate ligands present in 1i7q; the RMSD for all matching atoms was 0.713 Å. Structures of mutant stTrpE variants were generated by *in silico* mutating residues 263, 364, and 365 of the stTrpE homology model. Mutated residues were rotamer-optimized employing the SCWALL method of YASARA [13]. To remove conformational stress, all homology models were equilibrated by means of a short molecular dynamics (MD) simulation, resulting in equilibrated homology models (EHMs). For details of the protocol see below.

Molecular dynamics simulations.

MD simulations were performed at 298 K under periodic boundary conditions with explicit water, using a multiple time step of 1 fs for intramolecular and 2 fs for intermolecular forces. Lennard-Jones and long-range electrostatic interactions were treated with a 7.86 Å cut-off, the latter were calculated using the Particle Mesh Ewald method [14]. Temperature was adjusted using a Berendsen thermostat based on the time-averaged temperature; the atmospheric pressure was kept constant. Each simulation cell was 5 Å larger than the protein along each axis; cells were filled with water to a density of 0.997 g/ml and counter ions were added to a final concentration of 0.9% NaCl. Protonation states of protein side chains were assigned as described [12] and parameterization of glutamine and chorismate was performed using the AM1BCC method. Each snapshot was energy-minimized as follows: After an initial steepest descent phase, energy of the system was further minimized by means of simulated annealing (time step 2 fs, atom velocities scaled down by 0.9 every 10th step) until energy did not improve by more than 0.05 kJ/mol per atom during 200 consecutive steps.

Computing putative nucleophile paths in wild-type stTrpE and stTrpE variants.

An intrinsic feature of the stAS complex is the intermolecular transport of ammonia from the active site of stTrpG to the active site of stTrpE. To identify the most likely paths of ammonia in wild-type stAS and of the nucleophile in stAS complexes with mutated stTrpE variants, the respective EHMs were subjected to MD simulations and nucleophile channels were computed as follows. EHMs of wild-type stTrpE and of each stTrpE variant were simulated in three production MD runs. To reassign initial atom velocities and seed independent calculations, the temperature was slightly changed (±0.1 K) for the second and third run. Trajectories were sampled at intervals of 10 ps for a total of 2 ns, resulting in 600 snapshots for each stTrpE variant. These structures were further energy minimized prior to the computation of channels. Moreover, for visual inspection, average 3D models were generated for each MD trajectory as follows: An EHM served as a reference structure and average positions for all atoms were deduced after superimposing the protein structures from all snapshots.

Nucleophile channels were computed utilizing MOLE 2.0, version 2.13.9.6 [15]. Default values were used except a probe radius of 2.14 Å, which is the size of ammonia in the given force-field. The starting point was the all-atom centroid of the ligand glutamine and Cys87 in stTrpG that approximates the location where nascent ammonia is generated. The endpoint was the all-atom centroid of the ligand CH and Ala327 in stTrpE that approximates the location of the CH-C2 atom where the initial nucleophilic attack in the AS reaction occurs. For each of the 600 resulting channels per variant, the channels centerlines served to specify a putative nucleophile trajectory (PNT). As the MD simulations induced small translational and rotational movements of the stAS complexes, a direct comparison of related PNTs was not possible. To compensate for this effect, all PNTs were superimposed on the respective EHM [16] and the resulting PNT-bundles were analyzed further.

Visual inspection of PNT-bundles by means of PyMol [17] in the region near the CH ligand indicated a preference for two major paths (Figure S4); one proceeding alongside Val265 and the other one alongside residues 365 and 425 (Figure S5a). Due to their prevalence in Leu365 and Ala365 variants, these paths were termed L-path and A-path, respectively. The spatial distribution of PNTs observed in a variant (Table S2) was determined by counting the number of PNTs that proceed along the L-path or the A-path, as follows: First, for each PNT *j*, the segment with a distance of 3 to 7 Å from CH-C2 was identified (Figure S5b). Due to the complexity of the local curvature of individual PNTs, these segments were represented by a different number of 3D coordinates. Thus, we binned the coordinates $coord_k^{i,j}$ in 16 shells i (thickness ¼ Å, centered on CH-C2) and computed the vector $\mathbf{pv}_k^{i,j}$ that starts at $coord_k^{i,j}$ and ends in CH-C2. Then a plane $\mathbf{P_j}$ (with normal vector $\mathbf{nv_j}$) was defined by CH-C2 and the Cα atoms of the PNT-lining residues Met364 and Thr425. Each $\mathbf{pv}_k^{i,j}$ was multiplied with the normal vector $\mathbf{nv_j}$; the sign of the scalar product (+, -) $s_k^{i,j}$ indicates the position of $coord_k^{i,j}$ relative to $\mathbf{P_j}$. For each shell i and PNT j, the mean $\overline{s^{i,j}} = \frac{1}{l}\sum_{k=1}^{l} s_k^{i,j}$, where l is the number of coordinates $coord_k^{i,j}$, was computed and normalized to [0, 1]. A value of 0 indicates that all PNT coordinates are located on the L-path side of $\mathbf{P_j}$ and a mean of 1 shows that all coordinates are on the A-path side of $\mathbf{P_j}$. The shell-wise computed mean $\overline{s^i} = \frac{1}{m}\sum_{j=1}^{m} \overline{s^{i,j}}$, where m is the number of PNTs in this shell, was then used to determine $\bar{s} = \frac{1}{n}\sum_{i=1}^{n} \overline{s^i}$, where n is the number of shells, which is the percentage of all PNTs along the A-path in the 3 - 7 Å shell around CH-C2. This fraction was considered indicative for the prevalent localization of PNTs and proposes the overall putative nucleophile path of each variant.

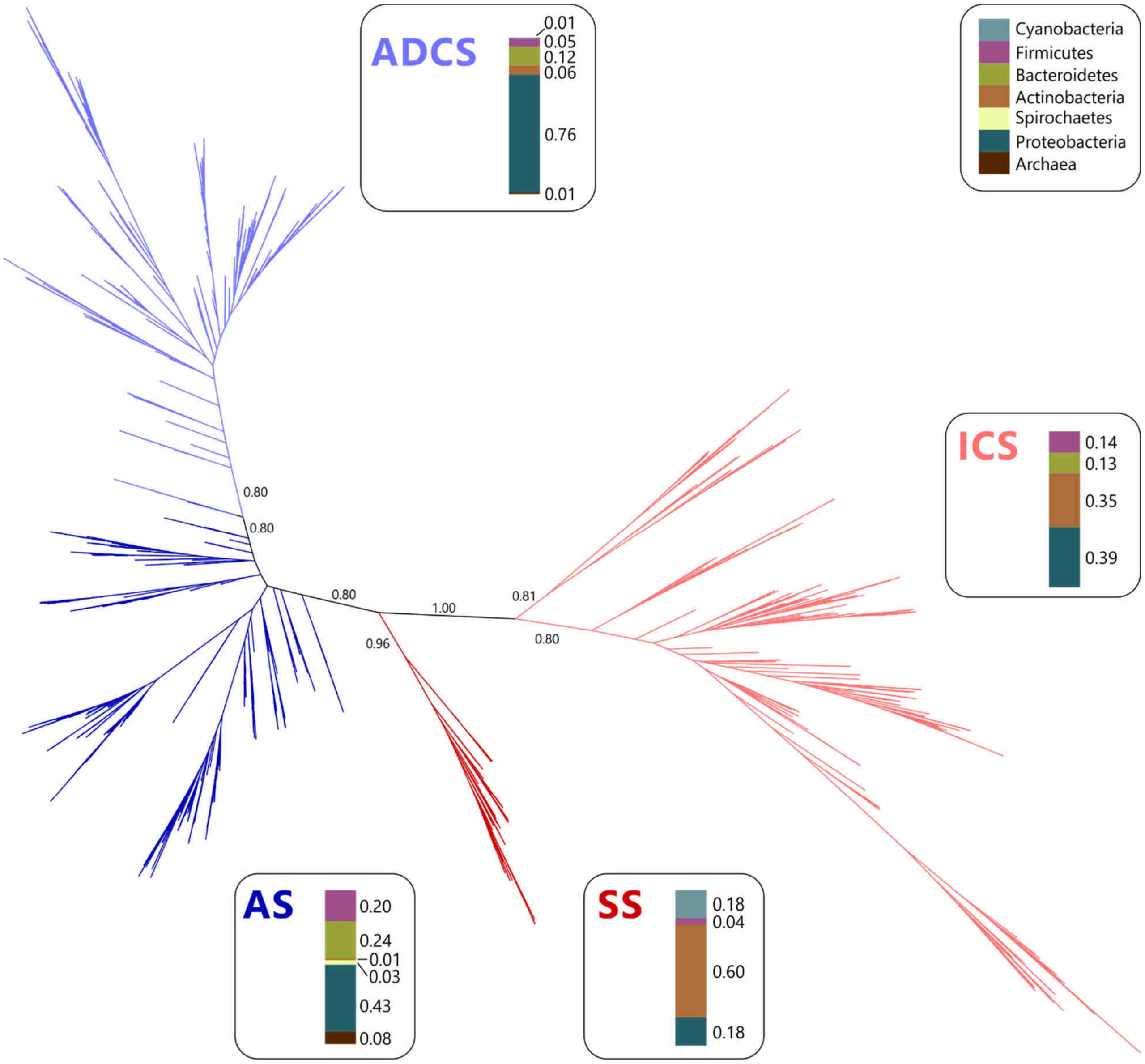


Figure S1. Unrooted phylogenetic tree comprising the four groups of MST enzymes. The edges leading to the AS, ADCS, ICS, and SS leaves, respectively, are colored accordingly. Posterior probabilities are shown for important splits. For each MST group, the contribution of six predominant bacterial phyla as well as archaea is given by color bars that represent the corresponding fractions. Importantly, the four MST groups form distinct subtrees and each group is represented by sequences from at least four different major phyla.

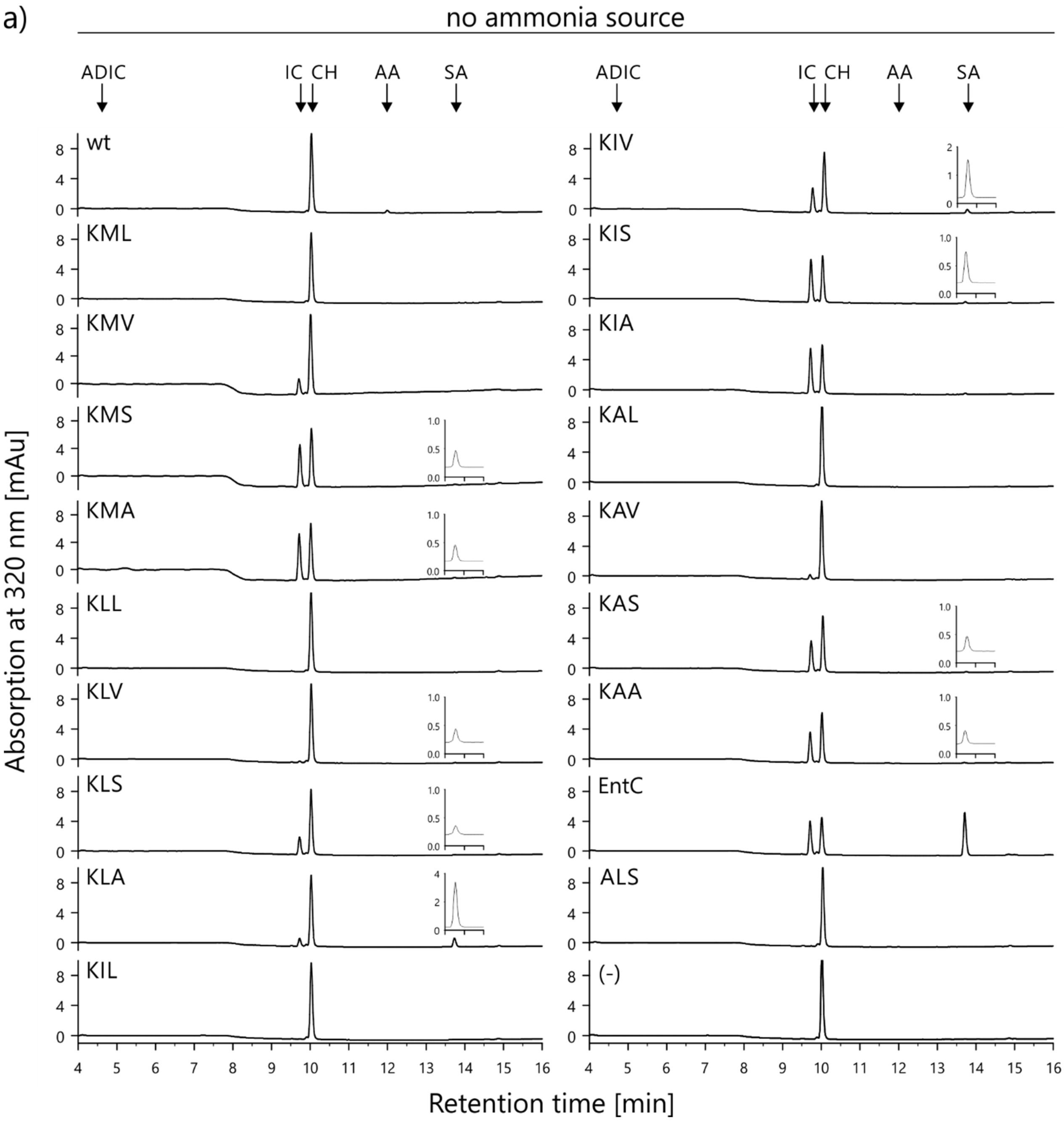


Figure S2. HPLC chromatograms of product mixtures from reactions of stTrpE variants in the presence of stTrpG. The retention times of aminodeoxyisochorismate (ADIC), isochorismate (IC), chorismate (CH), anthranilate (AA), and salicylate (SA) are indicated by arrows. In cases where only minimal SA amounts were detected, an inset, placed at the correct retention time, shows the fluorescence signal at 400 nm after excitation at 310 nm. StTrpE variants are denoted by their three-letter-abbreviations (cf. Table S1). “wt”, “EntC”, “ALS” and “(-)” refer to control reactions with wild-type stTrpE, with the native ICS EntC from *E. coli*, with a Lys263Ala knockout variant and without any enzyme, respectively. a) Product mixtures in the absence of an ammonia source.

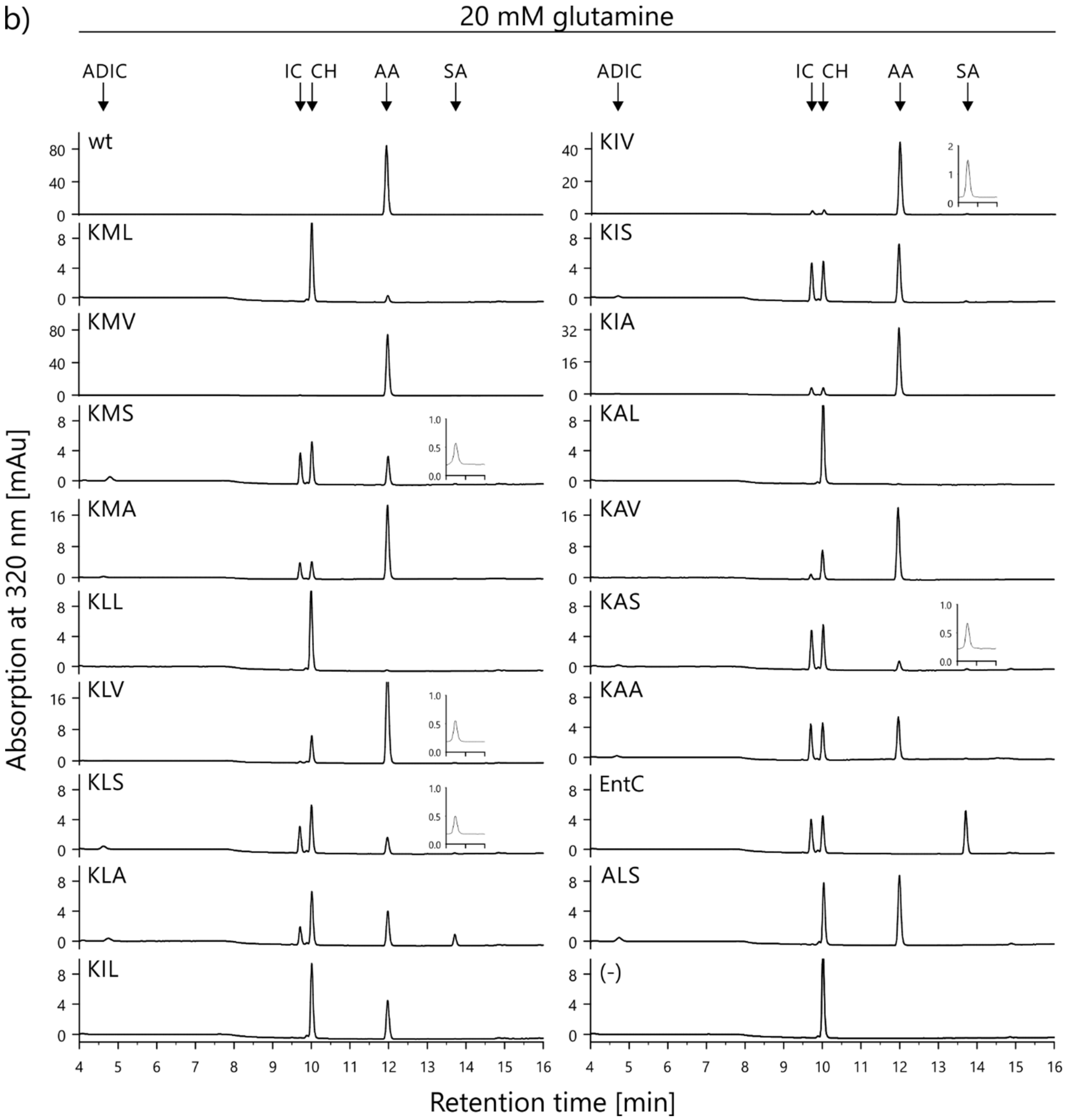


Figure S2 continued. b) Product mixtures in the presence of 20 mM glutamine as an ammonia source. Note the different y-axis scales. Small variations in ADIC retention times are due to minor differences in chromatographic conditions.

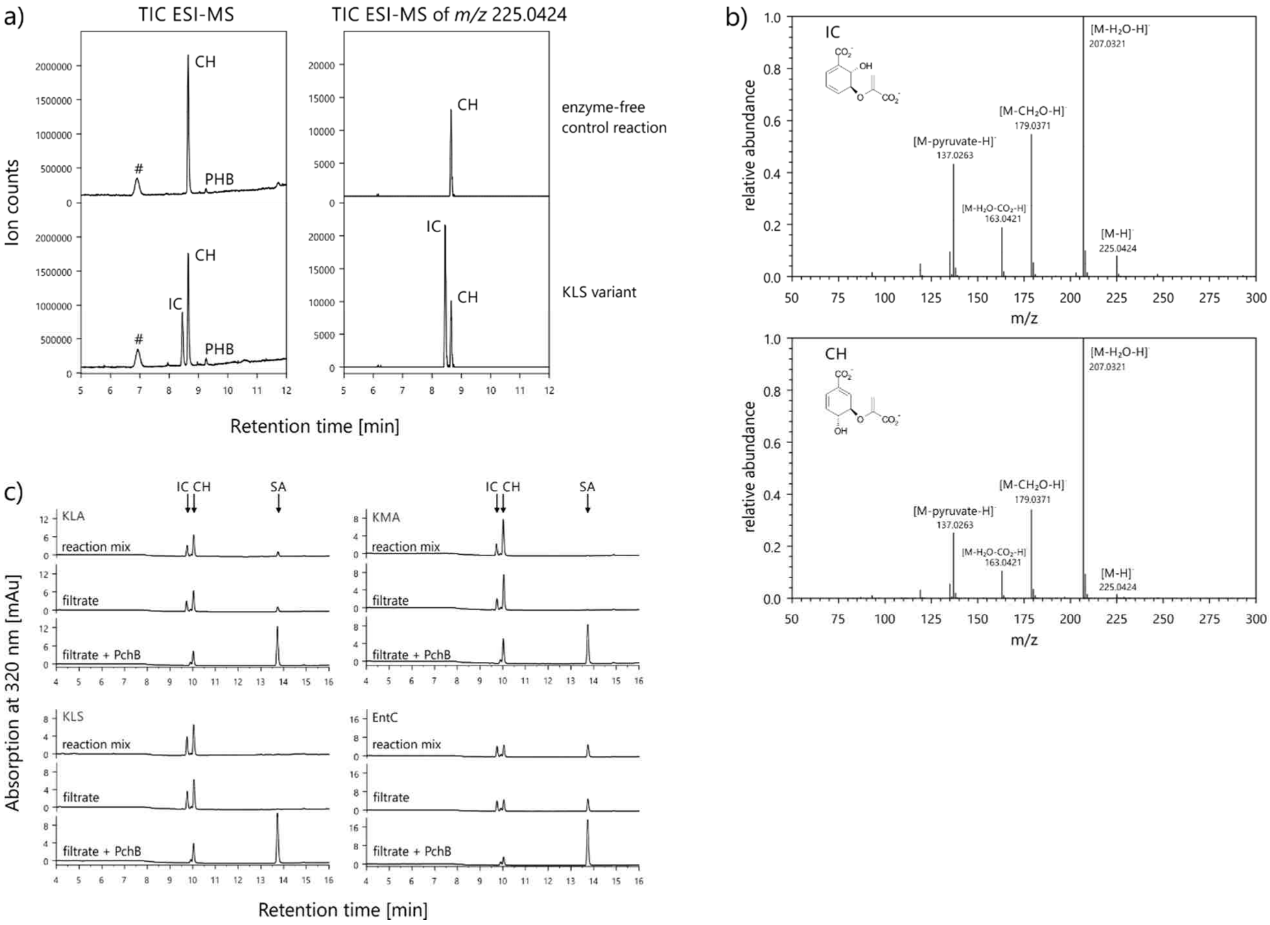


Figure S3. Identification of isochorismate (IC) by HPLC-ESI-MS and by enzymatic conversion to salicylate (SA). a) Total ion count (TIC) chromatograms for all m/z values (left) and for m/z value 225.0424 (right). In an enzyme-free control reaction (top), only chorismate (CH) and two impurities denoted by # and PHB (*p*-hydroxybenzoate), but no isochorismate (IC) were detected. In contrast, IC was detected at an m/z value of 226.0497 (the m/z-value of 225.0424 corresponds to the [M-H]$^-$-ions of IC and CH) in the reaction mixture of the stTrpE KLS variant. The m/z value is in good agreement with the theoretical value of 226.0477. b) ESI scans of IC and CH peaks from the TIC chromatograms shown in panel a). As IC and CH are constitutional isomers, both share the same m/z value of 226.0497 and appear in the corresponding TIC chromatogram of the reaction mixture of the stTrpE KLS variant. However, the ESI scans of the IC and CH peaks, selected based on retention times of 8.419-8.497 min and 8.619-8.696 min, respectively, reveal differences in their ion compositions. Although both patterns are dominated by ions that underwent the loss of H_2O (-18), CH_2O (-46), H_2O + CO_2 (-62), and pyruvate (-88), the relative peak heights differ. The fraction of “intact” molecular ions ([M-H]$^-$) is greater for IC, whereas the fraction of [M-H_2O-H]$^-$ ions is greater for CH. c) HPLC chromatograms depicting the enzymatic conversion of IC to SA by the isochorismate-pyruvate-lyase PchB. The retention times of IC, CH, and SA are indicated by arrows. “Reaction mix” denotes the product mixture after incubation of the enzymes (stTrpE variants KLA, KMA, and KLS, as well as the native ICS EntC from *E. coli*) with

500 µM CH at pH 8.5 and 25 °C for three hours. “Filtrate” denotes the product mixture after removal of enzymes by ultrafiltration and subsequent incubation at 25 °C for 30 minutes. “Filtrate + PchB” denotes the product mixture after removal of enzymes and subsequent incubation at 25 °C for 30 minutes with supplementation of PchB. Upon addition of PchB to the filtrate, IC is entirely converted to SA. The supplementation of PchB also leads to a slight decrease in the amount of residual CH in the filtrate. This is most likely due to the promiscuous chorismate mutase activity of PchB [9].

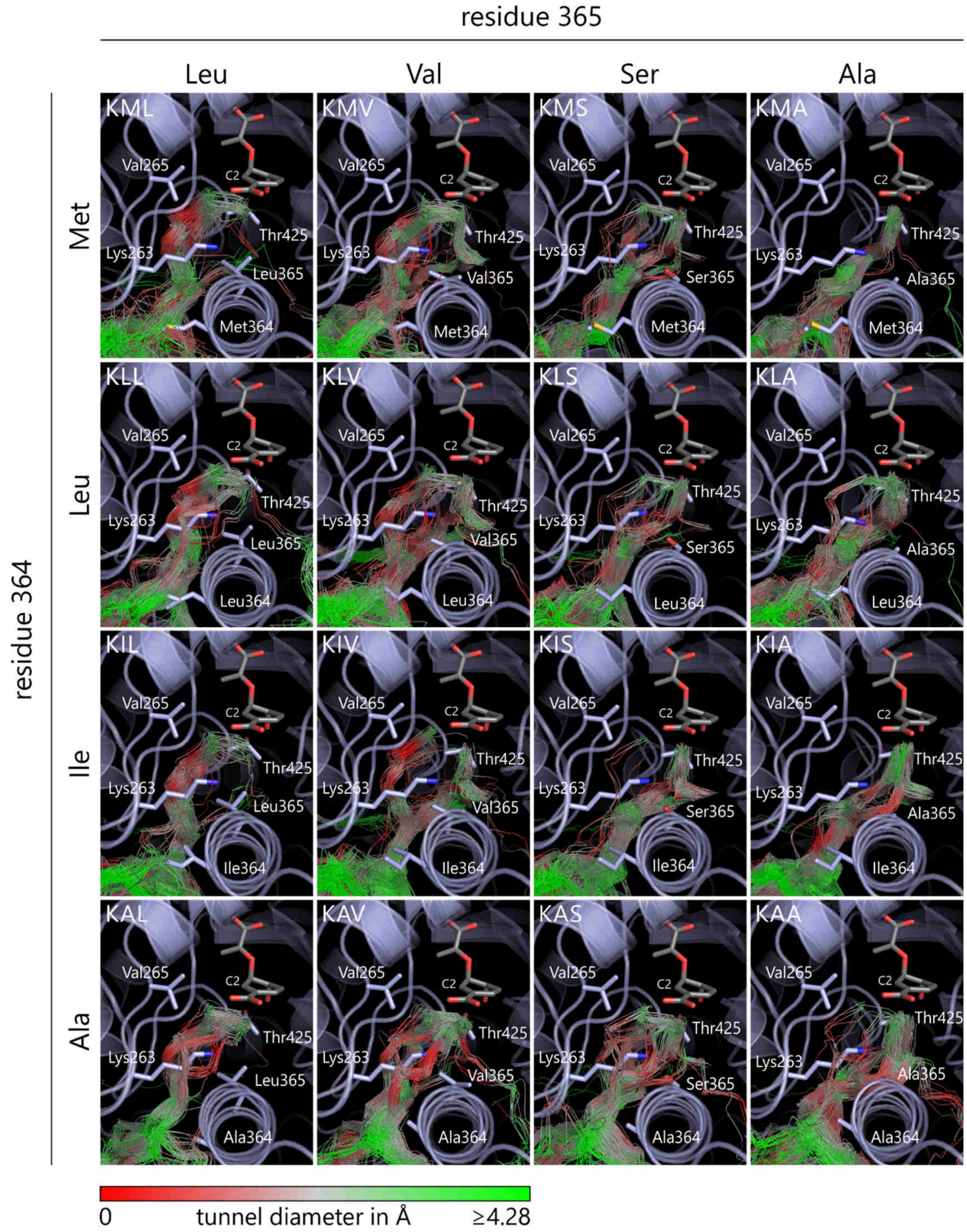


Figure S4. Comparison of PNTs in 16 stTrpE variants containing Lys263. PNT localization is almost exclusively determined by residue 365. To a great extent, PNTs of Leu365-variants proceed along Val265, i.e. along the L-path (cf. Figure S5). In variants with Val365, approximately half of the PNTs proceed along the L-path; in the KAV variant this distribution is less pronounced. In variants with Ser365, PNTs further shift towards the A-path (cf. Figure S5). This effect is most pronounced in the KIS variant. In the Ala365-variants, nearly all PNTs follow the A-path. Diameters of the PNT associated channels can be deduced from the red-grey-green color scale.

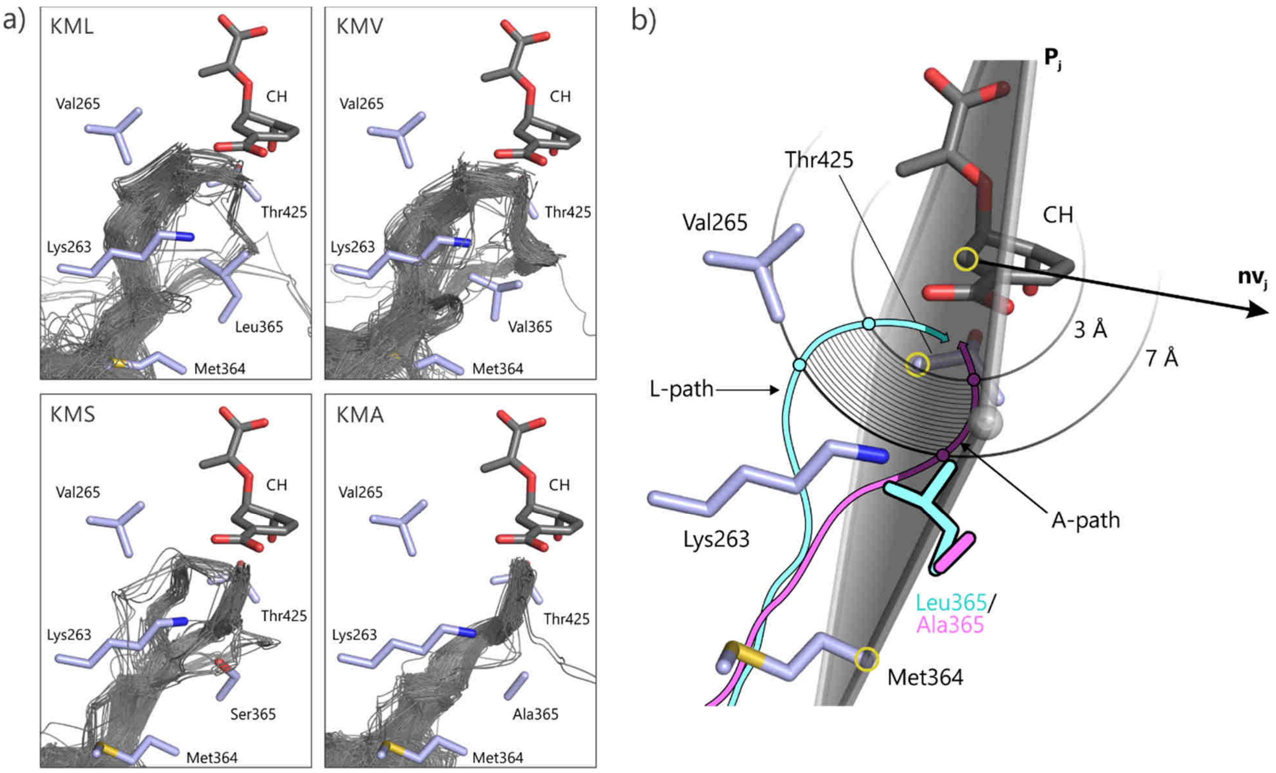


Figure S5. Principles used to quantify the spatial distribution of PNTs. a) Distribution of PNTs near Lys263 and residues 364 and 365 in the KML, KMV, KMS, and KMA variants. The localization of PNTs shifts with decreasing size of residue 365 (Leu > Val > Ser ≈ Ala) from one predominant path to a second one. As the majority of PNTs in Leu365-variants proceed along the path visible in the KML panel, the respective path was termed L-path. Accordingly, the predominant path deduced from the KMA variant was termed A-path. b) Graphical representation of L- and A-path (cyan and magenta, respectively) as well as of the plane $\mathbf{P}_j$ and its normal vector $\mathbf{nv}_j$ used for quantifying the spatial distribution of PNTs. $\mathbf{P}_j$ was specified by CH-C2 and the Cα atoms of residues 364 and 425 (yellow circles). For binning of PNT coordinates, sixteen ¼ Å shells (black circles) spanning a segment of 3 to 7 Å from CH-C2 were defined.

Table S1. List of plasmids and oligonucleotides used for generating st*trpE* variants.

Variant	*Template*	*Oligonucleotide*	*Sequence (5’→3’)*
st*trpE*_ Q263K (=KML)	pET21a_st*trpE*	5’st*trpE*_Q263K 3’st*trpE*_Q263K	GCGGGCGAGATATTT**A**A**A**GTGGTGCCGTCGC GCGACGGCACCAC**T**T**T**AAATATCTCGCCCGC
st*trpE* _ Q263K_L365V (=KMV)	pET21a_st*trpE*_ Q263K	5’st*trpE*_L365V 3’st*trpE*_L365V	GCTTTCCGAACATCTGATG**G**TGGTCGATCTGGCGCG CGCGCCAGATCGACCA**C**CATCAGATGTTCGGAAAGC
st*trpE* _ Q263K_L365S (=KMS)	pET21a_st*trpE*_ Q263K	5’st*trpE*_L365S 3’st*trpE*_L365S	GCTTTCCGAACATCTGATG**TCT**GTCGATCTGGCGCG CGCGCCAGATCGAC**AGA**CATCAGATGTTCGGAAAGC
st*trpE* _ Q263K_L365A (=KMA)	pET21a_st*trpE*_ Q263K	5’st*trpE*_L365A 3’st*trpE*_L365A	GCTTTCCGAACATCTGATG**GCG**GTCGATCTGGCGCG CGCGCCAGATCGAC**CGC**CATCAGATGTTCGGAAAGC
st*trpE* _ Q263K_M364L (=KLL)	pET21a_st*trpE*_ Q263K	5’st*trpE*_M364L 3’st*trpE*_M364L	GCTTTCCGAACATCTG**C**TGCTGGTCGATCTGGC GCCAGATCGACCAGCA**G**CAGATGTTCGGAAAGC
st*trpE* _ Q263K_M364L_L365V (=KLV)	pET21a_st*trpE*_ Q263K_M364L_L365S	5’st*trpE*_S365V 3’st*trpE*_S365V	CTTTCCGAACATCTGCTG**GTG**GTCGATCTGGCGCG CGCGCCAGATCGAC**CAC**CAGCAGATGTTCGGAAAG
st*trpE* _ Q263_M364L_L365S (=KLS)	pET21a_st*trpE*_ Q263K_L365S	5’st*trpE*_M364L_v2 3’st*trpE*_M364L_v2	GAGCTTTCCGAACATCTG**C**TGTCTGTCGATCTGGCGCG CGCGCCAGATCGACAGACA**G**CAGATGTTCGGAAAGCTC
st*trpE* _ Q263K_M364L_L365A (=KLA)	pET21a_st*trpE*_ Q263K_M364L_L365S	5’st*trpE*_S365A 3’st*trpE*_S365A	CTTTCCGAACATCTGCTG**G**CTGTCGATCTGGCGCG CGCGCCAGATCGACAG**C**CAGCAGATGTTCGGAAAG
st*trpE* _ Q263K_M364I (=KIL)	pET21a_st*trpE*_ Q263K	5’st*trpE*_M364I 3’st*trpE*_M364I	GCTTTCCGAACATCTGAT**C**CTGGTCGATCTGGC GCCAGATCGACCAG**G**ATCAGATGTTCGGAAAGC
st*trpE* _ Q263K_M364I_L365V (=KIV)	pET21a_st*trpE*_ Q263K	5’st*trpE*_M364I_L365V 3’st*trpE*_M364I_L365V	GAGCTTTCCGAACATCTGAT**CG**TGGTCGATCTGGCGCG CGCGCCAGATCGACCA**CG**ATCAGATGTTCGGAAAGCTC
st*trpE* _ Q263K_M364I_L365S (=KIS)	pET21a_st*trpE*_ Q263K_M364I_L365V	5’st*trpE*_V365S 3’st*trpE*_V365S	GCTTTCCGAACATCTGATC**TC**GGTCGATCTGGCGCG CGCGCCAGATCGACC**GA**GATCAGATGTTCGGAAAGC
st*trpE* _ Q263K_M364I_L365A (=KIA)	pET21a_st*trpE*_ Q263K_M364I_L365V	5’st*trpE*_V365A 3’st*trpE*_V365A	GCTTTCCGAACATCTGATCG**C**GGTCGATCTGGCGCG CGCGCCAGATCGACC**G**CGATCAGATGTTCGGAAAGC
st*trpE* _ Q263K_M364A (=KAL)	pET21a_st*trpE*_ Q263K	5’st*trpE*_M364A 3’st*trpE*_M364A	GCTTTCCGAACATCTG**GC**GCTGGTCGATCTGGC GCCAGATCGACCAGC**GC**CAGATGTTCGGAAAGC
st*trpE* _ Q263K_M364A_L365V (=KAV)	pET21a_st*trpE*_ Q263K_M364L_L365V	5’st*trpE*_L364A_v2 3’st*trpE*_L364A_v2	GCTTTCCGAACATCTG**GC**GGTGGTCGATCTGGC GCCAGATCGACCACC**GC**CAGATGTTCGGAAAGC
st*trpE* _ Q263K_M364A_L365S (=KAS)	pET21a_st*trpE*_ Q263K_M364L_L365S	5’st*trpE*_L364A 3’st*trpE*_L364A	GCTTTCCGAACATCTG**GCG**TCTGTCGATCTGGCG CGCCAGATCGACAGA**CGC**CAGATGTTCGGAAAGC
st*trpE* _ Q263K_M364A_L365A (=KAA)	pET21a_st*trpE*_ Q263K_M364A_L365S	5’st*trpE*_S365A_v2 3’st*trpE*_S365A_v2	GCTTTCCGAACATCTGGCG**GC**TGTCGATCTGGCGCGC GCGCGCCAGATCGACA**GC**CGCCAGATGTTCGGAAAGC
st*trpE* _ Q263A_M364L_L365S (=ALS)	pET21a_st*trpE*_ Q263K_M364L_L365S	5’st*trpE*_K263A 3’st*trpE*_K263A	GCGGGCGAGATATTT**GC**AGTGGTGCCGTCGC GCGACGGCACCACT**GC**AAATATCTCGCCCGC

Table S2. Fractions of PNTs ($\bar{s}$) approaching the CH ligand along the A-path.

		Residue 365			
		L	V	S	A
Residue 364	M	5%	32%	79%	100%
	L	2%	61%	86%	98%
	I	5%	58%	99%	98%
	A	4%	23%	66%	97%

Table S3. List of oligonucleotides used for amplification of st*trpE*, st*trpG,* ec*entC* and pa*pchB* genes.

Oligonucleotide	Sequence (5’→3’)
5’st*trpE_NdeI*	GGAATTCCATATGCAAACACCAAAACCC
3’st*trpE_XhoI*	AAACTCGAGGAAGGTCTCCTGT
5’st*trpG_NdeI*	GGAATTCCATATGGCTGATATTCTGCT
3’st*trpG_XhoI*	AATCTCGAGCTTTTGCTGCGCCCAG
5’ec*entC_NdeI*	CAGGGCATATGGATACGTCACTGGCTGAG
3’ec*entC_XhoI*	CCCTGCTCGAGATGCAATCCAAAAACGTTC
5’pa*pchB_NdeI*	CAGGGCATATGAAAACTCCCGAAGAC
3’pa*pchB_XhoI*	CCCTGCTCGAGTGCGGCACCCCGTGTCTG

Restriction sites are underlined.

Supplementary References

[1] S. Hunter, P. Jones, A. Mitchell, R. Apweiler, T. K. Attwood, A. Bateman, T. Bernard, D. Binns, P. Bork, S. Burge, E. de Castro, P. Coggill, M. Corbett, U. Das, L. Daugherty, L. Duquenne, R. D. Finn, M. Fraser, J. Gough, D. Haft, N. Hulo, D. Kahn, E. Kelly, I. Letunic, D. Lonsdale, R. Lopez, M. Madera, J. Maslen, C. McAnulla, J. McDowall, C. McMenamin, H. Mi, P. Mutowo-Muellenet, N. Mulder, D. Natale, C. Orengo, S. Pesseat, M. Punta, A. F. Quinn, C. Rivoire, A. Sangrador-Vegas, J. D. Selengut, C. J. Sigrist, M. Scheremetjew, J. Tate, M. Thimmajanarthanan, P. D. Thomas, C. H. Wu, C. Yeats, S. Y. Yong, *Nucleic Acids Res.* **2012**, *40*, D306-312.

[2] K. Katoh, D. M. Standley, *Mol. Biol. Evol.* **2013**, *30*, 772-780.

[3] J. Castresana, *Mol. Biol. Evol.* **2000**, *17*, 540-552.

[4] N. Lartillot, H. Philippe, *Mol. Biol. Evol.* **2004**, *21*, 1095-1109.

[5] G. E. Crooks, G. Hon, J. M. Chandonia, S. E. Brenner, *Genome Res.* **2004**, *14*, 1188-1190.

[6] W. Wang, B. A. Malcolm, *BioTechniques* **1999**, *26*, 680-682.

[7] a) A. A. Morollo, R. Bauerle, *Proc. Natl. Acad. Sci. U S A* **1993**, *90*, 9983-9987; b) J. Zwahlen, S. Kolappan, R. Zhou, C. Kisker, P. J. Tonge, *Biochemistry* **2007**, *46*, 954-964.

[8] a) Z. He, K. D. Stigers Lavoie, P. A. Bartlett, M. D. Toney, *J. Am. Chem. Soc.* **2004**, *126*, 2378-2385; b) Z. He, M. D. Toney, *Biochemistry* **2006**, *45*, 5019-5028.

[9] C. Gaille, P. Kast, D. Haas, *J. Biol. Chem.* **2002**, *277*, 21768-21775.

[10] A. A. Morollo, M. J. Eck, *Nat. Struct. Biol.* **2001**, *8*, 243-247.

[11] G. Spraggon, C. Kim, X. Nguyen-Huu, M. C. Yee, C. Yanofsky, S. E. Mills, *Proc. Natl. Acad. Sci. U S A* **2001**, *98*, 6021-6026.

[12] E. Krieger, K. Joo, J. Lee, S. Raman, J. Thompson, M. Tyka, D. Baker, K. Karplus, *Proteins* **2009**, *77 Suppl 9*, 114-122.

[13] A. A. Canutescu, A. A. Shelenkov, R. L. Dunbrack, Jr., *Protein Sci.* **2003**, *12*, 2001-2014.

[14] U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee, L. G. Pedersen, *J. Chem. Phys.* **1995**, *103*.

[15] D. Sehnal, R. Svobodova Varekova, K. Berka, L. Pravda, V. Navratilova, P. Banas, C. M. Ionescu, M. Otyepka, J. Koca, *J. Cheminfo.* **2013**, *5*, 39.

[16] W. Kabsch, *Acta Cryst.* **1976**, *32*, 922-923.

[17] A. B. Chowdry, K. A. Reynolds, M. S. Hanes, M. Voorhies, N. Pokala, T. M. Handel, *J. Comput. Chem.* **2007**, *28*, 2378-2388.

7.3 Publication C

Evolutionary diversification of protein-protein interactions by interface add-ons

Maximilian G. Plach, Florian Semmelmann, Florian Busch, Markus Busch, Leonhard Heizinger, Vicky H. Wysocki, Rainer Merkl, and Reinhard Sterner (2017).

Submitted for Publication

Evolutionary diversification of protein-protein interactions by interface add-ons

Maximilian G. Plach[1], Florian Semmelmann[1], Florian Busch[2], Markus Busch[1], Leonhard Heizinger[1], Vicki H. Wysocki[2], Rainer Merkl[1*], and Reinhard Sterner[1*]

[1]Institute of Biophysics and Physical Biochemistry, University of Regensburg, D-93040 Regensburg, Germany
[2]Department of Chemistry and Biochemistry, The Ohio State University, 460 West 12th Avenue, OH-43210 Columbus, USA
*Correspondence: Rainer Merkl: +49-941-3086; Rainer.Merkl@ur.de, Reinhard Sterner: +49-941 943 3015; Reinhard.Sterner@ur.de

SUMMARY

Cells contain a multitude of protein complexes whose subunits interact with high specificity. However, the number of different protein folds and interface geometries found in Nature is limited. This raises the question how protein-protein interaction specificity is achieved on the structural level and how the formation of non-physiological complexes is avoided. Here we describe structural elements called interface add-ons that fulfill this function and elucidate their role for the diversification of protein-protein interactions during evolution. By performing a systematic computational survey, we identified interface add-ons in 10% of bacterial, heteromeric protein complexes. The paradigmatic experimental characterization of over 30 cognate and hybrid glutamine amidotransferase complexes in combination with comprehensive genetic profiling and protein design demonstrated that interface add-ons are determinants of protein-protein interaction specificity. Moreover, in vivo experiments showed that the lack of interface add-ons can lead to physiologically harmful cross-talk between essential biosynthetic pathways. We conclude that interface add-ons are an effective evolutionary tool to facilitate interaction specificity when a protein has to discriminate between several potential partners that share similar interface geometries.

INTRODUCTION

Protein-protein interactions are essential for key cellular processes, ranging from the formation of molecular machineries to the assembly of signal transduction networks. A huge number of interactions have evolved to accomplish the various biological functions. For example, experimental and *in silico* methods projected the interactome of yeast to comprise approximately 18 000 binary protein-protein interactions (Yu et al., 2008). In the light of such dense protein networks, it is not trivial for a cell to guarantee interaction specificity in crucial cases such as toxin-antitoxin, antibody-antigen, protease-inhibitor or multi-enzyme complexes. This problem is aggravated by the limited number of different interface geometries that mediate protein-protein interactions. It is estimated that not more than 1000-4000 such geometries exist (Gao and Skolnick, 2010; Garma et al., 2012) and that their number is restricted by the same biophysical constraints that limit the number of protein folds (Chothia, 1992; Zhang et al., 2006).

Understanding the principles of protein-protein interactions in general and the assurance of interaction specificity in spite of the limited number of interface geometries in particular is an important biological challenge. A recent *in silico* analysis indicated that relatively small insertions and deletions in protein interfaces can differentiate between monomers and homodimers and that these elements may preclude undesired interactions (Hashimoto and Panchenko, 2010). However, the interfaces in homo-oligomers are unique with respect to amino acid composition and residue-residue contact preferences and differ significantly from those found in other types of protein complexes such as hetero-oligomers (Ofran and Rost, 2003a). For the latter, the assurance of interaction specificity is most demanding in cases where two or more proteins with similar interface geometries exist that in principle compete for the same interaction partner (Schreiber and Keating, 2011). An example would be a complex A:B, in which A can interact with several homologous potential interaction partners B, B′, and B″. The putative interfaces of B′ and B″, whose binding has to be avoided, are often similar to that of the genuine partner B, creating the risk of erroneous and potentially harmful A-B′ and A-B″ interactions.

In order to contribute to the understanding of interaction specificity in hetero-oligomers, we started with a systematic *in silico* survey of the interfaces from 305 representative heteromeric protein complexes. Not considering small differences in secondary structure elements or slightly different quaternary structures, in about 10% of this sample interface geometries vary significantly between related complexes that share homologous subunits. In these cases, interfaces are extended by additional loops and entire secondary structure elements that contain residues crucial for complex stability. We designated these elements “interface add-ons” and presumed that they differentiate interfaces between related complexes that share homologous subunits and thus contribute to interaction specificity by negative design (Schreiber and Keating, 2011), much like additional bits turn a master key into a special key.

In order to experimentally back this assumption, we comprehensively analyzed protein interaction specificities in a family of glutamine amidotransferase complexes (GATases) that are part of the tryptophan and folate biosynthesis pathways. These heteromeric enzyme complexes comprise glutaminase and synthase subunits which interact to transfer ammonia from glutamine to an acceptor substrate (Raushel et al., 1999). The synthase subunits of these GATases, as well as the glutaminase subunits, respectively, are homologs, share high sequence similarity, and belong to the same folds. A subset of synthase subunits exclusively involved in tryptophan biosynthesis contains an interface add-on, which is absent in all other homologous synthase subunits, including those exclusively involved in folate biosynthesis. We experimentally characterized 54 combinations of nine synthases (three containing the interface add-on) with six different glutaminases, as well as a rationally designed synthase with a deletion in its interface add-on. Our results show that glutaminase-synthase interaction specificity is determined by the presence or absence of the interface add-on, independent of the phylogenetic origin of the proteins and their participation in tryptophan or folate biosynthesis.

The profiling of more than 15 000 bacterial and archaeal genomes highlights the greater biological relevance of this finding: Most species possess two homologous synthases that do not contain an interface add-on for tryptophan and folate biosynthesis and thus rely on a single type of glutaminase to interact with the two similar synthases. In contrast, in species that possess a synthase with an interface add-on and a second synthase without an interface add-on, an additional, specifically adapted type of glutaminase is present. We assume that positive selection favored this diversification of the synthases as it allowed for an effective separation of tryptophan and folate biosynthesis. *In vivo* experiments show that this separation is physiologically relevant as its override is detrimental for cellular growth. From a broader perspective, our complementary *in silico* and experimental analysis argues that interface add-ons are an evolutionary strategy to prevent the formation of non-physiological complexes by specializing protein-protein interactions.

RESULTS

Identification of interface add-ons

Interfaces of protein complexes can be partitioned into highly conserved core and more variable rim regions (Bouvier et al., 2009; Guharoy and Chakrabarti, 2005). We speculated that this variability can lead to diverse peripheral geometries which might contribute to protein-protein interaction specificities within heteromeric complexes. For a systematic computational assessment of interface geometries, we combined structural information from PDB entries (Berman et al., 2000) with the sequences provided by the corresponding InterPro families (Hunter et al., 2012), which contained on average 12 000 homologs. Our protocol (**Figure 1A**) comprises seven filter routines that systematically refine the specification of an interface add-on and simultaneously narrow down the number of candidates (**Table S1**).

The survey was based on those 1739 heteromeric bacterial protein complex structures deposited in PDB that were devoid of non-protein macromolecules and had subunit stoichiometries of AB, A_2B_2, A_3B_3, A_4B_4, A_6B_6, ABC, and $A_2B_2C_2$. Removing identical proteins and focusing on structures that are associated with family entries in InterPro reduced the number to 305 complexes. In the following, we name them "reference complexes" and use *SU* to address one of the subunits A, B, or C. For each subunit *SU* and all its InterPro homologs *H* we computed pairwise sequence alignments PW(*SU*, *H*) (**Figure 1B**) and eliminated all PW(*SU*, *H*) of poor quality as well as heavily fragmented ones. Subsequently, those PW(*SU*, *H*) were identified that showed in *SU* or *H* an additional fragment containing at least eight residues. Thus, our approach excludes minor alterations of interface topologies that are related to shorter indels and identifies fragments that can fold into defined secondary structure elements.

For the computational analysis of the remaining PW(*SU*, *H*), two cases had to be distinguished: First, *SU* can contain one or more additional fragments (insertions) that are not present in the aligned homolog *H*. Second, *H* can contain one or more additional fragments that are not present in *SU* (deletions). To assess the first case, the insertions resulting from all PW(*SU*, *H*) were mapped onto the sequence of *SU* and a histogram was computed that specifies the number of insertions overlapping each residue position (**Figure 1B**); for details see Experimental Procedures. Subsequently only those insertions were considered further that exceeded a pre-defined significance threshold; compare grey area in the histogram shown in **Figure 1B** and the examples in **Figure S1A**. Thus, 209 insertions were found in 117 of the reference complexes with most of insertions comprising between 10 and 20 amino acids (**Figure 1C**). To assess the second case, a histogram specifying the number of deletions beginning at each residue position of *SU* was analyzed analogously. In total, we identified 392 deletions associated with 212 reference complexes; examples are shown in **Figure S1B**. A detailed characterization of the corresponding insertions in *H* is difficult, because their structures are not known and in homology models the local topology of such large insertions is often unreliable (Webb and Sali, 2014). Moreover the subsequent docking of the model into the given reference complex introduces further errors; thus we did not further examine these cases.

In contrast, the 209 fragments that correspond to an insertion in *SU* could be analyzed in detail. To begin with, we identified 162 insertions in 95 reference complexes that contained at least one interface residue (IFR); see Experimental Procedures. Next, we assessed *in silico* the contribution of these insertions to complex stability. The algorithm mCSM estimates the effect of mutations on protein stability and protein-protein affinity and the resulting $\Delta\Delta G$ values correlate well with experimental findings (Pires et al., 2014).

We used mCSM for an alanine scanning of all 162 insertions by mutating each non-alanine IFR individually to alanine. According to mCSM classification, a large number of the *IFR→Ala* mutations are highly destabilizing for the complex ($\Delta\Delta G_{IFR\rightarrow Ala}^{complex}$ < 2 kcal/mol) (**Figure 1D**). Consistently, alanine mutations of interface residues that are crucial for complex stability (hot-spots) typically result in similar $\Delta\Delta G$ values (Bogan and Thorn, 1998; Thorn and Bogan, 2001). It is important to note that these effects are not caused by destabilized subunits, because 90% of the corresponding $\Delta\Delta G_{IFR\rightarrow Ala}^{subunit}$ values deduced from subunit structures are above -1.0 kcal/mol (**Figure 1D**).

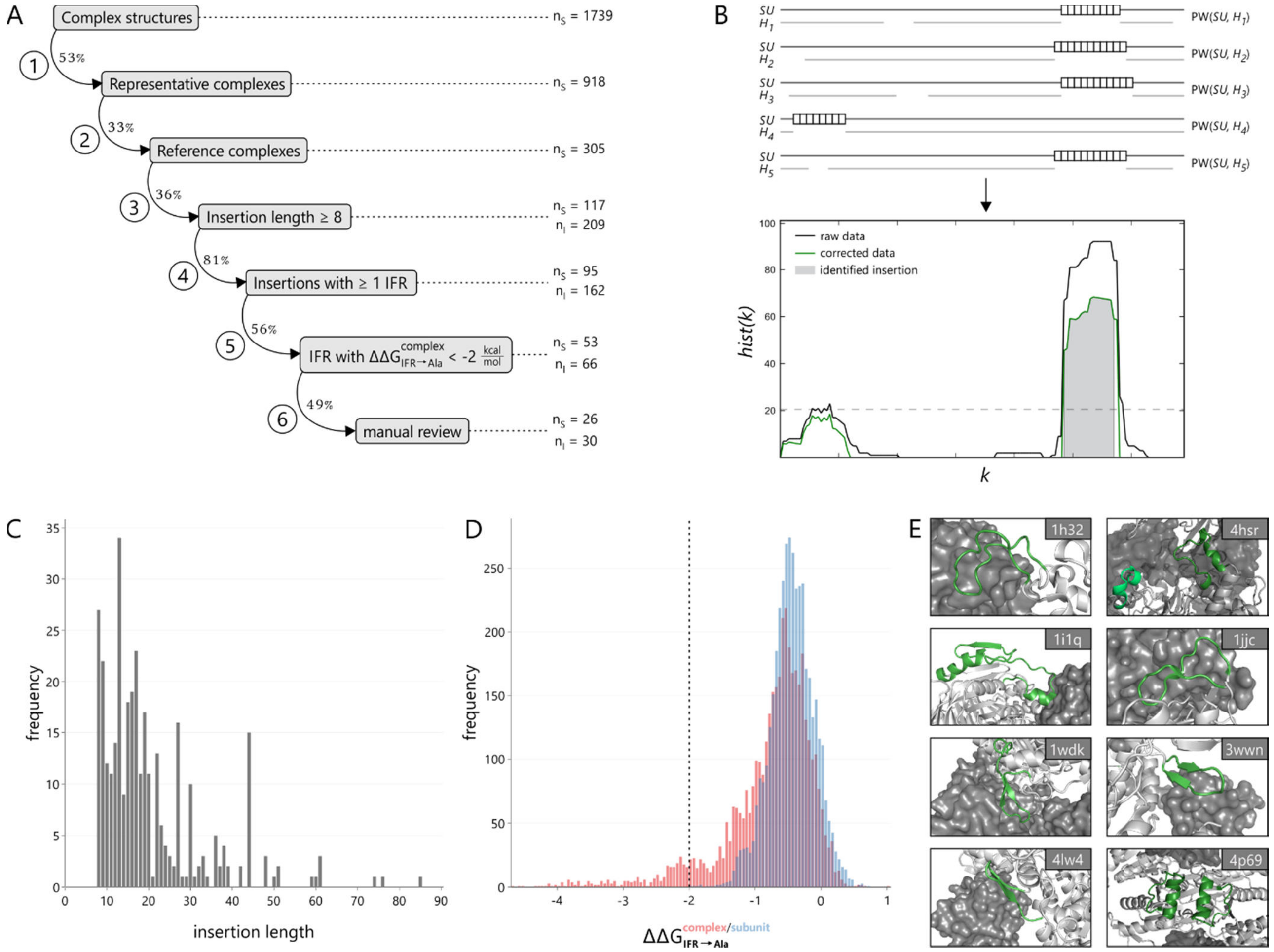


Figure 1: Survey of interface add-ons in heteromeric protein complexes.

A) Step diagram providing an overview of the survey for interface add-ons in 1739 bacterial, heteromeric protein complexes from the Protein Data Bank (PDB). For each of the six steps the proportion of reference structures passed to the next step is given. The total number of structures (n_S) and identified insertions (n_I) is given for steps 3-6. A detailed listing is provided in **Table S1** and the final set of interface add-ons is listed in **Table S2**. **B)** Analysis of pairwise alignments PW(*SU*, *H_n*) for the identification of insertions in a subunit of a reference complex (*SU*), relative to its homologs (H_1-H_5) from the corresponding InterPro family entry. Insertion with a length of at least eight residues (white boxes) were mapped onto each residue position *k* of *SU*, resulting in the counts *hist*(*k*). After correcting for potential noise, insertions (grey areas) were identified as described in **Experimental Procedures**. **C)** Histogram of insertion lengths. The majority of insertions contain between ten and twenty residues. One outlier with a length of 165 residues is not shown. **D)** Histogram of predicted protein-protein affinity change values $\Delta\Delta G_{IFR\rightarrow Ala}^{complex}$ (red) and $\Delta\Delta G_{IFR\rightarrow Ala}^{subunit}$ (blue). The dashed vertical line indicates the threshold of -2 kcal/mol used for classifying interface add-ons. **E)** Examples of heteromeric protein complexes that contain interface add-ons. Subunits with interface add-ons are shown in cartoon representation; other subunits of the complexes are shown in surface representation. The add-ons are colored green. A detailed description of the examples is provided in **Table S3**.

Assuming that interface add-ons contribute little to the stability of the subunits but considerably to the stability of the complex, we classified an insertion as an interface add-on, if at least one $\Delta\Delta G_{IFR\rightarrow Ala}^{complex}$ value was below -2 kcal/mol. In the end, a total of 30 interface add-ons in 26 reference complexes passed a final manual control. Thus, a conservative estimation suggests that about 10% of bacterial, heteromeric protein complexes contain interface add-ons (**Table S2**).

Interface add-ons in heteromeric, bacterial protein complexes

Interface add-ons generally form well defined secondary-structure elements (**Figure 1E**) and, on average, comprise 23 amino acids of which 63% are involved in protein-protein interactions. The corresponding complexes are involved in a variety of key biological functions like amino acid and pyrimidine biosynthesis, β-oxidation of fatty acids, biosynthesis of aminoacyl-tRNAs, sulfur metabolism, respiratory chains, biosynthesis of antibiotics, and the citric acid cycle. A detailed description of selected examples in the context of the respective protein families is provided in **Table S3**. On average, these interface add-ons are present in about 8% of the InterPro homologs and show high sequence conservation.

Three examples shall be highlighted in the following: The reference structure 4lw4 (**Figure 1E, Table S3**) is a hetero-tetrameric CsdA$_2$:CsdE$_2$ cysteine desulfurase. CsdA catalyzes the transfer of sulphur from cysteine or selenocysteine to the acceptor subunit CsdE in the cysteine-sulfinite-desulfinase system (Loiseau et al., 2005). CsdA contains an interface add-

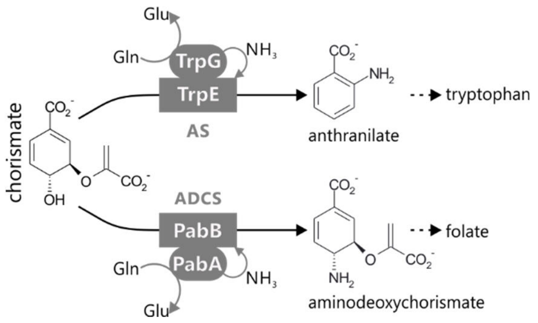


Figure 2: Reactions catalyzed by AS and ADCS in tryptophan and folate biosynthesis.
Both the glutaminase subunits (ellipses) and the synthase subunits (rectangles) are homologs.

on of 14 residues that form two anti-parallel β-strands of the interface with CsdE. Importantly, *E. coli* not only contains the CsdA$_2$:CsdE$_2$ complex but also the IscS$_2$:TusA$_2$ cysteine desulfurase complex in which IscS is homologous to CsdA. The IscS$_2$:TusA$_2$ complex is part of a different biosynthetic branch, namely biogenesis of iron-sulfur clusters (Mihara and Esaki, 2002). IscS does not contain the CsdA interface add-on and binds to the acceptor subunit TusA instead of the structurally different CsdE. A second example is the reference structure 4p69 (**Figure 1E, Table S3**), which is the hetero-tetrameric complex of the *E. coli* isocitrate dehydrogenase IcdA and the isocitrate-dehydrogenase-kinase/phosphatase AceK. The latter is important for phosphorylating and dephosphorylating IcdA as a means to regulate isocitrate flux between the citric acid cycle and the glyoxylate cycle (LaPorte, 1993; LaPorte and Koshland, 1982). IcdA contains an interface add-on consisting of 24 residues that fold into an α-helix located in the interface between the two IcdA and the two AceK chains. The isopropylmalate dehydrogenase LeuB from *E. coli*, which is part of leucine biosynthesis, is homologous and functionally related to IcdA. However, LeuB lacks the interface add-on of IcdA and no interactions with kinases/phosphatases like AceK have been described.

The third example is the reference structure 1i1q (**Figure 1E**, **Table S3**), which is the anthranilate synthase from *Salmonella typhimurium*. Anthranilate synthases (AS) are hetero-tetrameric GATase complexes that catalyze the initial step in the biosynthesis of the essential amino acid tryptophan (Zalkin, 1973). Their glutaminase subunits TrpG hydrolyze glutamine to glutamate. The concomitantly formed ammonia is subsequently channeled to the synthase subunits TrpE (Raushel et al., 2003), where it reacts with chorismate to anthranilate (**Figure 2**). The interface add-on is located in TrpE and, with a length of 51 residues, is one of the largest add-ons indentified in our survey. It folds into two α-helices connected by two β-strands and prominently protrudes into the TrpE:TrpG dimer interface with loop L2 and α-helix H2 (**Figure 3A**). Although the interface add-on as a whole contains only few conserved residues, the two α-helices contain several conserved hydrophobic (H1) and charged/polar (H2) amino acids, respectively (**Figure 3B**). In contrast to the AS from *S. typhimurium*, those from *Mycobacterium tuberculosis* and *Sulfolobus solfataricus*, for example, contain TrpE subunits without this interface add-on. In the following, we thus use the term "TrpE" to refer to AS synthase subunits that do not contain the interface add-on and use the term "TrpEx" ("extended") to refer to AS synthase subunits that contain the interface add-on.

Some organisms not only possess AS (TrpEx$_2$:TrpG$_2$) but also the related complex aminodeoxychorismate synthase (ADCS, PabB:PabA), which catalyzes the first step of folate biosynthesis (**Figure 2**) (Dosselaere and Vanderleyden, 2001). Its synthase subunit PabB is homologous to TrpEx, but lacks the interface add-on, and its glutaminase subunit PabA is homologous to TrpG. A sequence-similarity-network (Gerlt et al., 2015) of the common InterPro family of TrpEx, TrpE, and PabB sequences shows that TrpE and PabB proteins, which both do not contain the interface add-on, are highly similar in sequence as they tightly cluster together (**Figure 3C**). In contrast, TrpEx sequences are grouped into a distinct subcluster. A comparison of representative TrpEx, TrpE, and PabB structures shows that the interface add-on is the only major structural difference between these three homologs (**Figure 3D**).

Three observations suggest that the interface add-on in TrpEx acts as an explicit negative design element to differentiate TrpEx from PabB and thus ensure the specific formation of AS (TrpEx$_2$:TrpG$_2$) and ADCS (PabB:PabA) complexes: First, its location in the interface between TrpEx and TrpG. Second, the high conservation of charged and polar residues in its interface helix H2, a characteristic of insertions that modulate the association of protein complexes (Hashimoto and Panchenko, 2010). Third, the fact that some organisms like *E. coli* contain homologous ADCS complexes whose glutaminase subunit PabA is highly similar to the TrpG glutaminase subunit of AS. Consequently, the interface add-on in TrpEx may prevent the putative cross-interactions TrpEx-PabA and PabB-TrpG and thus the formation of non-physiological complexes.

Phylogenetic distribution of AS and ADCS complexes

In contrast to the genomes of γ-Proteobacteria like *E. coli* and *S. typhimurium* that contain the four genes coding for the individual AS and ADCS complexes, the genome of *Bacillus subtilis*, for example, contains the genes for TrpE and PabB, but only a single glutaminase gene annotated as *pab*A. This specific situation in *B. subtilis* has been investigated extensively (Gollnick et al., 2005; Slock et al., 1990) and it has been shown that the single PabA glutaminase serves both TrpE and PabB. To get an overview of such gene co-occurrences, we determined phylogenetic distributions of *trp*Ex, *trp*E, *trp*G, *pab*B, and *pab*A across more than 15 000 bacterial and archaeal species. To this end, we developed a computational genetic profiling routine that utilizes BLAST and Hidden-Markov-Models (HMMs, **Figure S3**) to find, annotate, and distinguish the homologous TrpEx/TrpE/PabB synthases and TrpG/PabA glutaminases in these species with high sensitivity and selectivity (**Figure S4**). HMM-approaches have proven successful to distinguish highly similar sequences or to find distantly related homologs (Söding, 2005; Yoon, 2009).

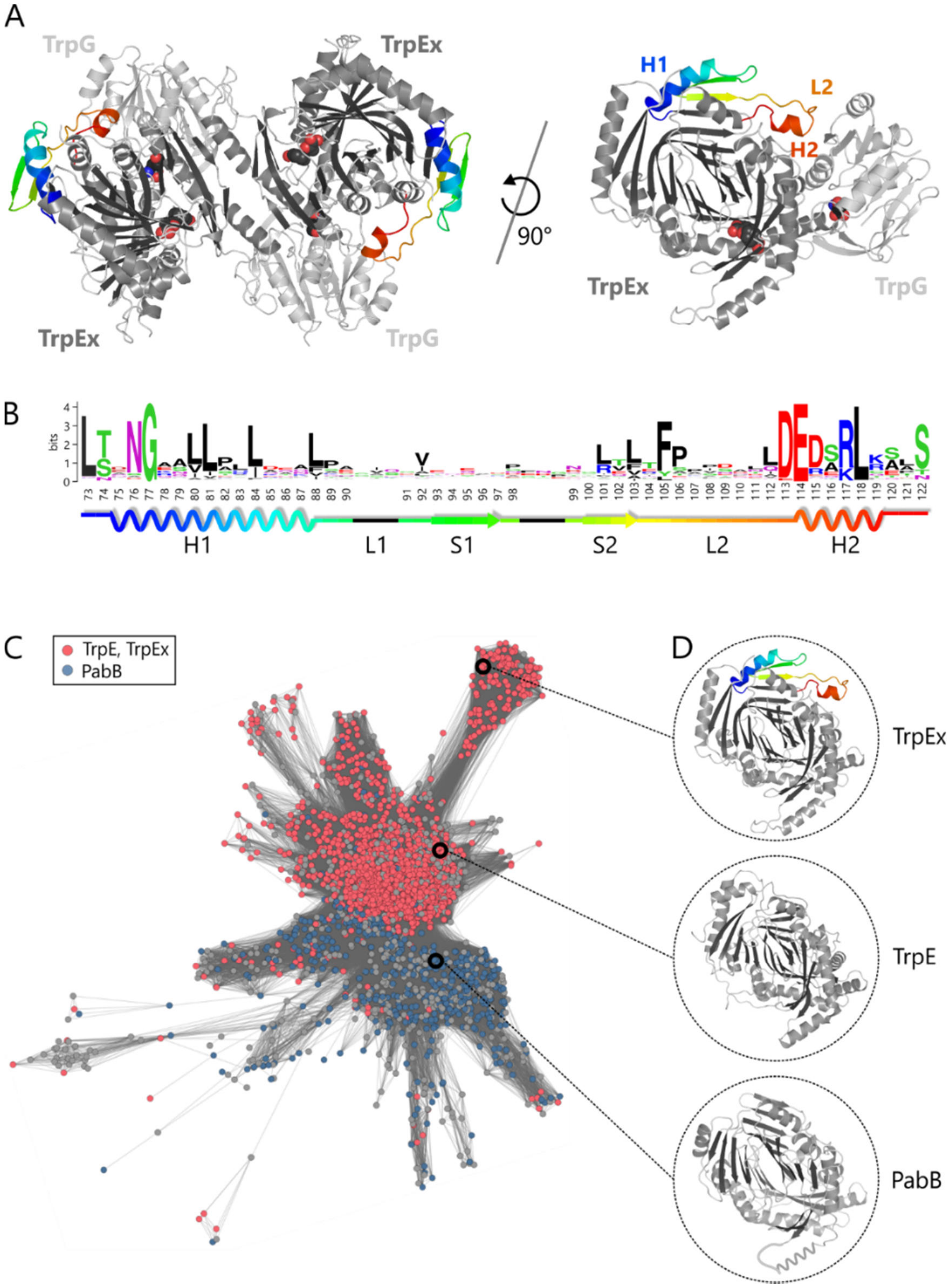


Figure 3: Sequence-structure relationships in the TrpE(x)/PabB family.

A) The TrpEx$_2$:TrpG$_2$ AS complex from *S. typhimurium* (PDB ID 1i1q) is composed of two TrpEx:TrpG hetero-dimers. The interface add-on is colored in a rainbow gradient. Active sites of TrpEx and TrpG are indicated by superimposed space-filling ligand models (from 1i7q). The structure of the TrpEx:TrpG heterodimer on the right hand side shows that loop L2 and α-helix H2 of the interface add-on protrude into the dimer interface. **B)** Sequence logo of the TrpEx interface add-on. Amino acids are colored according to their chemical properties. Residue numbers are according to 1i1q. The 2D elements of the interface add-on are shown in a rainbow-color cartoon representation. **C)** The largest cluster of the SSN generated for this family (IPR019999) at an E-value cut-off of 1E-77. It contains all TrpE(x) and most PabB sequences. Nodes are colored according to the annotation of InterPro. Grey nodes represent sequences with ambiguous annotation. The full network is provided in the **Supplemental Information** (**Figure S2**). **D)** Crystal structures of TrpEx from *S. typhimurium* (1i1q), TrpE from *M. tuberculosis* (4pen), and PabB from *E. coli* (1k0e, a helix that is not resolved is sketched in cartoon representation). The structures are linked to the corresponding nodes in the SSN. The TrpEx interface add-on is colored in a rainbow gradient.

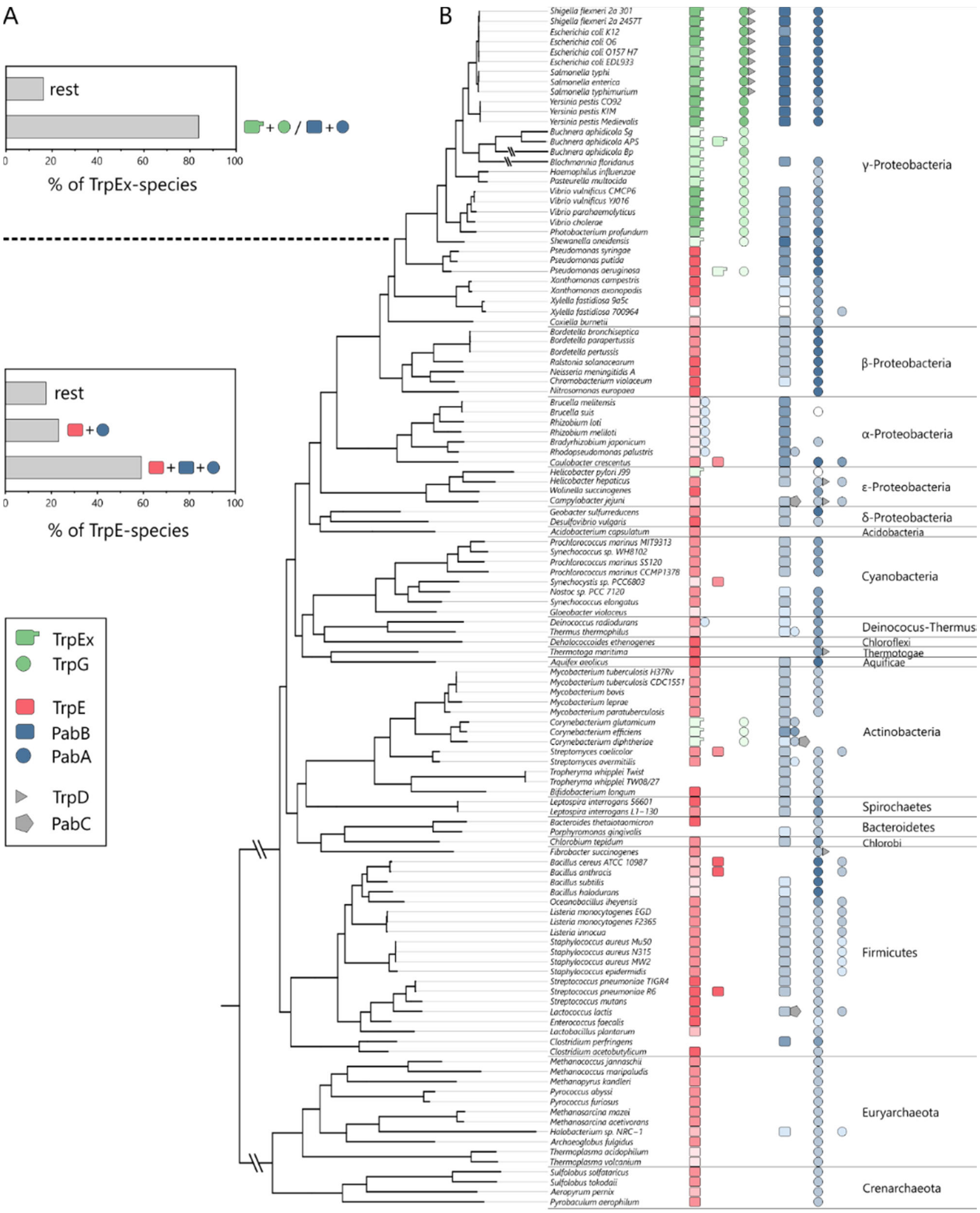


Figure 4: Co-occurrences of synthase and glutaminase subunits of AS and ADCS and their phylogenetic distribution.
A) Fractions of TrpEx- and TrpE-species that possess a certain pattern of co-occurring synthases and glutaminases. For details see **Table S4**. **B)** Phylogenetic tree showing the co-occurrences in 120 representative archaeal and bacterial species. The split between TrpEx- and TrpE-species is indicated by a dashed line. For ease of representation some edges were shorted. Protein function is represented by different shapes and colors. The HMM assignment scores from our computational routine are indicated for each function in four color shades that represent the four quartiles of scores from less reliable (lighter shades) to highly reliable (darker shades). For TrpEx scores ranged from 85 to 171, for TrpE and PabB from 2 to 99, for TrpG from 10 to 74, and for PabA from 2 to 73.

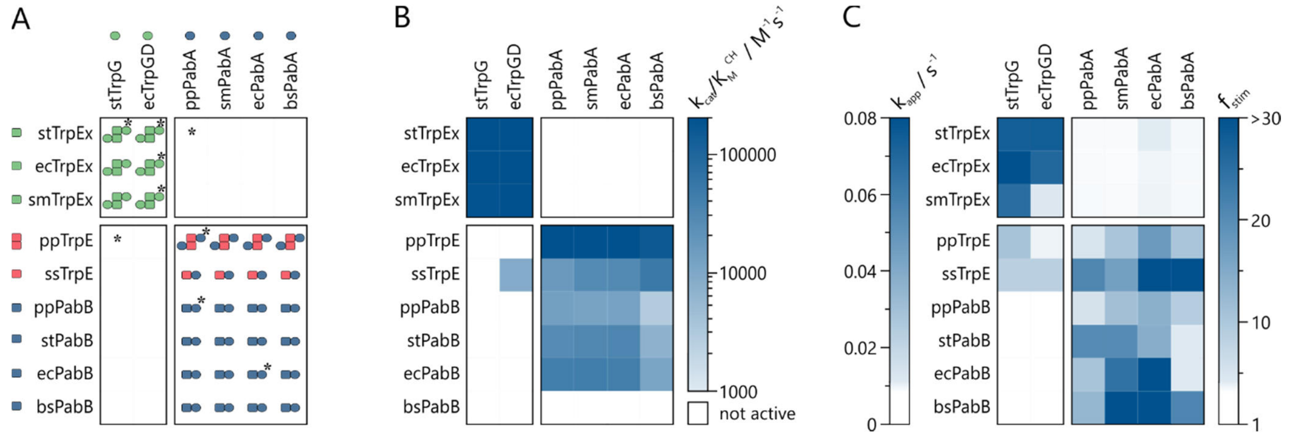


Figure 5: Structural and functional characterization of interactions between TrpEx/TrpE/PabB and TrpG/PabA.
A) Oligomeric states of glutaminase and synthase subunits as well as of cognate and hybrid complexes as determined by SEC-SLS. Blank spaces indicate no complex formation. Combinations that were additionally characterized by native mass spectrometry are marked by asterisks. Lists of oligomeric states together with determined molecular weights are provided in **Table S5** for SEC-SLS and in **Table S6** for native mass spectrometry. Spectra are provided in **Figure S6**. **B)** Catalytic efficiencies for the glutamine-dependent conversion of chorismate (CH) to anthranilate (combinations involving TrpEx and TrpE) and aminodeoxychorismate (combinations involving PabB). Each combination was assayed in triplicates. Exact values with standard deviations are provided in **Table S7**. bsPabB did not display ADCS-activity with any of the available glutaminases under the applied experimental conditions. **C)** Apparent turnover rates of glutamine hydrolysis by TrpG (k_{app}) and stimulation of PabA glutamine hydrolysis by the presence of TrpEx, TrpE, and PabB (f_{stim}). Each combination was assayed in triplicates in the presence of 4 mM glutamine. Exact values with standard deviations are provided in **Table S8**.

In brief, we created five HMMs to represent each type of synthase and each type of glutaminase. We applied BLAST to scan bacterial and archaeal genomes for homologs of the synthases (glutaminases) and compared the respective three (two) HMM scores of the hits to assign them as TrpEx, TrpE, or PabB (TrpG or PabA). The dataset was made non-redundant and sequences with low HMM scores (ambiguous assignments) were rejected. Eventually, we determined the co-occurrences of synthases and glutaminases for 1463 species that constitute the TrpEx sub-cluster (TrpEx-species) and for 4386 species that constitute the TrpE sub-cluster (TrpE-species) of the SSN shown in **Figure 3C**. The phylogenetic distribution of the co-occurrences was derived by mapping them onto a tree-of-life comprising a representative set of archaeal and bacterial species (Letunic and Bork, 2007). representative set of archaeal and bacterial species (Letunic and Bork, 2007).

As expected, the majority of TrpEx-species (84%) possess both the synthase-glutaminase pairs TrpEx-TrpG and PabB-PabA (**Figure 4A**). In not more than 16% of the genomes, corresponding genes were present in multiple copies or missing (**Table S4**). Most of the TrpEx-species are γ-Proteobacteria and belong preferentially to the orders Vibrionales and Enterobacteriales (**Figure 4B**). The evolutionary oldest TrpEx-species are *Shewanella*, an offspring of Pseudomonadales and Xanthomonadales. Most plausibly, TrpEx and TrpG have emerged in *Shewanella* and have been conserved in the γ-proteobacterial lineage since then. Apart from γ-Proteobacteria, TrpEx and/or TrpG are only present in *Helicobacter pylori*, *Pseudomonas aeruginosa*, and the Corynebacteria *C. diphteriae*, *C. efficiens*, and *C. glutamicum*, which is a result from horizontal gene transfer (Farrow and Pesci, 2007; Xie et al., 2003).

All other species, including archaea and all major bacterial phyla are TrpE-species. 59% of them possess, like *B. subtilis*, TrpE and PabB, but only PabA and no TrpG glutaminases (**Figure 4A**). Some TrpE-species (23%) lack PabB and possess only TrpE and PabA. Among them are Euryarchaeota and Crenarchaeota, which lack the classical folate-biosynthesis genes and use alternative biosynthetic pathways or rely on methanopterin-related methyl donors instead of tetrahydrofolate (White, 1988; Worrell and Nagle, 1988). In not more than 18% of TrpE-species genomes, corresponding genes were present in multiple copies or missing (**Table S4**). Taken together, TrpEx-species generally contain a full set of synthases (TrpEx, PabB) and glutaminases (TrpG, PabA). In contrast, TrpE-species generally contain one or both synthases (TrpE and/or PabB) and only one type of glutaminase (PabA). Consequently, in TrpE species, PabA has to interact with both synthases, whereas a more specific interaction pattern seems plausible for TrpEx-species.

Protein-protein interaction specificity in AS and ADCS complexes

We experimentally analyzed the influence of the TrpEx interface add-on on the specificity of glutaminase-synthase interactions in AS and ADCS complexes. For this purpose, we expressed and purified three TrpEx, two TrpE, and four PabB representatives: TrpEx from *S. typhimurium* (stTrpEx), *E. coli* (ecTrpEx), and *S. marcescens* (smTrpEx); TrpE from *Pseudomonas putida* (ppTrpE) and *Sulfolobus solfataricus* (ssTrpE); PabB from *P. putida* (ppPabB), *S. typhimurium* (stPabB), *E. coli* (ecPabB), and *B. subtilis* (bsPabB). The identity of the synthases was validated by an HPLC assay showing that all *bona fide* TrpEx, TrpE, and PabB enzymes formed the expected anthranilate and aminodeoxychorismate, respectively (**Figure S5**).

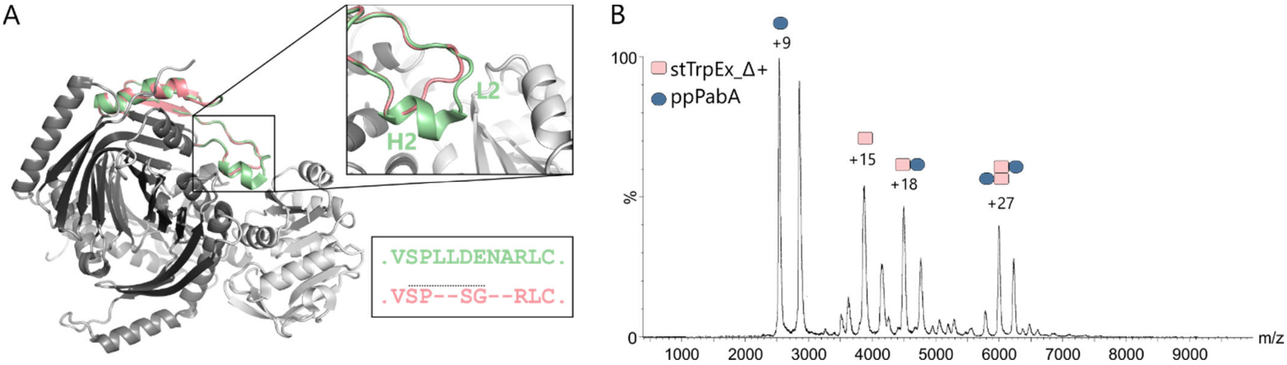


Figure 6: Design of stTrpEx_Δ and mass-spectrometric characterization of stTrpEx_Δ:ppPabA complexes.
A) Overlay of the structure of the stTrpEx:stTrpG dimer from *S. typhimurium* (PDB ID 1i1q) with a model of stTrpEx_Δ. The TrpEx subunits are colored dark grey, the TrpG subunit light grey. The interface add-on of stTrpEx is colored green, the corresponding region of stTrpEx_Δ red. The model of stTrpEx_Δ was generated with the homology modelling module of YASARA 16.4.6 using 1i1q, chain A as template with default settings (Krieger et al., 2009). The cutout shows a detailed view of the modified interface add-on part. The insert shows an alignment of the interface add-on regions detailed in the cutout. The residue sequence LLDENA in stTrpEx (green) was replaced with SG in stTrpEx_Δ (red) to mimic a type-I β-turn (dashed line). **B)** Representative mass spectrum of an equimolar mixture of stTrpEx_Δ and ppPabA (20 μM each). Pictograms denote monomers and complexes. Charges of the most populated charge species are included.

Furthermore, we expressed and purified two TrpG and four PabA homologs: TrpG from *S. typhimurium* (stTrpG) and *E. coli* (ecTrpGD); PabA from *P. putida* (ppPabA), *S. marcescens* (smPabA), *E. coli* (ecPabA), and *B. subtilis* (bsPabA). ecTrpG could only be solubly expressed as a fusion construct with TrpD.

We analyzed the ability of the synthases to form complexes with the glutaminases by analytical size exclusion chromatography in combination with static light scattering (SEC-SLS) and by native mass spectrometry. We made three striking observations, which are schematically illustrated in **Figure 5A**.

First, all three TrpEx synthases exclusively interacted with the two TrpG glutaminases. These combinations resulted in tetrameric complexes, in accordance with the oligomeric states deduced from AS crystal structures (Morollo and Eck, 2001; Spraggon et al., 2001). In other words, none of the TrpEx synthases formed complexes with any of the PabA glutaminases. The absence of any interaction between TrpEx and PabA was confirmed by native mass spectrometry for the stTrpEx/ppPabA combination at a concentration similar to those that produced native TrpEx:TrpG complexes (**Figure S6**).

Second, all TrpE and PabB synthases interacted with all PabA but none of the TrpG glutaminases. The TrpE:PabA complexes were dimers or tetramers, whereas the PabB:PabA complexes were exclusively dimeric. The complexes ppTrpE:ppPabA, ppPabB:ppPabA, and ecPabB:ecPabA were validated by native mass spectrometry, as was the absence of a complex for the ppTrpE-stTrpG combination (**Figure S6**). Based on these results, we regard the denomination PabA (Huang and Gibson, 1970) as anecdotal, because the respective glutaminases are not only part of ADCS but also of AS complexes in most microbial species.

Third, interactions between TrpEx and TrpG on the one hand and between TrpE/PabB and PabA on the other hand are conserved across species and kingdom borders. All 15 tested enzymes were able to interact with a partner from a different organism, independent of their positions in the tree of life. For instance, TrpE from the Crenarchaeon *S. solfataricus* formed complexes with PabA from phylogenetically distant Firmicutes (bsPabA) and γ-Proteobacteria (ppPabA, smPabA, and ecPabA).

Functional characterization of cognate and hybrid AS and ADCS complexes

We followed the glutamine-dependent conversion of chorismate (CH) to anthranilate (AA) and aminodeoxychorismate (ADC) for all 54 possible combinations of synthases and glutaminases by continuous fluorimetric assays. The results are summarized in **Figure 5B**. Notably, all stable complexes detected by SEC-SLS, except those containing bsPabB, were catalytically active. AS complexes displayed catalytic efficiencies k_{cat}/K_M^{CH} between $8.7 \cdot 10^3$ and $1.3 \cdot 10^6$ $M^{-1}s^{-1}$ with practically no differences in the highest efficiencies of TrpEx:TrpG and TrpE:PabA complexes. ADCS complexes (PabB:PabA) converted CH to ADC with catalytic efficiencies k_{cat}/K_M^{CH} between $9.9 \cdot 10^2$ and $2.8 \cdot 10^4$ $M^{-1}s^{-1}$. These data show that functional AS and ADCS complexes can be formed by synthases and glutaminases from different species. This indicates a strong conservation of the synthase-glutaminase interface, because functional complexes require efficient channeling of nascent ammonia between the subunits. We could not detect measurable AS or ADCS activity for all "non-interacting" pairs. The only exception was ssTrpE which did not only display AS activity with all four PabA glutaminases but also to a certain degree with ecTrpGD, indicating the formation of a transient complex during catalysis.

It is known that glutaminase activity in GATase complexes is allosterically stimulated by the synthase (Bera et al., 1999; Goto et al., 1976; Miles et al., 1998). In order to quantify this effect for the 54 combinations of glutaminases and synthases, we incubated TrpG and PabA with glutamine and monitored its hydrolysis before and after the addition of TrpEx, TrpE, or PabB (**Figure S7**). The results are illustrated in **Figure 5C**. We found that both TrpG representatives do not

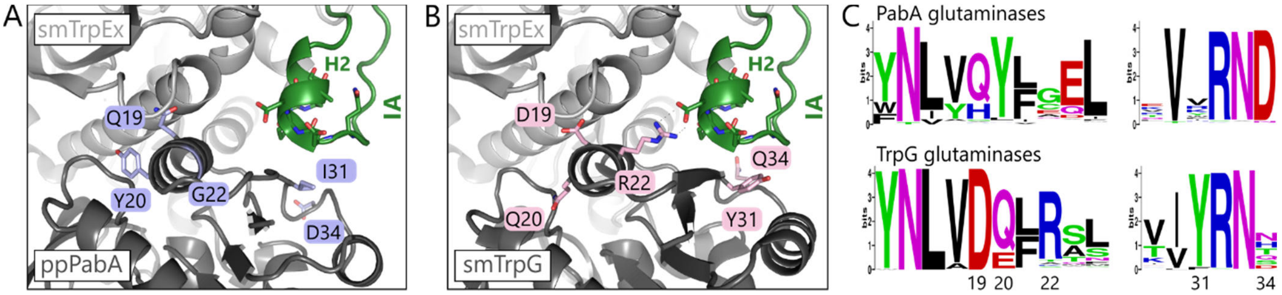


Figure 7: Generation of ppPabA* by interface design.
A) Predicted interface region in an artificial smTrpEx:ppPabA complex. A 3D model of ppPabA was generated with iTASSER (Zhang, 2008) using default parameters and superimposed onto smTrpG in the crystal structure of the smTrpEx:smTrpG complex (PDB ID 1i7q). Positions of ppPabA selected for mutation are highlighted in blue, the smTrpEx interface add-on in green. **B)** Interface region of the smTrpEx:smTrpG dimer (1i7q). Residues highlighted in pink correspond to those chosen for mutations in ppPabA. **C)** Sequence logos of the highlighted residues and their surroundings.

display measurable activity in the absence of TrpEx, in accordance with previous findings on stTrpG (List et al., 2012) and other related glutaminases (Beismann-Driemeyer and Sterner, 2001; Miles et al., 1998; Strohmeier et al., 2006). The presence of each of the TrpEx synthases, however, leads to apparent turnover rates for glutamine hydrolysis of up to 0.07 s^{-1}. In most cases the presence of TrpE/PabB, as expected, did not lead to glutamine hydrolysis activity. Only the presence of ppTrpE and ssTrpE activated the TrpGs to a certain degree.

In contrast to a previous report (Roux and Walsh, 1992), all PabA glutaminases proved to be slightly active already in the absence of any synthase; the apparent turnover rates were between 0.003 and 0.015 s^{-1}. When supplemented with TrpE or PabB, the rates increased up to 70-fold. Thus, PabA glutaminases are stimulated by allosteric signals from both TrpE and PabB. As expected, the presence of the three TrpEx synthases did not stimulate their glutaminase activity.

The interface add-on determines interaction specificity in AS and ADCS complexes

The hitherto presented data strongly indicate that the interface add-on in TrpEx synthases is a negative design element that prevents the binding of PabA glutaminases and only allows the formation of correct TrpEx-TrpG pairs. As an ultimate test, we deleted six residues in α-helix H2 and the adjacent loop L2 of the interface add-on from stTrpEx while leaving the remaining interface to stTrpG untouched (**Figure 6A**). The resulting stTrpEx_Δ variant could be expressed in soluble form and purified. It was enzymatically active for the formation of anthranilate with ammonium chloride as the nitrogen source; the catalytic efficiency (k_{cat}/K_M^{CH} = 1.6 ± 0.1 · 10^3 $M^{-1}s^{-1}$) was identical to that of stTrpEx (k_{cat}/K_M^{CH} = 1.5 ± 0.01 · 10^3 $M^{-1}s^{-1}$).

Strikingly, stTrpEx_Δ was able to bind to ppPabA, with mass spectrometry demonstrating a tetrameric stTrpEx_Δ$_2$:ppPabA$_2$ complex (**Figure 6B, Table 1**). Moreover, the binding of stTrpEx_Δ to ppPabA leads to a 2-fold stimulation of ppPabA glutaminase activity (**Table 1**), compared to the 4.5-fold stimulation of ppPabA by its native interaction partners ppTrpE and ppPabB. This demonstrates that the interaction between stTrpEx_Δ and ppPabA is productive as the catalytic activity of the glutaminase is enhanced by the binding of the synthase. At the same time, the propensity of stTrpEx_Δ to form a tetrameric complex with stTrpG was significantly reduced (**Figure S8**). Neither the stTrpEx_Δ$_2$:ppPabA$_2$, nor the stTrpEx_Δ$_2$:stTrpG$_2$ complex was catalytically active in the context of glutamine-dependent anthranilate formation. Most plausibly, the channel required for transferring ammonia from the glutaminases to stTrpEx_Δ is not functional. In any case, our data clearly show that the TrpEx interface add-on determines interaction specificity in AS and ADCS complexes. A deletion of only six residues was sufficient to convert stTrpEx into a TrpE-like synthase, allowing for binding ppPabA.

Generation of functional TrpEx-PabA interactions by interface design

So far we have demonstrated that the interactions in TrpEx$_2$:TrpG$_2$ and PabB/TrpE:PabA complexes are specific and that synthase-glutaminase cross-interactions are prevented by the interface add-on in TrpEx. Thus possible cross-interactions are physiologically highly unlikely. This raises the question how the transition from TrpE to TrpEx took place and how the orthogonal TrpEx$_2$:TrpG$_2$ – PabB:PabA system has evolved in γ-Proteobacteria. One possible scenario is that TrpG evolved in response to the emergence of the TrpEx interface add-on. However, this scenario implies a cell that temporarily tolerates the existence of a non-interacting and thus non-functional TrpEx-PabA pair, which seems improbable given the essential metabolic function of AS in tryptophan biosynthesis.

Table 1: Structural and enzymatic characteristics of complexes comprising stTrpEx or stTrpEx_Δ and stTrpG or ppPabA

	complex formation[a]	glutaminase activity[b]	
	native MS	f_{stim}	k_{app} /s^{-1}
stTrpEx_Δ + ppPabA	Ex_Δ$_2$:A$_2$ tetramer	2.0 ± 0.1	n.a.
stTrpEx_Δ + stTrpG	Ex_Δ$_2$:G$_2$ tetramer	n.a.	0.160 ± 0.004

[a] Representative spectra are provided in **Figure S8**. A corresponding list of deduced oligomeric states together with determined molecular weights is provided in **Table S9**.
[b] Stimulation factors f_{Stim} are based on the apparent glutamine hydrolysis rate of ppPabA: k_{app} 0.015 s^{-1}. For comparison, f_{Stim} for the combinations of ppPabA with ppPabB and ppTrpE are 5.0 and 4.7, respectively. n.a.: not applicable.

Table 2: Structural and enzymatic characteristics of complexes comprising ppPabA* and TrpEx, TrpE, or PabB synthases.

	complex formation[a]	kinetic parameters of AS formation[b]			glutaminase activity[c]
	native MS	k_{cat} /s^{-1}	$K_{\mathrm{M}}^{\mathrm{CH}}$ /µM	$k_{\mathrm{cat}}/K_{\mathrm{M}}^{\mathrm{CH}}$ /M^{-1}s^{-1}	f_{stim}
ppPabA* + stTrpEx[d]	Ex_2:A_2 tetramer	0.31 ± 0.05	6.2 ± 3.5	$5.6 \cdot 10^4$	41.8 ± 3.0
ppPabA* + ecTrpEx	Ex_2:A_2 tetramer	0.63 ± 0.05	9.6 ± 2.0	$6.7 \cdot 10^4$	45.2 ± 3.3
ppPabA* + smTrpEx	Ex_2:A_2 tetramer	0.04 ± 0.004	5.2 ± 2.8	$8.8 \cdot 10^3$	49.0 ± 2.7
ppPabA* + ppTrpE[e]	E_2:A_2 tetramer	0.9 ± 0.06	8.4 ± 1.8	$1.1 \cdot 10^5$	11.5 ± 3.0
ppPabA* + ppPabB[f]	B:A dimer	0.2 ± 0.03	25.0 ± 2.6	$6.7 \cdot 10^3$	19.1 ± 0.6

[a] Representative spectra are provided in **Figure S9**. A corresponding list of deduced oligomeric states together with determined molecular weights is provided in **Table S10**.

[b] Values are the mean and standard deviation from at least three independent measurements.

[c] Stimulation factors f_{Stim} are based on the apparent glutamine hydrolysis rate of ppPabA*: k_{app} 0.005 s^{-1} (wild-type ppPabA: k_{app} 0.015 s^{-1}).

[d] For comparison: stTrpEx$_2$:stTrpG$_2$ complex: k_{cat} 3.7 ± 0.3 s^{-1}, K_M^{CH} 11.3 ± 2.9 µM, and $k_{cat}/K_M^{\mathrm{CH}}$ $3.5 \cdot 10^5$ M^{-1}s^{-1}.

[e] For comparison: ppTrpE$_2$:ppPabA$_2$ complex: k_{cat} 3.4 ± 0.3 s^{-1}, K_M^{CH} 6.3 ± 1.5 µM, and $k_{cat}/K_M^{\mathrm{CH}}$ $5.4 \cdot 10^5$ M^{-1}s^{-1}.

[f] For comparison: ppPabB:ppPabA complex: k_{cat} 0.2 ± 0.08 s^{-1}, K_M^{CH} 33.3 ± 2.3 µM, and $k_{cat}/K_M^{\mathrm{CH}}$ $6.7 \cdot 10^3$ M^{-1}s^{-1}.

A more plausible evolutionary route is in line with recent findings by Laub and co-workers (Aakre et al., 2015). They showed that novel toxin-antitoxin complexes, which are important for cell defense and viability, can evolve without non-functional inter-stages by passing through promiscuous intermediates with relaxed interaction specificity. A similar evolutionary path might have led from PabA to TrpG via intermediates that display relaxed interaction specificity towards synthases. This would imply that few mutations in PabA are sufficient to allow for TrpEx binding without compromising its function and the interaction with its cognate partner PabB.

To test the plausibility of this evolutionary scenario we first generated a homology model of ppPabA and superimposed it with smTrpG in the smTrpEx$_2$:smTrpG$_2$ crystal structure to yield an artificial smTrpEx:ppPabA complex (**Figure 7A**). Using this model, we identified five interface residues of ppPabA that are located in structural elements that lie close to the interface add-on. We replaced these five residues with the corresponding ones of smTrpG (**Figure 7B**), which resulted in the variant ppPabA*. Four of the substitutions (Q19D, Y20Q, G22R, I31Y) lead to residues that are mainly conserved within TrpG glutaminases whereas the D34Q substitution leads to a residue that is only present in a minor fraction of TrpG homologs (**Figure 7C**) but removes the highly conserved aspartate found in PabA glutaminases.

The ppPabA* variant could be expressed in soluble form and purified. It formed tetrameric complexes with all three TrpEx homologs as shown by native mass spectrometry (**Table 2**). Moreover, all three complexes were functional and converted CH to AA in a glutamine-dependent assay with catalytic efficiencies $k_{\mathrm{cat}}/K_{\mathrm{M}}^{\mathrm{CH}}$ between $8.8 \cdot 10^3$ and $6.7 \cdot 10^4$ M^{-1}s^{-1}. These values are only one order of magnitude lower than those of native TrpEx$_2$:TrpG$_2$ complexes. Moreover, the glutaminase activity of ppPabA* is stimulated about 45-fold by the presence of stTrpEx, smTrpEx, or ecTrpEx. Therefore, ppPabA* possesses the three hallmark features of wild-type glutaminases: (i) Formation of stable glutaminase-synthase complexes, (ii) channeling of nascent ammonia from glutaminase to synthase, and (iii) allosteric stimulation of glutaminase activity.

Notably, ppPabA* displays a relaxed interaction specificity towards synthases, as it retained its ability to form stable and functional complexes with both ppTrpE and ppPabB, which do not contain the interface add-on (**Table 2**). The composition and catalytic efficiencies of these complexes were comparable to those of genuine ppTrpE:ppPabA and ppPabB:ppPabA complexes. Thus, the five substitutions only marginally impair the native ppPabA function but permit an additional functional interaction with TrpEx. PpPabA* is therefore promiscuous with regard to interacting with synthases that contain an interface add-on and with such that do not contain an interface add-on. It may thus represent a promiscuous evolutionary intermediate that fills the gap between the TrpE/PabB-specific PabA glutaminases and the TrpEx-specific TrpG glutaminases.

Experimental evidence for harmful metabolic cross-talk in TrpE species

Our genetic profiling of over 15 000 bacterial and archaeal species has shown that Vibrionales and Enterobacteria have evolved a conserved orthogonal system of glutamine amidotransferase complexes for the two important biosynthetic pathways leading to tryptophan and folate. The observation that TrpEx and TrpG have been retained in all descendants after the split between Pseudomonas and Shewanella species approximately 950 million years ago (Battistuzzi and Hedges, 2009) suggests that two orthogonal AS and ADCS complexes entail some kind of selective advantage like the prevention of metabolic cross-talk.

In principle, such cross-talk is conceivable in bacteria that do not possess TrpEx but TrpE. As shown above, these species only contain PabA glutaminases that form functional AS and ADCS complexes with both TrpE and PabB. This raises the question how ammonia flow is directed towards either tryptophan or folate biosynthesis in these organisms. The existence of sophisticated regulatory mechanisms in Firmicutes illustrates the enormous effort that has been invested by Nature to solve this problem (**Figure S10**).

The central player in regulating tryptophan and folate biosynthesis in *B. subtilis* is the tryptophan-sensing protein TRAP (Babitzke, 1997), which binds excessive tryptophan and exercises transcriptional attenuation and translational control

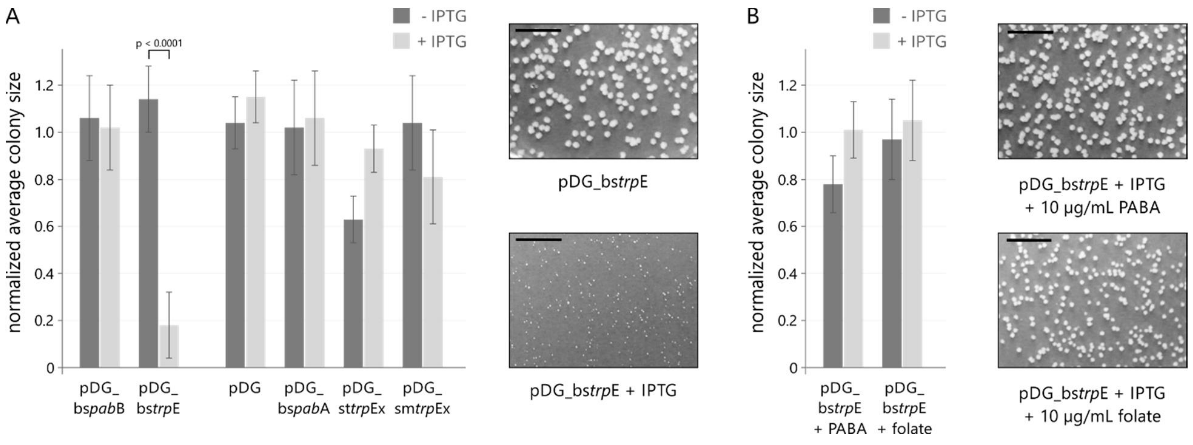


Figure 8: *B. subtilis* growth experiments.
The bars charts show the average sizes of *B. subtilis* colonies grown for 48 hours on minimal medium lacking (dark grey bars) or containing (light grey bars) 2 mM IPTG. The values for the cells transformed with the indicated plasmids are normalized to the average colony size of cells with the empty plasmid (pDG) and grown in the absence of IPTG. Error bars indicate the standard deviation from at least four independent replicates using different transformants. **A**) The overexpression of bs*trp*E leads to a significant reduction of colony size (p-value < 0.0001 at a confidence interval of 99%). The insets show the respective colonies grown in the absence and presence of IPTG. The black line is a reference scale of 1 cm. For representative images of all transformants see **Figure S11**. **B)** The presence of *p*-aminobenzoic acid (PABA) or folate offsets the effect of bs*trp*E overexpression. The insets show the respective colonies grown in the presence of IPTG and either PABA or folate.

both on the tryptophan and the folate operon. A problematic situation develops if biosynthesis of folate is required while cellular levels of tryptophan are high. Under such conditions the production of the single available glutaminase PabA required for folate biosynthesis is blocked by TRAP. However, a dual promoter system allows for the translation-mediated displacement of TRAP resulting in the full ADCS complex required for folate biosynthesis (Yakhnin et al., 2007).

As long as a cell contains TrpE and PabB concurrently, the TRAP-based mechanism cannot guarantee specific direction of ammonia flow to one pathway or the other. Moreover, this sophisticated regulatory network could easily be disturbed resulting in potentially harmful metabolic cross-talk. We simulated such a situation by overexpressing *trp*E in *B. subtilis* and hypothesized that the resulting high amounts of TrpE synthase will take up all PabA glutaminase and thus deduct it from folate biosynthesis. Vice versa, we also assumed that the overexpression of *pab*B will lead to a shortage of PabA glutaminases in tryptophan biosynthesis.

We transformed the prototrophic *B. subtilis* strain SB491 (Zeigler et al., 2008) with expression plasmids that contained either bs*pab*B or bs*trp*E, or bs*pab*A, st*trp*Ex, or sm*trp*Ex as controls. To test the phenotypic effects of IPTG-induced overexpression, we let the transformants grow on defined minimal medium and determined average colony sizes (**Figure 8A**). Overexpression of bs*pab*B had no effect on the average colony size. However, overexpression of bs*trp*E resulted in significantly smaller colonies, which grew to only about 20% of the size observed in the absence of IPTG.

We did not observe similar growth deficiencies for the empty plasmid, nor for the overexpression of bs*pab*A, st*trp*Ex, and sm*trp*Ex. Thus, the growth deficiency under overexpression of bs*trp*E is not caused by toxic effects from IPTG or high protein concentrations. To test if it this effect is caused by a compromised folate biosynthesis, we performed the same growth experiments in the presence of folate or its precursor *p*-aminobenzoic acid (**Figure 8B**). Indeed, the supple-mentation of these metabolites resulted in normally sized colonies, thus offsetting the effect of bs*trp*E overexpression.

DISCUSSION

Interface add-ons are an evolutionary tool for the diversification of protein-protein interfaces and protein-protein interactions

It is intriguing how protein complexes form with the specificity and selectivity required for their proper function in almost all biological processes. A profound understanding of this specificity and selectivity is not possible without detailed knowledge about how protein-protein interfaces and interactions between proteins change and evolve. Given the limited number of different protein architectures (Chothia, 1992; Hashimoto and Panchenko, 2010), quaternary structure topologies (Ahnert et al., 2015), and interface geometries (Gao and Skolnick, 2010; Garma et al., 2012), the diversification of most protein-protein interactions is the result of adaptational mutations (Capra et al., 2012) of interface residues that did not change interface geometry. Although interfaces possess such mutational variability (DePristo et al., 2005; Harms and Thornton, 2013), mutational adaptation is inherently restrained if the interface does not only serve as a binding surface but also performs an additional function like the propagation of an allosteric signal, the completion of an active site, or the channeling of reaction intermediates between the interacting proteins. In such cases, mutations of interface residues, while increasing interaction specificity, may concurrently compromise function. This greater selective constraint imposed on interfaces compared to other regions of proteins is reflected in the lower mutational

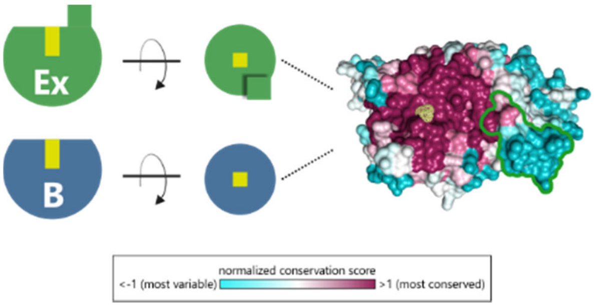


Figure 9: Conservation of the interface region in TrpEx and PabB.
A) On the left TrpEx and PabB are depicted as circle segments with their half of the ammonia channel marked by a yellow box and the TrpEx interface add-on is indicated by a green square. On the right hand side, a top view of the interface region of TrpEx is shown in surface representation (PDB ID 1i7q) with the interface add-on marked by a green outline and the entry to the ammonia channel marked by a yellow mesh. The normalized CONSURF conservation score ranges from below -1 (cyan, most variable) to above 1 (magenta, most conserved) and were calculated from 5849 TrpEx and PabB sequences with a pairwise identity <90%.

rate of interface residues relative to non-interface residues (Ames et al., 2016; Mintseris and Weng, 2005).
Our systematic survey of protein-protein interfaces in heteromeric complexes highlights one solution to such dilemmas: the addition of add-ons to existing interfaces. These interface add-ons have a typical length of 10-20 amino acids and also mostly a well-defined secondary structure. All of them contain at least one residue that falls into the category of a binding hot spot, which is that a mutation of this residue to alanine decreases the binding free energy of at least 2 kcal/mol (Thorn and Bogan, 2001). In fact, many of the interface add-ons contain three or more and some even up to nine hot spot residues. Interface add-ons seem to be quite frequent, consistent with the assumption that negative design elements are important evolutionary traits (Schreiber and Keating, 2011). Under very stringent filter conditions we found them in about 10% of the structures in our representative dataset. Interface add-ons are also not limited to certain phyla as they are present in complexes from Actinobacteria, the Deinococcus-Thermus group, Firmicutes, Thermotogae, as well as several classes of Proteobacteria.

Large interface insertions with similar effects on protein-protein interaction specificity have, to the best of our knowledge, not been identified so far. Though significant insertions have been described in other enzymes, they merely modulate self-association of homo-oligomers (Hashimoto and Panchenko, 2010) or allosteric regulation (Sintchak et al., 2002), but have no impact on interaction specificity. Moreover, interface add-ons are to be discriminated from small, independently folding interaction-mediating domains like ankyrin-repeats (Li et al., 2006), POZ-domains (Bardwell and Treisman, 1994), or bromodomains (Filippakopoulos and Knapp, 2012). These elements are stand-alone mediators of protein-protein interactions and do not change the specificity of an already existing interaction.

The case of the AS and its TrpEx subunit is particularly interesting for several reasons: First, with a length of 51 amino acids, the TrpEx interface add-on is particularly extensive. Such large insertions are rare because of their high risk of impairing protein stability. Consequently, insertions or deletions commonly comprise only one residue and most are shorter than eight residues (Hashimoto and Panchenko, 2010; Pascarella and Argos, 1992). Second, the synthase-glutaminase interface in AS does not only mediate complex formation but is also crucial for ammonia channeling and allosteric communication between the two subunits. Obviously, the TrpEx interface add-ons extends the interface in a way that does not compromise these functional properties. Third, and most importantly, many bacteria, in addition to AS, also contain the homologous ADCS complex, which catalyzes a similar reaction in folate biosynthesis. In these organisms, the TrpEx interface add-on determines the specific formation of AS and ADCS complexes, despite the highly similar core interfaces of PabB and TrpEx (**Figure 9**). This is exemplified by the properties of the stTrpEx_Δ variant where the deletion of six residues allowed for the binding of a PabA glutaminase.

The interface add-on in TrpEx has far-reaching physiological consequences

The *in vivo* growth experiments with *B. subtilis* clearly show that the overexpression of bs*trp*E and the resulting high cellular concentration of TrpE synthases sequesters the available PabA glutaminases, whereby impairing folate biosynthesis and significantly affecting cellular fitness (**Figure 8**). From a broader perspective, we have demonstrated that cross-talk between metabolic pathways can be harmful for an organism. A metabolic conflict comparable to that in Firmicutes does not exist in species that possess TrpEx and its specific, associated glutaminase TrpG. In these species, *trp*G is an integral part of the *trp*-operon and is translationally coupled to *trp*Ex by overlapping start and stop codons, which facilitates an equimolar synthesis of the both AS components (Oppenheim and Yanofsky, 1980). Consequently, the synthesis of the two AS components can be controlled independent of cellular folate levels and a simple repression and attenuation mechanism is sufficient to regulate the *trp*-operon (Kelley and Yanofsky, 1985; Roesser and Yanofsky, 1991; Yanofsky et al., 1984).

Glutaminase intermediates with relaxed interaction specificity enabled the evolution of TrpEx-species

Phylogenetic distribution and sequence similarities of the glutaminases suggest that TrpG has evolved from a PabA ancestor. It is however highly unlikely that the appearance of TrpG was a compensatory response to the emergence of TrpEx. Our experimental characterization of TrpEx shows that it cannot interact with PabA. Thus, an organism containing TrpEx, PabB, and only PabA would be non-viable due to non-functional tryptophan biosynthesis.

Plausible scenarios that lead from a TrpE to a TrpEx species avoid such evolutionary dead-ends by assuming promiscuous PabA* glutaminase intermediates with relaxed interaction specificity. Promiscuity in this context refers to their ability to interact with synthases that contain an interface add-on, i.e. TrpEx, and likewise with synthases that do not contain an interface add-on, i.e. TrpE and PabB. Our designed ppPabA* variant displays such a relaxed interaction specificity: It contains five amino acid substitutions that are sufficient to establish stable and functional interactions with TrpEx. Moreover, it forms stable and functional complexes also with ppTrpE and ppPabB, which both lack the TrpEx interface add-on.

A parsimonious approach building on comparable intermediates leads to two different plausible evolutionary trajectories (**Figure 10**). In the neo-functionalization trajectory (**Figure 10**, left path), the gene of an ancestral PabA was duplicated, thus allowing the acquisition of mutations that lead to a relaxed interaction specificity. Mutational drift and co-evolution (Pazos and Valencia, 2008) finally led to a specialization of PabA* to the contemporary TrpG glutaminase found in γ-Proteobacteria. In the alternative sub-functionalization trajectory (**Figure 10**, right path), the presence of a promiscuous PabA* intermediate made it possible to tolerate the integration of the interface add-on into TrpE. After gene duplication the copies co-evolved with TrpEx and PabB, respectively, and sub-functionalization led to specific, contemporary AS and ADCS complexes. With the data at hand, it is not possible to decide which evolutionary trajectory might more accurately describe the evolutionary history of TrpG. For example, it is known that most duplicated genes do not stay in the gene pool of a population long enough to accumulate function- or specificity-changing mutations and are lost instead (Hughes, 1994; Lynch and Conery, 2000), arguing against the neo-functionalization path. On the other hand, the promiscuity-inducing mutations in a PabA* variant must not be at the expense of catalytic efficiency or protein stability; an important point to consider with the sub-functionalization path. However, irrespective of which trajectory reflects the actual evolutionary path, a promiscuous PabA* intermediate is required to interact with both TrpEx and PabB, ensuring that ammonia is made available for tryptophan and folate biosynthesis. This principle of evolutionary utilization of interaction-promiscuity and the concurrent prevention of non-functional protein complexes has recently also been highlighted on the example of bacterial toxin-antitoxin complexes (Aakre et al., 2015). There, promiscuous interaction partners allowed the expansion and diversification of toxin-antitoxin systems without impairing cell viability. Moreover, the exploit of interaction promiscuity and the gain of interaction specificity by just few changes in sequence have been highlighted by studies of complexes between colicin endonucleases and immunity proteins (Levin et al., 2009).

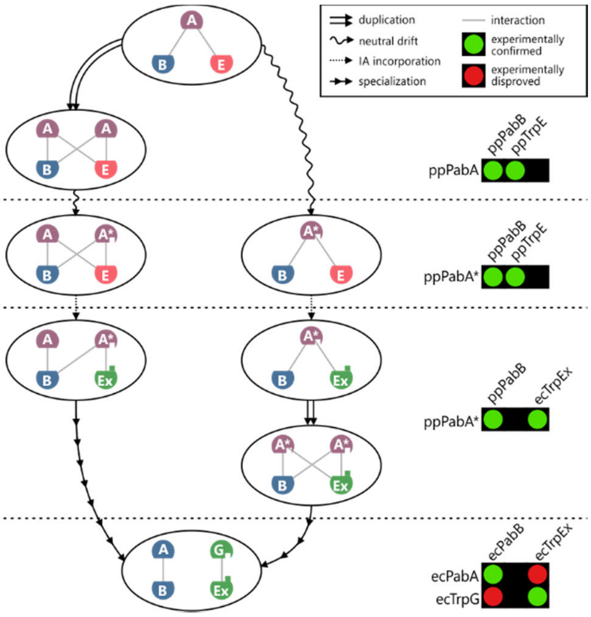


Figure 10: Possible evolutionary trajectories leading to the orthogonal complexes of γ-Proteobacteria.

Two evolutionary scenarios leading to separate AS and ADCS complexes in TrpEx-species. The left path consists of the following steps: duplication of PabA, neutral drift of PabA to PabA*, integration of the interface add-on into TrpE, and specialization of PabA* to TrpG. Within the right path, neutral drift initially leads from PabA to PabA*. After duplication of PabA*, specialization and co-evolution gives rise to contemporary PabA and TrpG enzymes. On the right hand side experimentally characterized interactions from this work are shown that represent the respective evolutionary phases.

Final perspective

To conclude, our work adds to the understanding how highly conserved protein interfaces are tinkered towards novel interaction specificities in the course of evolution. Our systematic characterization of homologous synthases and glutaminases in a GATase family shows that diversification can be achieved by the integration of interface add-ons. We also provide *in vivo* evidence that the incorporation of the interface add-on in TrpEx entails selective advantages by assuring independent regulation of ammonia influx into tryptophan and folate biosynthesis. Importantly, our *in silico* survey of bacterial heteromeric protein complexes suggest that TrpEx is not a single case but that interface add-ons are common mediators of interaction specificity in various physiological contexts. From a biomedical perspective, the TrpEx interface add-on could provide an anchor point for the grafting of peptides that may inhibit the formation of the AS complex and thus impair tryptophan biosynthesis, specifically targeted at pathogenic TrpEx-species. In the context of protein design, interface add-ons might prove as valuable tools for the generation of orthogonal pairs of interacting proteins that all utilize the same core interface.

EXPERIMENTAL PROCEDURES

Survey of interface add-ons in heteromeric protein complexes

The initial dataset contained 1739 heteromeric bacterial protein complex structures deposited in PDB that were devoid of non-protein macromolecules and had subunit stoichiometries of AB, A_2B_2, A_3B_3, A_4B_4, A_6B_6, ABC, and $A_2B_2C_2$. To avoid redundancy, identical proteins crystallized under different experimental conditions were excluded, leaving a subset of 918 complex structures. The InterPro dataset distinguishes protein families, which represent groups of homologous proteins at different levels of functional and structural similarity, and domains, which often occur in numerous non-homologous proteins (Hunter et al., 2012). We thus removed all complex structures whose subunits are only associated with “domain”, “repeat”, and “site” entries and selected those that were assigned to highest-level InterPro families. For the remaining 305 complex structures we extracted all bacterial sequences from the corresponding InterPro families. In the following, we name these complexes "reference complexes" and use *SU* to address one of the subunits A, B, or C; a homologous sequence from the corresponding family is denoted by *H*.

Due to the average number of 12 000 homologs per family, it is difficult to create a reliable multiple sequence alignment. To identify insertions in *SU* or *H*, we thus computed individual pairwise sequence alignments PW(*SU*, *H*) by means of MAFFT (Katoh and Standley, 2013) with the -localpair option and a gap-opening penalty of 3.0. In order to exclude heavily fragmented alignments, only those were considered further that contained a maximum of 40 isolated gaps in *SU*. Furthermore, PW(*SU*, *H*) that contained N- or C-terminal indels were also ignored because these may arise from erroneous sequence annotations. Finally, all PW(*SU*, *H*) were selected that showed in *SU* or *H* an additional fragment comprising at least eight residues.

The insertions resulting from all PW(*SU*, *H*) were mapped onto the sequence of *SU* and a histogram *hist* was computed; *hist*(*k*) specifies for each residue position *k* of *SU* how often it is part of an insertion (for examples see **Figure S1A**). Due to the high number of mappings and the sequence variability of the chosen InterPro families, most residue positions *k* are part of insertions, which results in noisy histograms. After correcting for this noise, the maximal number *max_hist* of all *hist*(*k*) values was determined and all continuous sections with *hist*(*k*) > 0.3 *max_hist* were identified as significant insertions. Analogously, for each *SU* a histogram was computed that specifies how often an insertion in *H* starts at residue position *k* (**Figure S1B**). Again the maximal value *max_hist* of all *hist*(*k*) values was determined and all positions *k* with *hist*(*k*) < 0.5 *max_hist* were identified as starting points of significant insertions in *H*. In total, 209 insertions were identified in 117 reference complexes and 418 insertions were identified in InterPro homologs associated with 226 reference complexes (**Table S1**).

From these 209 insertions we selected those that are part of a protein-protein interface in the respective reference complex structure. To this end for each reference structure the biological quaternary assembly was generated using the PDBePISA server (Krissinel and Henrick, 2007) or taken from author-provided assembly files from the PDB. Then, we calculated for each residue of an insertion the shortest distance to any residue belonging to the other subunits based on the position of their heavy atoms. A residue was designated as an IFR, if this distance was less than 4.5 Å which is a commonly chosen cut-off (Barlow and Thornton, 1983; Gao and Skolnick, 2010; Ofran and Rost, 2003b; Xu et al., 1997).

The contribution of these insertions to protein-protein interactions in the respective complexes was analyzed using mCSM (Pires et al., 2014) with the protein-protein option as follows: All *i* non-alanine IFRs of an insertion were mutated *in silico* to alanine in the full quaternary assembly resulting in *i* predicted protein-protein affinity change values $\Delta\Delta G^{complex}_{IFR\rightarrow Ala}$. Analogously, these *i* IFRs were mutated to alanine in the isolated subunit structure giving *i* $\Delta\Delta G^{subunit}_{IFR\rightarrow Ala}$ values. An insertion was designated as a candidate interface add-on, if at least one mutation resulted in $\Delta\Delta G^{complex}_{IFR\rightarrow Ala} < -2\,\frac{kcal}{mol}$. Finally, after manual inspection, candidate interface add-ons were removed based on the following conditions: i) No 3D structure available for a homolog *H* that does not contain the interface add-on, ii) highly similar duplicates, iii) location of an interface add-on in a homomeric interface, which is present for example in complexes with the stoichiometry A_2B_2, and iv) candidates were N- or C-terminal extensions, not detected by the previous sequence filtering. In the end, 30 interface add-ons in 26 different reference structures remained.

Computation of sequence similarity networks

SSNs of the InterPro entries IPR019999 and IPR015890 were computed according to (Atkinson et al., 2009; Gerlt et al., 2015) and visualized with Cytoscape 3.3 (Shannon et al., 2003; Smoot et al., 2011). A detailed description is available in the **Supplemental Experimental Procedures**.

Computation of amino acid conservation and sequence logos

TrpEx sequences were extracted from the SSN of IPR015890 via their UniProt identifiers and aligned with MAFFT (Katoh and Standley, 2013). The MSA was made non-redundant at 90% identity and contained 210 TrpE sequences. The amino-acid conservation of the TrpEx interface add-on was derived from the MSA and visualized as a sequence logo by means of WebLogo (Crooks et al., 2004). The amino acids in the sequence logo are colored according to their chemical properties: purple, amido functionality; red, acidic; blue, basic; black, hydrophobic; green, hydroxyl functionality and glycine. Sequence logos of PabA and TrpG glutaminases were generated accordingly from MSAs of corresponding sequences in $TrpEx_{repr}$ and $TrpE_{repr}$ (see below).

Genetic profiling of Archaea and Bacteria to determine the phylogenetic distribution of AS and ADCS complexes

To investigate whether the presence of TrpEx or TrpE affects the distribution of PabB and the type and number of associated glutaminases, we determined the occurrence of TrpEx, TrpE, PabB, TrpG, and PabA for all species associated with the TrpEx and TrpE sub-clusters in the SSN of IPR015890. The Hidden-Markov-Model based grouping routine that we developed for this application is described in detail in the **Supplemental Experimental Procedures** and sketched in **Figure S3**. In brief, all species that contribute sequences to the TrpEx and TrpE sub-clusters were individually scanned for the presence of TrpEx, TrpE, and PabB as well as for the presence of TrpG and PabA using BLAST (Altschul et al., 1990). Hits were assigned as either of these five enzymes by sensitive comparison with enzyme-specific Hidden Markov Models (**Figure S12**). Finally, the two representative datasets $TrpEx_{repr}$ and $TrpE_{repr}$ containing the co-occurring proteins for TrpEx- and TrpE-species, respectively, were generated.

Computation of interface conservation

The similarity of the interface regions between TrpEx and PabB was computed from structure-based MSAs comprising sequences from $TrpEx_{repr}$ and $TrpE_{repr}$. A detailed description is available in the **Supplemental Experimental Procedures**.

Cloning and mutagenesis

The genes of TrpEx, TrpE, PabB, TrpG, and PabA proteins were amplified from genomic DNA or whole cell lysate in standard PCR reactions and cloned into pET21a, pET28a, or pMAL-c5X vectors. St*trp*Ex_Δ was generated by deletion of codons 111 to 116 inclusive and inserting the codons AGC and GGC coding for serine and glycine, respectively, in pET21a_st*trp*Ex. using the NEB Q5® site-directed mutagenesis kit. The gene of ppPabA* was optimized for expression in *E. coli* and synthesized (Life Technologies). A detailed description of all cloning procedures is available in the **Supplemental Experimental Procedures**.

Expression and purification of proteins

If not stated otherwise, all proteins were produced by gene expression in *E. coli* BL21-Gold (DE3) cells. ssTrpE was produced by gene expression in *E. coli* BL21-CodonPlus (DE3)-RIPL cells. Cells were grown in Luria broth medium at 20° C or 37° C over night. Proteins were purified from the soluble fraction of the cell extracts by Ni^{2+}-affinity and size exclusion chromatography. A detailed description is available in the **Supplemental Experimental Procedures**.

HPLC analysis of anthranilate and aminodeoxychorismate formation

TrpEx, TrpE, and PabB were assayed in 20 mM bicine buffer pH 8.5, 5 mM $MgCl_2$, 1 mM DTT, 200 mM NH_4Cl, and 500 µM CH. PabB-assays additionally contained 10 µM aminodeoxychorismate-lyase from *E. coli* (ecPabC) for conversion of the PabB product aminodeoxychorismate (ADC) to *p*-aminobenzoate (PABA). A detailed description of the HPLC setup is provided in the **Supplemental Experimental Procedures**.

Analysis of complex formation between different synthases and glutaminases

The ability of the different synthases and glutaminases to form heteromeric complexes was examined by a combination of size exclusion chromatography and static light scattering (SEC-SLS). The experimental setup of SEC comprised Superdex 75 10/300 GL or S200 10/300 GL columns (GE Healthcare) operated on an ÄKTAmicro system (GE Healthcare) connected to an ALIAS autosampler (Spark Holland). For SLS a Viscotek TDA 305 detector array (Malvern), including right-angle light scattering and refractive index detectors was used. The system was operated at 25 °C with a flow-rate of 0.5 mL/min of degassed buffer (50 mM Tris-HCl, pH 7.5, 150 mM KCl, 5 mM $MgCl_2$) and was calibrated with Ribonuclease A and rabbit muscle aldolase from the GE Healthcare SEC Low-Molecular-Weight calibration kit for experiments with S75 and S200 columns, respectively. Individual synthases and glutaminases were assayed at a concentration of 50 µM (applied volume 100 µL). For analysis of complex formation, synthases and glutaminases were equimolarilly mixed to a final concentration of 50 µM. Data were analyzed with the OmniSec software (version 4.7, Malvern).

Several synthases and glutaminases as well as combinations of them were additionally analyzed by native mass spectrometry. Proteins were exchanged into 100 mM ammonium acetate using Micro Bio-Spin columns (Bio-Rad). Single proteins and mixtures of synthases and glutaminases were analyzed on a Synapt G2 quadrupole ion mobility time-of-flight mass spectrometer (Waters), equipped with a nano-electrospray ionization (nESI) source. Nanospray borosilicate capillaries were prepared in-house and filled with 5 µL of a mixture of 20 µM synthase and 20 µM glutaminase (30 µM ppPabA*). Analytes were sprayed by applying a capillary voltage of 1-1.2 kV. The following instrument settings were used: 30 V sample cone voltage, 10 V trap collision voltage, and 5 mbar backing pressure. The source was kept at room temperature. Data were analyzed with MassLynx 4.1 (Waters) and spectra were assigned based on m/z values with separation of overlapping charge or oligomeric states by ion mobility.

Steady-state enzyme kinetics

The TrpEx/TrpE reaction was measured at 25 °C in a fluorimetric assay monitoring AA fluorescence (excitation 313 nm, emission 390 nm). A standard glutamine-dependent assay contained 100 mM potassium phosphate buffer, pH 7.0, 5 mM $MgCl_{2,}$ 1 mM DTT, 20 mM glutamine, and 90, 100, or 110 µM CH, respectively. After preincubation, 0.1 µM TrpEx or TrpE and 0.3 µM glutaminase were added. Entire progress curves were recorded for the three different substrate concentrations. For ammonia-dependent assays glutamine was substituted with 200 mM ammonium chloride and 100 mM bicine buffer, pH 8.5 was used. The PabB reaction was measured in an analogous fluorimetric assay at 25 °C monitoring PABA fluorescence (excitation 320 nm, emission 350 nm). The PabB product ADC was converted *in situ* to PABA by PabC. A standard assay contained 100 mM potassium phosphate buffer, pH 7.0, 5 mM $MgCl_{2,}$ 1 mM DTT, 20 mM glutamine, 5 µM PabC, and 90, 100, or 110 µM CH, respectively. After preincubation, 0.5 µM PabB and 1.5 µM glutaminase were added. Entire progress curves were recorded for the three different substrate concentrations. The k_{cat} and K_M^{CH} values were determined by fitting the progress curves with the Michaelis-Menten equation included in COSY (Eberhard, 1990).

The glutaminase activity was measured spectrophotometrically in a coupled enzymatic assay. Glutamate formed by the glutaminases was converted to α-ketoglutarate by glutamate-dehydrogenase (GDH) with simultaneous reduction of NAD^+ to NADH (**Figure S7A**). A standard assay contained 50 mM tricine-KOH buffer, pH 8.0, 5 mM $MgCl_2$, 1 mM DTT, 10 mM NAD^+, 1 mg/mL GDH, and 4 mM glutamine. Following preincubation, 1 µM glutaminase was added and the reaction was monitored at 340 nm and 25 °C for at least 15 minutes. After making sure that the progress curve proceeded with a constant slope, 3 µM synthase were added and the reaction was again monitored for at least 15 minutes. The slopes of the linear parts of the progress curves before and after the addition of the synthase were used to calculate a stimulation factor that describes the increase of the apparent turnover rate upon the addition of the synthase (**Figure S7B**). As TrpG glutaminases were not active without any synthase, no stimulation factor could be calculated. Therefore, only the apparent turnover rates after the addition of the synthases were used for further evaluation.

Bacillus subtilis growth experiments

The bs*pab*A, bs*pab*B, bs*trp*E, st*trp*Ex, and sm*trp*Ex genes were cloned into the pDG148 vector following the ligation-independent cloning protocol from (Joseph et al., 2001) using standard PCR reaction conditions and the oligonucleotides listed in the **Supplemental Experimental Procedures**. Electro-competent *B. subtilis* SB 491 cells were generated and transformed with plasmid DNA as described in the **Supplemental Experimental Procedures**. *In vivo* competition assays with pDG148_bs*pab*A, pDG148_bs*pab*B, pDG148_bs*trp*E, pDG148_st*trp*Ex, and pDG148_sm*trp*Ex were performed on Spizizen's minimal medium agar plates. For a detailed description please see the **Supplemental Experimental Procedures**.

AUTHOR CONTRIBUTIONS

Conceptualization, M.G.P., R.M., R.S.; Methodology, M.G.P., F.S., R.S., and R.M.; Software, M.G.P., M.B., and L.H.; Investigation, M.G.P., F.S., F.B.; Writing – Original Draft, M.G.P., F.S.; Writing – Review & Editing, M.G.P, R.S., and R.M.; Visualization, M.G.P.; Supervision, V.H.W., R.M., and R.S.; Funding Acquisition, M.G.P., F.S., V.H.W., R.M., and R.S.

ACKNOWLEDGEMENTS

M.G.P. and F.S. were supported by PhD fellowships from the *Fonds der Chemischen Industrie* and the *Konrad-Adenauer-Stiftung*, respectively. F.B. was supported by a National Institutes of Health grant to V.H.W. (R01 GM113658). We thank Alexandra Holinski for providing the pMAL-c5X expression vector, and Christiane Endres, Sonja Fuchs, Sabine Laberer, and Jeannette Ueckert for expert technical assistance.

REFERENCES

Aakre, C.D., Herrou, J., Phung, T.N., Perchuk, B.S., Crosson, S., and Laub, M.T. (2015). Evolving new protein-protein interaction specificity through promiscuous intermediates. *Cell 163*, 594-606.

Ahnert, S.E., Marsh, J.A., Hernandez, H., Robinson, C.V., and Teichmann, S.A. (2015). Principles of assembly reveal a periodic table of protein complexes. *Science 350*, aaa2245.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. *J. Mol. Biol. 215*, 403-410.

Ames, R.M., Talavera, D., Williams, S.G., Robertson, D.L., and Lovell, S.C. (2016). Binding interface change and cryptic variation in the evolution of protein-protein interactions. *BMC Evol. Biol. 16*, 40.

Atkinson, H.J., Morris, J.H., Ferrin, T.E., and Babbitt, P.C. (2009). Using sequence similarity networks for visualization of relationships across diverse protein superfamilies. *PLoS One 4*, e4345.

Babitzke, P. (1997). Regulation of tryptophan biosynthesis: Trp-ing the TRAP or how *Bacillus subtilis* reinvented the wheel. *Mol. Microbiol*. *26*, 1-9.

Bardwell, V.J., and Treisman, R. (1994). The POZ domain: A conserved protein-protein interaction motif. *Genes Dev*. *8*, 1664-1677.

Barlow, D.J., and Thornton, J.M. (1983). Ion-pairs in proteins. *J. Mol. Biol*. *168*, 867-885.

Battistuzzi, F.U., and Hedges, S.B. (2009). Eubacteria. In *The timetree of life*, S.B. Hedges, and S. Kumar, eds. (Oxford: Oxford University Press), pp. 106-115.

Beismann-Driemeyer, S., and Sterner, R. (2001). Imidazole glycerol phosphate synthase from *Thermotoga maritima*. Quaternary structure, steady-state kinetics, and reaction mechanism of the bienzyme complex. *J. Biol. Chem*. *276*, 20387-20396.

Bera, A.K., Chen, S., Smith, J.L., and Zalkin, H. (1999). Interdomain signaling in glutamine phosphoribosylpyrophosphate amidotransferase. *J. Biol. Chem*. *274*, 36498-36504.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The Protein Data Bank. *Nucleic Acids Res*. *28*, 235-242.

Bogan, A.A., and Thorn, K.S. (1998). Anatomy of hot spots in protein interfaces. *J. Mol. Biol*. *280*, 1-9.

Bouvier, B., Grunberg, R., Nilges, M., and Cazals, F. (2009). Shelling the Voronoi interface of protein-protein complexes reveals patterns of residue conservation, dynamics, and composition. *Proteins 76*, 677-692.

Capra, E.J., Perchuk, B.S., Skerker, J.M., and Laub, M.T. (2012). Adaptive mutations that prevent crosstalk enable the expansion of paralogous signaling protein families. *Cell 150*, 222-232.

Chothia, C. (1992). Proteins. One thousand families for the molecular biologist. *Nature 357*, 543-544.

Crooks, G.E., Hon, G., Chandonia, J.M., and Brenner, S.E. (2004). WebLogo: a sequence logo generator. *Genome Res*. *14*, 1188-1190.

DePristo, M.A., Weinreich, D.M., and Hartl, D.L. (2005). Missense meanderings in sequence space: A biophysical view of protein evolution. *Nat Rev Genet 6*, 678-687.

Dosselaere, F., and Vanderleyden, J. (2001). A metabolic node in action: chorismate-utilizing enzymes in microorganisms. *Crit. Rev. Microbiol*. *27*, 75-131.

Eberhard, M. (1990). A set of programs for analysis of kinetic and equilibrium data. *Comput. Appl. Biosci*. *6*, 213-221.

Farrow, J.M., 3rd, and Pesci, E.C. (2007). Two distinct pathways supply anthranilate as a precursor of the *Pseudomonas quinolone* signal. *J. Bacteriol*. *189*, 3425-3433.

Filippakopoulos, P., and Knapp, S. (2012). The bromodomain interaction module. *FEBS Lett*. *586*, 2692-2704.

Gao, M., and Skolnick, J. (2010). Structural space of protein-protein interfaces is degenerate, close to complete, and highly connected. *Proc. Natl. Acad. Sci. USA 107*, 22517-22522.

Garma, L., Mukherjee, S., Mitra, P., and Zhang, Y. (2012). How many protein-protein interactions types exist in nature? *PLoS One 7*, e38913.

Gerlt, J.A., Bouvier, J.T., Davidson, D.B., Imker, H.J., Sadkhin, B., Slater, D.R., and Whalen, K.L. (2015). Enzyme Function Initiative-Enzyme Similarity Tool (EFI-EST): A web tool for generating protein sequence similarity networks. *Biochim. Biophys. Acta 1854*, 1019-1037.

Gollnick, P., Babitzke, P., Antson, A., and Yanofsky, C. (2005). Complexity in Regulation of Tryptophan Biosynthesis in *Bacillus subtilis*. *Annu. Rev. Genet*. *39*, 47-68.

Goto, Y., Zalkin, H., Keim, P.S., and Heinrikson, R.L. (1976). Properties of anthranilate synthetase component II from *Pseudomonas putida*. *J. Biol. Chem*. *251*, 941-949.

Guharoy, M., and Chakrabarti, P. (2005). Conservation and relative importance of residues across protein-protein interfaces. *Proc. Natl. Acad. Sci. USA 102*, 15447-15452.

Harms, M.J., and Thornton, J.W. (2013). Evolutionary biochemistry: revealing the historical and physical causes of protein properties. *Nat Rev Genet 14*, 559-571.

Hashimoto, K., and Panchenko, A.R. (2010). Mechanisms of protein oligomerization, the critical role of insertions and deletions in maintaining different oligomeric states. *Proc. Natl. Acad. Sci. USA 107*, 20352-20357.

Huang, M., and Gibson, F. (1970). Biosynthesis of 4-aminobenzoate in *Escherichia coli*. *J. Bacteriol*. *102*, 767-773.

Hughes, A.L. (1994). The evolution of functionally novel proteins after gene duplication. *Proceedings. Biological sciences 256*, 119-124.

Hunter, S., Jones, P., Mitchell, A., Apweiler, R., Attwood, T.K., Bateman, A., Bernard, T., Binns, D., Bork, P., Burge, S., *et al.* (2012). InterPro in 2011: New developments in the family and domain prediction database. *Nucleic Acids Res*. *40*, D306-312.

Joseph, P., Fantino, J.R., Herbaud, M.L., and Denizot, F. (2001). Rapid orientated cloning in a shuttle vector allowing modulated gene expression in *Bacillus subtilis*. *FEMS Microbiol. Lett*. *205*, 91-97.

Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol*. *30*, 772-780.

Kelley, R.L., and Yanofsky, C. (1985). Mutational studies with the *trp* repressor of *Escherichia coli* support the helix-turn-helix model of repressor recognition of operator DNA. *Proc. Natl. Acad. Sci. USA 82*, 483-487.

Krieger, E., Joo, K., Lee, J., Raman, S., Thompson, J., Tyka, M., Baker, D., and Karplus, K. (2009). Improving physical realism, stereochemistry, and side-chain accuracy in homology modeling: Four approaches that performed well in CASP8. *Proteins 77 Suppl 9*, 114-122.

Krissinel, E., and Henrick, K. (2007). Inference of macromolecular assemblies from crystalline state. *J. Mol. Biol*. *372*, 774-797.

LaPorte, D.C. (1993). The isocitrate dehydrogenase phosphorylation cycle: regulation and enzymology. *J. Cell. Biochem*. *51*, 14-18.

LaPorte, D.C., and Koshland, D.E., Jr. (1982). A protein with kinase and phosphatase activities involved in regulation of tricarboxylic acid cycle. *Nature 300*, 458-460.

Letunic, I., and Bork, P. (2007). Interactive Tree Of Life (iTOL): An online tool for phylogenetic tree display and annotation. *Bioinformatics 23*, 127-128.

Levin, K.B., Dym, O., Albeck, S., Magdassi, S., Keeble, A.H., Kleanthous, C., and Tawfik, D.S. (2009). Following evolutionary paths to protein-protein interactions with high affinity and selectivity. *Nat. Struct. Mol. Biol*. *16*, 1049-1055.

Li, J., Mahajan, A., and Tsai, M.D. (2006). Ankyrin repeat: A unique motif mediating protein-protein interactions. *Biochemistry 45*, 15168-15178.

List, F., Bocola, M., Haeger, M.C., and Sterner, R. (2012). Constitutively active glutaminase variants provide insights into the activation mechanism of anthranilate synthase. *Biochemistry 51*, 2812-2818.

Loiseau, L., Ollagnier-de Choudens, S., Lascoux, D., Forest, E., Fontecave, M., and Barras, F. (2005). Analysis of the heteromeric CsdA-CsdE cysteine desulfurase, assisting Fe-S cluster biogenesis in *Escherichia coli*. *J. Biol. Chem*. *280*, 26760-26769.

Lynch, M., and Conery, J.S. (2000). The evolutionary fate and consequences of duplicate genes. *Science 290*, 1151-1155.

Mihara, H., and Esaki, N. (2002). Bacterial cysteine desulfurases: their function and mechanisms. *Appl. Microbiol. Biotechnol*. *60*, 12-23.

Miles, B.W., Banzon, J.A., and Raushel, F.M. (1998). Regulatory control of the amidotransferase domain of carbamoyl phosphate synthetase. *Biochemistry 37*, 16773-16779.

Mintseris, J., and Weng, Z. (2005). Structure, function, and evolution of transient and obligate protein-protein interactions. *Proc. Natl. Acad. Sci. USA 102*, 10930-10935.

Morollo, A.A., and Eck, M.J. (2001). Structure of the cooperative allosteric anthranilate synthase from *Salmonella typhimurium*. *Nat. Struct. Biol*. *8*, 243-247.

Ofran, Y., and Rost, B. (2003a). Analysing six types of protein-protein interfaces. *J. Mol. Biol*. *325*, 377-387.

Ofran, Y., and Rost, B. (2003b). Predicted protein-protein interaction sites from local sequence information. *FEBS Lett*. *544*, 236-239.

Oppenheim, D.S., and Yanofsky, C. (1980). Translational coupling during expression of the tryptophan operon of *Escherichia coli*. *Genetics 95*, 785-795.

Pascarella, S., and Argos, P. (1992). Analysis of insertions/deletions in protein structures. *J. Mol. Biol*. *224*, 461-471.

Pazos, F., and Valencia, A. (2008). Protein co-evolution, co-adaptation and interactions. *The EMBO journal 27*, 2648-2655.

Pires, D.E., Ascher, D.B., and Blundell, T.L. (2014). mCSM: predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics 30*, 335-342.

Raushel, F.M., Thoden, J.B., and Holden, H.M. (1999). The amidotransferase family of enzymes: molecular machines for the production and delivery of ammonia. *Biochemistry 38*, 7891-7899.

Raushel, F.M., Thoden, J.B., and Holden, H.M. (2003). Enzymes with molecular tunnels. *Acc. Chem. Res. 36*, 539-548.

Roesser, J.R., and Yanofsky, C. (1991). The effects of leader peptide sequence and length on attenuation control of the *trp* operon of *E. coli*. *Nucleic Acids Res. 19*, 795-800.

Roux, B., and Walsh, C.T. (1992). p-aminobenzoate synthesis in *Escherichia coli:* kinetic and mechanistic characterization of the amidotransferase PabA. *Biochemistry 31*, 6904-6910.

Schreiber, G., and Keating, A.E. (2011). Protein binding specificity versus promiscuity. *Curr. Opin. Struct. Biol*. *21*, 50-61.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: A software environment for integrated models of biomolecular interaction networks*. Genome Res*. *13*, 2498-2504.

Sintchak, M.D., Arjara, G., Kellogg, B.A., Stubbe, J., and Drennan, C.L. (2002). The crystal structure of class II ribonucleotide reductase reveals how an allosterically regulated monomer mimics a dimer. *Nat. Struct. Biol*. *9*, 293-300.

Slock, J., Stahly, D.P., Han, C.Y., Six, E.W., and Crawford, I.P. (1990). An apparent *Bacillus subtilis* folic acid biosynthetic operon containing *pab*, an amphibolic *trp*G gene, a third gene required for synthesis of *para*-aminobenzoic acid, and the dihydropteroate synthase gene. *J. Bacteriol*. *172*, 7211-7226.

Smoot, M.E., Ono, K., Ruscheinski, J., Wang, P.L., and Ideker, T. (2011). Cytoscape 2.8: New features for data integration and network visualization. *Bioinformatics 27*, 431-432.

Söding, J. (2005). Protein homology detection by HMM-HMM comparison. *Bioinformatics 21*, 951-960.

Spraggon, G., Kim, C., Nguyen-Huu, X., Yee, M.C., Yanofsky, C., and Mills, S.E. (2001). The structures of anthranilate synthase of *Serratia marcescens* crystallized in the presence of (*i*) its substrates, chorismate and glutamine, and a product, glutamate, and (*ii*) its end-product inhibitor, $_{\text{L}}$-tryptophan. *Proc. Natl. Acad. Sci. USA 98*, 6021-6026.

Strohmeier, M., Raschle, T., Mazurkiewicz, J., Rippe, K., Sinning, I., Fitzpatrick, T.B., and Tews, I. (2006). Structure of a bacterial pyridoxal 5'-phosphate synthase complex. *Proc. Natl. Acad. Sci. USA 103*, 19284-19289.

Thorn, K.S., and Bogan, A.A. (2001). ASEdb: A database of alanine mutations and their effects on the free energy of binding in protein interactions. *Bioinformatics 17*, 284-285.

Webb, B., and Sali, A. (2014). Protein structure modeling with MODELLER. *Methods Mol Biol 1137*, 1-15.

White, R.H. (1988). Analysis and characterization of the folates in the nonmethanogenic archaebacteria. *J. Bacteriol. 170*, 4608-4612.

Worrell, V.E., and Nagle, D.P., Jr. (1988). Folic acid and pteroylpolyglutamate contents of archaebacteria. *J. Bacteriol*. *170*, 4420-4423.

Xie, G., Bonner, C.A., Brettin, T., Gottardo, R., Keyhani, N.O., and Jensen, R.A. (2003). Lateral gene transfer and ancient paralogy of operons containing redundant copies of tryptophan-pathway genes in *Xylella* species and in heterocystous cyanobacteria*. Genome Biol 4*, R14.

Xu, D., Tsai, C.J., and Nussinov, R. (1997). Hydrogen bonds and salt bridges across protein-protein interfaces. *Prot. Eng*. *10*, 999-1012.

Yakhnin, H., Yakhnin, A.V., and Babitzke, P. (2007). Translation control of *trp*G from transcripts originating from the folate operon promoter of *Bacillus subtilis* is influenced by translation-mediated displacement of bound TRAP, while translation control of transcripts originating from a newly identified *trp*G promoter is not. *J. Bacteriol*. *189*, 872-879.

Yanofsky, C., Kelley, R.L., and Horn, V. (1984). Repression is relieved before attenuation in the *trp* operon of *Escherichia coli* as tryptophan starvation becomes increasingly severe. *J. Bacteriol*. *158*, 1018-1024.

Yoon, B.J. (2009). Hidden Markov Models and their Applications in Biological Sequence Analysis. *Curr. Genomics 10*, 402-415.

Yu, H., Braun, P., Yildirim, M.A., Lemmens, I., Venkatesan, K., Sahalie, J., Hirozane-Kishikawa, T., Gebreab, F., Li, N., Simonis, N*., et al*. (2008). High-quality binary protein interaction map of the yeast interactome network. *Science 322*, 104-110.

Zalkin, H. (1973). Anthranilate synthetase. *Adv. Enzymol. Relat. Areas Mol. Biol*. *38*, 1-39.

Zeigler, D.R., Pragai, Z., Rodriguez, S., Chevreux, B., Muffler, A., Albert, T., Bai, R., Wyss, M., and Perkins, J.B. (2008). The origins of 168, W23, and other *Bacillus subtilis* legacy strains. *J. Bacteriol*. *190*, 6983-6995.

Zhang, Y. (2008). I-TASSER server for protein 3D structure prediction. BMC *Bioinformatics 9*, 40.

Zhang, Y., Hubner, I.A., Arakaki, A.K., Shakhnovich, E., and Skolnick, J. (2006). On the origin and highly likely completeness of single-domain protein structures. *Proc. Natl. Acad. Sci. USA 103*, 2605-2610.

Supporting Information for Publication C

Evolutionary diversification of protein-protein interactions by interface add-ons

Maximilian G. Plach, Florian Semmelmann, Florian Busch, Markus Busch, Leonhard Heizinger, Vicky H. Wysocki, Rainer Merkl, and Reinhard Sterner (2017).

Submitted for Publication

Supplemental Information

Evolutionary diversification of protein-protein interactions by interface add-ons

Maximilian G. Plach[1], Florian Semmelmann[1], Florian Busch[2], Markus Busch[1], Leonhard Heizinger[1], Vicki H. Wysocki[2], Rainer Merkl[1*], and Reinhard Sterner[1*]

[1]Institute of Biophysics and Physical Biochemistry, University of Regensburg, D-93040 Regensburg, Germany
[2]Department of Chemistry and Biochemistry, The Ohio State University, 460 West 12th Avenue, OH-43210 Columbus, USA
*Correspondence: Rainer Merkl: +49-941-3086; Rainer.Merkl@ur.de, Reinhard Sterner: +49-941 943 3015; Reinhard.Sterner@ur.de

submitted for publication

Supplemental Experimental Procedures

Materials

Glutamate dehydrogenase was purchased from Roche. Chorismate was purchased from Sigma Aldrich as the barium salt and barium ions were precipitated by the addition of a slight excess of sodium sulfate. All other chemical reagents were purchased from Sigma Aldrich in analytical or HPLC grade and used without further purification.

Computation of sequence similarity networks

The SSN of the InterPro entry IPR019999 was computed according to (Atkinson et al., 2009; Gerlt et al., 2015). To exclude sequence fragments and multi-domain proteins, only sequences between 320 and 620 amino-acids in length were chosen for the initial all-by-all `BLAST`. An E-value cut-off of 1E-77 was applied and the complexity of the resulting network was reduced by computing a representative network in which sequences with >75% identity were grouped into single nodes. Networks were visualized with the `organic y-files` layout in `Cytoscape 3.3` (Shannon et al., 2003; Smoot et al., 2011). Similarly, we computed a SSN for the InterPro entry IPR015890, which is a "domain" entry that contains the sequences from IPR019999 as well as additional sequences belonging to the "chorismate binding domain" fold. For this SSN we chose an E-value cut-off of 1E-80. Both SSNs showed the same aggregation of TrpE sequences in one large cluster with a noticeable separation of the TrpEx sequences to a distinct subcluster. At more stringent E-values, the TrpEx subcluster becomes separated from the TrpE cluster in both SSNs.

Genetic profiling of Archaea and Bacteria to determine the phylogenetic distribution of AS and ADCS complexes

BLAST-scans of TrpEx and TrpE species

In the following, we name the content of the TrpEx cluster $TrpEx_{SSN}$ and that of the TrpE cluster $TrpE_{SSN}$.

The species that comprise $TrpEx_{SSN}$ and $TrpE_{SSN}$ were extracted from the SSN of IPR015890 by their NCBI taxonomy identifiers (TaxIDs). After eliminating duplicate identifiers, 4297 and 11561 unique TaxIDs remained. The two datasets of TaxIDs are hereafter referred to as $TrpEx_{SSN}^{TaxID}$ and $TrpE_{SSN}^{TaxID}$. To scan each species in $TrpEx_{SSN}^{TaxID}$ and $TrpE_{SSN}^{TaxID}$ individually for the presence of TrpEx, TrpE, PabB, TrpG, and PabA, it was necessary to limit the BLAST-search space to the proteome of the individual species. To circumvent limitations of the command line version of blastp 2.2.30+ each search was restricted to a species-specific (i. e. TaxID-specific) subset of proteins from the *nr* database. For the BLAST-searches with default parameters the sequences of PabB and PabA from *E. coli* were used as queries for the synthases and glutaminases, respectively (see table below). Hits with E-values lower than 1E-20 were further processed and their full sequences were added to the two sets of candidate sequences $TrpEx_{candidates}^{TaxID}$ and $TrpE_{candidates}^{TaxID}$.

Query sequences used for BLAST searches.

ecPabA	MILLIDNYDSFTWNLYQYFCELGADVLVKRNDALTLADIDALKPQKIVISPGPCTPDEAG ISLDVIRHYAGRLPILGVCLGHQAMAQAFGGKVVRAAKVMHGKTSPITHNGEGVFRGLAN PLTVTRYHSLVVEPDSLPACFDVTAWSETREIMGIRHRQWDLEGVQFHPESILSEQGHQL LANFLHR
ecPabB	MKTLSPAVITLLWRQDAAEFYFSRLSHLPWAMLLHSGYADHPYSRFDIVVAEPICTLTTF GKETVVSESEKRTTTTDDPLQVLQQVLDRADIRPTHNEDLPFQGGALGLFGYDLGRRFES LPEIAEQDIVLPDMAVGIYDWALIVDHQRHTVSLLSHNDVNARRAWLESQQFSPQEDFTL TSDWQSNMTREQYGEKFRQVQEYLHSGDCYQVNLAQRFHATYSGDEWQAFLQLNQANRAP FSAFLRLEQGAILSLSPERFILCDNSEIQTRPIKGTLPRLPDPQEDSKQAVKLANSAKDR AENLMIVDLMRNDIGRVAVAGSVKVPELFVVEPFPAVHHLVSTITAQLPEQLHASDLLRA AFPGGSITGAPKVRAMEIIDELEPQRRNAWCGSIGYLSFCGNMDTSITIRTLTAINGQIF CSAGGGIVADSQEEAEYQETFDKVNRILKQLEK

Generation of representative HMMs

To annotate the sequences in $TrpEx_{candidates}^{TaxID}$ and $TrpE_{candidates}^{TaxID}$ as either one of the three eligible synthases, or as one of the two eligible glutaminases, each entry was compared to Hidden-Markov-Models representing the five enzymes. To parametrize the HMMs, sequence sets for TrpEx, TrpE, PabB, TrpG, and PabA were compiled and MSAs were generated with MAFFT in L-INS-i mode (Katoh and Standley, 2013). MSA_{TrpEx} contained sequences taken from $TrpEx_{SNN}$. Sequences for MSA_{TrpE} and MSA_{PabB} were taken from a previous study (Plach et al., 2015). Sequences for MSA_{TrpG} and MSA_{PabA} were retrieved from *nr* by means of PSI-BLAST searches with TrpG from *Serratia marcescens* and PabA from *E. coli* as queries, respectively. For HMM generation by means of hhmake (Söding, 2005) the MSAs were made non-redundant at 90% sequence identity and columns with more than 50% gaps were ignored. All other parameters of hhmake were kept default. $\mathrm{HMM}_{\mathrm{TrpEx}}$, $\mathrm{HMM}_{\mathrm{TrpE}}$, $\mathrm{HMM}_{\mathrm{PabB}}$, $\mathrm{HMM}_{\mathrm{TrpG}}$, and $\mathrm{HMM}_{\mathrm{PabA}}$ were based on 101, 107, 103, 28, and 133 sequences, respectively. All five HMM_{enz} were subjected to four-fold cross-validation to confirm their specificity (**Figure S12**). To this end, 25% of the sequences in an MSA were randomly selected and served as a test set. The remaining 75% served as a training set to generate HMM^{*}_{enz}. Each test set was compared to the corresponding HMM^{*}_{enz} and the other four unaltered HMM_{enz}. For example, 25% of PabA-sequences from MSA_{PabA} were chosen as a test set and the remaining 75% used to parametrize $\mathrm{HMM}^{*}_{\mathrm{PabA}}$. The PabA test-set was then compared to $\mathrm{HMM}_{\mathrm{PabA}}$, $\mathrm{HMM}_{\mathrm{TrpG}}$, $\mathrm{HMM}_{\mathrm{PabB}}$, $\mathrm{HMM}_{\mathrm{TrpE}}$, and $\mathrm{HMM}_{\mathrm{TrpEx}}$. This process was repeated four times. Each HMM^{*}_{enz} identified the corresponding test-set with E-values between 40 and 60 orders of magnitude lower than sequences belonging to the other four groups. These findings make clear that the five HMMs are well-suited to classify TrpEx, TrpE, PabB, TrpG, and PabA sequences.

Sequence-to-HMM comparisons

Using `hhsearch` (Söding, 2005) and the five HMM_{enz}, sequences from $TrpEx^{TaxID}_{candidates}$ and $TrpE^{TaxID}_{candidates}$ were assigned to the group giving the lowest E-value. Glutaminases were compared with $\mathrm{HMM}_{\mathrm{TrpG}}$ and $\mathrm{HMM}_{\mathrm{PabA}}$. Accordingly, synthases were compared with $\mathrm{HMM}_{\mathrm{TrpEx}}$, $\mathrm{HMM}_{\mathrm{TrpE}}$, and $\mathrm{HMM}_{\mathrm{PabB}}$. The assignments were stored in the sets $TrpEx_{predictions}$ and $TrpE_{predictions}$. The selectivity of the assignments was examined by calculating log-odds ratios S_{kl}, which relate the E-values of HMM_{k} and HMM_{l}:

$$S_{kl} = \left| \log_{10}\left(\frac{\text{E-val}(\mathrm{HMM}_{k})}{\text{E-val}(\mathrm{HMM}_{l})} \right) \right|$$

For the classification of PabA homologs, $\mathrm{HMM}_{k} = \mathrm{HMM}_{\mathrm{PabA}}$ and $\mathrm{HMM}_{l} = \mathrm{HMM}_{\mathrm{TrpG}}$ were used. For the classification of PabB homologs, $\mathrm{HMM}_{k} = \mathrm{HMM}_{\mathrm{PabB}}$ was used in all cases; for the set $TrpEx_{predictions}$ $\mathrm{HMM}_{l} = \mathrm{HMM}_{\mathrm{TrpEx}}$ and for the set $TrpE_{predictions}$ $\mathrm{HMM}_{l} = \mathrm{HMM}_{\mathrm{TrpE}}$ was used. The refinement of the HMMs based on an iterative sequence selection process similar to the `PSI-BLAST` approach did not improve their specificity.

Creating representative datasets $TrpEx_{repr}$ and $TrpE_{repr}$

Elements in $TrpEx^{TaxID}_{candidates}$ and $TrpE^{TaxID}_{candidates}$ were further processed to assess the phylogenetic distribution of the corresponding enzymes. At first, all entries were removed that were mapped to sequences with score $S_{kl} < 10$ in order to exclude species possessing sequences that could not be assigned unambiguously. Subsequently, for species which are over-represented in *nr* due to many strain-level entries, only one entry was chosen. All species with incomplete TaxIDs were removed and for all entries with an identical common name determined via Key2Ann (Pürzer et al., 2011) only one entry was chosen. The final datasets $TrpEx_{repr}$ and $TrpE_{repr}$ contained 1463 and 4386 entries, respectively.

Determining the phylogenetic distribution of synthase-glutaminase co-occurrences

To determine the phylogenetic distribution of the co-occurrences a tree-of-life (TOL) based on 31 concatenated orthologs (Ciccarelli et al., 2006) was used in combination with `iTOL 3.0` (Letunic and Bork, 2007). For cases where $TrpEx_{repr}$ and $TrpE_{repr}$ contain several elements for a species of the TOL (e.g. several *E. coli* strains) a consensus approach was used. The final TOL comprised 120 species from all major bacterial and archaeal phyla with mapped co-occurrences of TrpEx, TrpE, PabB, TrpG, and PabA.

Computation of interface conservation

The similarity of the interface regions between TrpEx and PabB was determined as follows: We first computed a structure-based sequence alignment of smTrpEx (PDB ID 1i7q) and ecPabB (1k0e) using `STRAP` (Gille and Frommel, 2001) with the TM-align algorithm (Zhang and Skolnick, 2005). This alignment was then supplemented with TrpEx and PabB sequences from $TrpEx_{repr}$ and $TrpE_{repr}$ by means of `MAFFT` (Katoh and Standley, 2013) with the "`keeplength`" option. The MSA was made non-redundant at 90% identity and was finally mapped onto the crystal structure of smTrpEx using CONSURF (Ashkenazy et al., 2010).

Cloning and mutagenesis

The genes of TrpEx, TrpE, PabB, TrpG, and PabA proteins were amplified from genomic DNA or whole cell lysate in standard PCR reactions using the combinations of oligonucleotides specified in the table below. If not stated otherwise, all genes were cloned into the pET21a vector (Stratagene, providing a C-terminal hexahistidine-tag) via the introduced restriction sites for *NdeI* and *XhoI*. St*trp*G and st*trp*E were already available cloned into the pET21a vector (Plach et al., 2015). Ec*pab*C was cloned into the pET28a vector (Stratagene, providing an N-terminal hexahistidine-tag). Ec*trp*E was initially amplified and cloned into the pET21a vector using *NheI* and *XhoI* restriction sites, because it contains an intragenic *NdeI* restriction site. This site was then deleted via a modified QuickChange mutagenesis protocol (Wang and Malcolm, 1999) introducing a silent mutation of a histidine codon (CAT → CAC). The resulting ec*trp*E* was then amplified with an *NdeI* forward primer and cloned into the pET21a vector via *NdeI* and *XhoI* restriction sites. bs*pab*B was cloned into a modified version of the pMAL-c5X vector (New England Biolabs). This vector is designed such that the gene of interest can be inserted downstream of the *mal*E gene (which encodes for the maltose-binding-protein, MBP) via restriction sites for *NdeI* and *XhoI.* The vector additionally features the coding sequence for a hexahistidine-tag upstream of *mal*E and a linker region between *mal*E and the gene of interest that contains a thrombin cleavage site. Gene expression from this vector results in the fusion of the protein of interest to the C-terminus of N-terminally hexahistidine-tagged MPB. The protein of interest can subsequently be cleaved off with thrombin.

The gene coding for stTrpEx_Δ was generated by deleting codons 111 to 116 inclusive in pET21a_st*trp*Ex and inserting the codons AGC and GGC coding for serine and glycine, respectively, using the NEB Q5® site-directed mutagenesis kit and the 5'-phosphorylated oligonucleotides listed in the table below. The gene coding for ppPabA* (see table below) was optimized for expression in *E. coli*, synthesized (LifeTechnologies), and cloned into the pET21a vector using the terminal restriction sites for *Nde*I and *Xho*I.

Template DNA and oligonucleotides used for cloning and site-directed mutagenesis. Restriction sites are underlined; mismatches for site-directed mutagenesis are in bold; 5'-phosphorylated oligonucleotides are marked with ℗; "n.a.": not applicable.

Gene	Organism	Template DNA	Primer (5′→3′)	
bs*pab*A	*Bacillus subtilis* (bs)	Whole cell lysate *Bacillus subtilis* strain 168	Fo_NdeI	CAGGGCATATGATTTTAATGATTGATAACTACG
			Re_XhoI	CCCTGCTCGAGCGCAATAACTTCCTTGCG
bs*pab*B			Fo_NdeI	CAGGGCATATGGCACAACGCAGACC
			Re_XhoI	CCCTGCTCGAGTCTAATTTTTGTCTCTTCTTCGC
bs*trp*E			Fo_NdeI	CAGGGCATATGAATTTCCAATCAAACATTTC
			Re_EcoRI	CCCTGGAATTCCTAGTGATGGTGATGATGATGACGCACAATTGTAGAAATC
ec*pab*A	*Escherichia coli* (ec)	Whole cell lysate *Escherichia coli* K12	Fo_NdeI	CAGGGCATATGATCCTGCTTATAGATAACTACG
			Re_XhoI	CCCTGCTCGAGGCGATGCAGGAAATTAGCCAGC
ec*pab*B			Fo_NdeI	CAGGGCATATGAAGACGTTATCTCCCGCTGTG
			Re_XhoI	CCCTGCTCGAGCTTCTCCAGTTGCTTCAGG
ec*pab*C			Fo_NdeI	CAGGGCATATGTCTTAATTAACGGTC
			Re_XhoI	CCCTGCTCGAGATTCGGGCGCTCACAAAG
ec*trp*GD			Fo_NdeI	CAGGGCATATGGCTGACATTCTGCTG
			Re_XhoI	CCCTGCTCGAGCCCTCGTGCCGCC
ec*trp*Ex			Fo_NheI	CAGGGGCTAGCATGCAAACACAAAAACCGACTCTCG
			Fo_NdeI	CAGGGCATATGCAAACACAAAAACCGACTCTCG
			Re_XhoI	CCCTGCTCGAGCGCAATAACTTCCTTGCG
			QCM_fo	GTTTCCGTGCCGCA**C**ATGCGTTGTGAATGTAATC
			QCM_re	GATTACATTCACAACGCATGT**G**CGGCACGGAAAC
pp*pab*A	*Pseudomonas putida* (pp)	Genomic DNA *Pseudomonas putida* F1 (DSMZ 6899)	Fo_NdeI	CAGGGCATATGTTACTGATGATCGACAATTACGACTC
			Re_XhoI	CCCTGCTCGAGACGGCGGCCGCC
pp*pab*B			Fo_NdeI	CAGGGCATATGCCGACCTGCACGCTAC
			Re_XhoI	CCCTGCTCGAGCAAGCCCTGCAGGGTCTG
pp*trp*E			Fo_NdeI	CAGGGCATATGAACCGCGAAGAATTCC
			Re_XhoI	CCCTGCTCGAGTCTGGCGGAAGTCTGC

sm*pab*A	*Serratia marcescens* (sm)	Genomic DNA *Serratia marcescens* subsp. *marcescens* (DSMZ 30121)	Fo_NdeI Re_XhoI	CAGGGCATATGCTGCTGCTGATCGATAAC CCCTGCTCGAGACGGTTGAGGAAGTTATCC
sm*pab*B			Fo_NdeI Re_XhoI	CAGGGCATATGAGCGTAACCGCCC CCCTGCTCGAGCGACAGGGCATACTCCC
sm*trp*Ex			Fo_NdeI Re_XhoI	CAGGGCATATGATGAACACCAAACCAC CCCTGCTCGAGGAACACCTCCTTGGC
ss*trp*E	*Sulfolobus solfataricus* (ss)	Genomic DNA *Sulfolobus solfataricus* P2 (DSMZ 1617)	Fo_NdeI Re_XhoI	CAGGGCATATGGAAGTTCATCCAATAAGTG CCCTGCTCGAGCCTCACCCCTATTGCTG
st*pab*B	*Salmonella typhimurium* (st)	Genomic DNA *Salmonella enterica* subsp. *enterica* serovar *typhimurium* LT2 (DSMZ 17058)	Fo_NdeI Re_XhoI	CAGGGCATATGATGAAGACGTTATCTCCC CCCTGCTCGAGGTTCTCCAGTGGGTGC
st*trp*G			Fo_NdeI Re_XhoI	GGAATTCCATATGGCTGATATTCTGCT AATCTCGAGCTTTTGCTGCGCCCAG
st*trp*Ex			Fo_NdeI Re_XhoI	GGAATTCCATATGCAAACACCAAAACCC AAACTCGAGGAAGGTCTCCTGT
sttrpEx_Δ	n.a.	pET21a_st*trp*Ex	Fo Re	ⓅGGCCGTTTATGCTCTCTGTCGGTATTTGATGC ⓅGCTTGGGCTGACGGGCGGGAAGC

Nucleotide and amino acid sequence of ppPabA*.
The nucleotide sequence was optimized for expression in *E. coli*. Restriction sites for *Nde*I (5′) and *Xho*I (3′) are underlined. The C-terminal hexahistidine-tag and two linker amino acids are in bold.

Nucleotide sequence
5′CATATGCTGCTGATGATCGACAACTATGATAGCTTTACCTATAACGTTGTTGATCAGCTGCGTGAACTGGG TGCAGAAGTTAAAGTTTATCGTAATCAAGAACTGACGATCGCACAGATTGAAGCACTGAATCCGGAACGTAT TGTTGTTAGTCCGGGTCCGTGTACCCCGAGCGAAGCCGGTGTTAGCATTGAAGCAATTCTGCATTTTGCAGG TAAACTGCCGATTCTGGGTGTTTGTCTGGGTCATCAGAGCATTGGTCAGGCATTTGGTGGTGATGTTGTTCG TGCACGTCAGGTTATGCATGGTAAAACCAGTCCGGTTTATCATCGTGATCTGGGTGTGTTTGCAAGCCTGAA TAATCCGCTGACCGTTACCCGTTATCATTCACTGGTTGTTAAACGTGAAACCCTGCCGGATTGTCTGGAAGT TACCGCATGGACCAGCCATGCAGATGGTAGCGTTGATGAAATTATGGGTCTGCGTCATAAAACCCTGAATAT TGAAGGTGTTCAGTTTCATCCGGAAAGCATTCTGACCGAACAGGGTCACGAACTGTTTGCAAATTTTCTGAA ACAGACCGGTGGTCGTCGCCTCGAG3′

Amino acid sequence
MLLMIDNYDSFTYNVVDQLRELGAEVKVYRNQELTIAQIEALNPERIVVSPGPCTPSEAGVSIEAILHFAGK LPILGVCLGHQSIGQAFGGDVVRARQVMHGKTSPVYHRDLGVFASLNNPLTVTRYHSLVVKRETLPDCLEVT AWTSHADGSVDEIMGLRHKTLNIEGVQFHPESILTEQGHELFANFLKQTGGRR**LEHHHHHH**

Expression and purification of proteins

If not stated otherwise, all proteins were produced by expression in *E. coli* BL21-Gold (DE3) cells (Agilent Technologies). SsTrpE was produced by expression of the ss*trp*E gene in *E. coli* BL21-CodonPlus (DE3)-RIPL cells (Agilent Technologies). For protein production, overnight cultures of individual clones were used to inoculate 4 L of Luria broth medium supplemented with 150 µg/mL ampicillin. Cells were grown at 37 °C to an OD_{600} of 0.6 and then cooled to 20 °C (except for ssTrpE). Expression was induced by adding 0.5 mM isopropyl β-D-1-thiogalactopyranoside and growth was continued overnight at 20 °C (37 °C for ssTrpE). Cells were harvested by centrifugation (2700 g, 4 °C), resuspended in 50 mM Tris-HCl, pH 7.5, 300 mM potassium chloride, 10 mM imidazole, and lysed by sonification. The insoluble fraction was removed by centrifugation (23000 g, 4 °C) and the soluble extracts were filtered through a 0.8 µm membrane.

Supernatants containing C-terminal or N-terminal hexahistidine-tagged proteins were loaded onto a HisTrapFF crude column (5 mL, GE Healthcare), which had been equilibrated with resuspension buffer, and eluted from the column by applying a linear gradient of 10 – 750 mM imidazole. Enzyme-containing fractions, as judged by SDS-PAGE, were pooled and further purified by preparative gel filtration (Superdex 75 HiLoad 26/60, 320 mL, GE Healthcare, 50 mM Tris-HCl, pH 7.5, 50 mM KCl, 5 mM MgCl2, 2 mM DTT, 4 °C). Elution fractions were analyzed by SDS-PAGE and the fractions containing pure protein were pooled. The enzymes were finally concentrated to 100-200 µM and flash frozen in liquid nitrogen. Protein concentrations were determined by measuring the absorbance at 280 nm, using the molar extinction coefficient calculated via ExPASy ProtParam (http://web.expasy.org/protparam/).

The MBP-bs*pab*B fusion protein was purified from the soluble cell extract by Ni^{2+}-affinity chromatography on a HisTrapFF crude column as described above. The pooled elution fractions were then dialyzed against 50 mM Tris-HCl pH 7.5, 300 mM KCl, 2.5 mM $CaCl_2$ for six hours at 4 °C. The MBP was cleaved off by digestion with thrombin (1 U/mL, 15 °C, overnight). BsPabB was separated from undigested fusion protein and MBP by Ni^{2+}-affinity chromatography. The flowthrough fractions (containing the untagged bsPabB) were analyzed by SDS-PAGE and appropriate fractions were pooled. BsPabB was further purified by preparative gel filtration, concentrated, and stored as described above. The final enzyme preparations contained 0.5 – 2 mg of >95% pure protein.

HPLC analysis of anthranilate and aminodeoxychorismate formation

The TrpEx, TrpE, and PabB synthases were assayed in 20 mM bicine buffer pH 8.5, 5 mM $MgCl_2$, 1 mM DTT, and 200 mM NH_4Cl. The enzyme and chorismate concentrations were 10 µM and 500 µM, respectively. PabB-assays additionally contained 10 µM aminodeoxychorismate-lyase from *E. coli* (ecPabC) for conversion of the PabB product aminodeoxychorismate (ADC) to *p*-aminobenzoate (PABA). The bsPabB assay additionally contained bsPabA, because without it, no formation of PABA could be observed. Enzymes were incubated with chorismate at 25 °C for three hours. Samples were ultrafiltrated (Amicon Ultra 0.5 mL centrifugal unit, 10 kDa molecular weight cut-off, 4 °C) and protein-free filtrates were analyzed on an Agilent 1200 HPLC system equipped with a 5 µm Agilent Zorbax Eclipse XDB C18 column (150 mm × 4.6 mm). Mobile phase A was 0.1% formic acid in H2O, mobile phase B 0.1% formic acid in acetonitrile. Separation was performed at 10 °C with a flow rate of 1 mL/min in the following manner: isocratic elution at 5% B from 0-5 minutes, linear gradient from 5-100% B from 6-20 minutes. Elution was monitored by absorbance at 280 nm and 320 nm and fluorescence emission at 400 nm following excitation at 310 nm. AA and PABA peaks were assigned based on a comparison of retention times with those of authentic chemical standards.

Bacillus subtilis growth experiments

The bs*pab*A, bs*pab*B, bs*trp*E, st*trp*Ex, and smt*rp*Ex genes were cloned into the pDG148 vector following the ligation-independent cloning protocol from (Joseph et al., 2001), using standard PCR reaction conditions and the plasmids and oligonucleotides listed in the table below. Electrocompetent cells of the prototrophic *B. subtilis* strain SB491 (DSM-No. 6397) were prepared as follows: a 50 mL over-night culture of *B. subtilis* SB491 in Luria broth (LB) medium was used to inoculate 1 L LB medium supplemented with 0.5 M sorbitol. Cells were grown at 37 °C to an OD_{600} of 1.0, harvested by centrifugation (4000 g, 4 °C) and washed four times with 250 mL poration medium (10% v/v glycerol, 9.1% w/v mannitol, 0.5 M sorbitol). The cells were finally resuspended in 5 mL poration medium and stored at -80 °C. Before electroporation, 100 µL frozen cells were thawed on ice and incubated with 100 ng of plasmid DNA for five minutes. Cell were electroporated at a voltage of 2500 V, immediately resuspended in 900 µL regeneration medium (LB medium with 7% w/v mannitol and 0.5 M sorbitol), cured at 37 °C for three hours and plated on LB agar plates containing 50 µg/mL kanamycin. Individual colonies were picked and the presence of the pDG148 plasmids containing the respective genes was checked in standard colony PCR reactions using the oligonucleotides listed in the table below. Positive colonies were used to inoculate 20 mL LB medium containing 50 µg/mL kanamycin. After growth over night cells were harvested and plasmid DNA was isolated and sequenced using the colony PCR primers to verify the presence of the correct plasmids and to exclude inadvertent mutations in the inserted genes. Correct clones were stored as glycerin stocks at -80 °C.

Growth experiments with pDG148, pDG148_bs*pab*A, pDG148_bs*pab*B, pDG148_bs*trp*E, pDG148_st*trp*Ex, and pDG148_smt*rp*Ex were performed on Spizizen's minimal medium (SMM) agar plates that were prepared as follows: 2 g $(NH_4)SO_4$, 14 g K_2HPO_4, 6 g KH_2PO_4, 1 g Na_3-citrate dihydrate, and 0.2 g $MgSO_4$ heptahydrate were dissolved in 700 mL water and autoclaved. After sterilization the solution was mixed with 300 mL of sterile 1.5% agar solution (USP grade, MP Biomedicals). Subsequently, L-arginine, L-proline, L-glutamic acid, and L-glutamine were added to a final concentration of 1 mg/mL each, glucose to a final concentration of 0.5% (w/v), and kanamycin to a final concentration of 50 µg/mL. All stock solutions were prepared separately and sterilized by ultrafiltration. If necessary, isopropyl β-D-1-thiogalactopyranoside (IPTG) was added to a final concentration of 2 mM and *para*-aminobenzoate and folic acid to a final concentration of 10 µg/mL (Wegkamp et al., 2010).

For the growth experiments an identical procedure was followed in all cases: Glycerin stocks of *B. subtilis* SB491 transformed with the different pDG148 plasmids were streaked out on LB agar plates and incubated over night at 37 °C. Single colonies were then used to inoculate 50 mL of LB medium containing 50 µg/mL kanamycin, which was incubated at 37 °C over night. This overnight culture was then used to inoculate 50 mL of LB medium containing 50 µg/mL kanamycin (to an optical density of 0.1 at 600 nm), which was incubated at 37 °C until an optical density of 1 at 600 nm was reached. Subsequently, 1 mL cell suspension was collected by centrifugation (4000 g, 4 °C) and washed three times with 1% sterile sodium chloride solution and finally dissolved in 1 mL of the sodium chloride solution to adjust the optical density to 1 mL^{-1}. Finally, $1{:}10^4$ and $1{:}10^5$ dilutions of the cell suspension in 1% sodium chloride were plated on SMM-agar plates and incubated at 37 °C. After 48 hours the average colony size of a plate was determined from high-quality *tiff* images using ImageJ (Schneider et al., 2012).

Plasmid templates and oligonucleotides used for ligation-independent cloning of genes into the pDG148 vector.
Adapter sequences for integration into the StuI-digested pDG148 vector are in lowercase. Regions complementary to the genes specified in the first column are in uppercase.

Gene	Template plasmid	Primer (5′→3′)

bs*pab*A	pET21a_bs*pab*A	`LIC_fo aaggaggaagcaggtATGATTTTAATGATTGATAACTACGATTC`
		`LIC_re gacacgcacgaggtTCACGCAATAACTTCCTTGCG`

bs*pab*B	pET21a_bs*pab*B	`LIC_fo aaggaggaagcaggtATGGCACAACGCAGACC`
		`LIC_re gacacgcacgaggtTCATCTAATTTTTGTCTCTTCTTCGC`

bs*trp*E	pET21a_bs*trp*E	`LIC_fo aaggaggaagcaggtATGAATTTCCAATCAAACATTTCCG`
		`LIC_re gacacgcacgaggtTCAACGCACAATTGTAGAAATCTGTTC`

st*trp*Ex	pET21a_st*trp*Ex	`LIC_fo aaggaggaagcaggtATGCAAACACCAAAACCCAC`
		`LIC_re gacacgcacgaggtTCAGAAGGTCTCCTGTGCATGATG`

sm*trp*Ex	pET21a_sm*trp*Ex	`LIC_fo aaggaggaagcaggtATGATGAACACCAAACCACAAC`
		`LIC_re gacacgcacgaggtTCAGAACACCTCCTTGGCATG`

Supplementary Tables

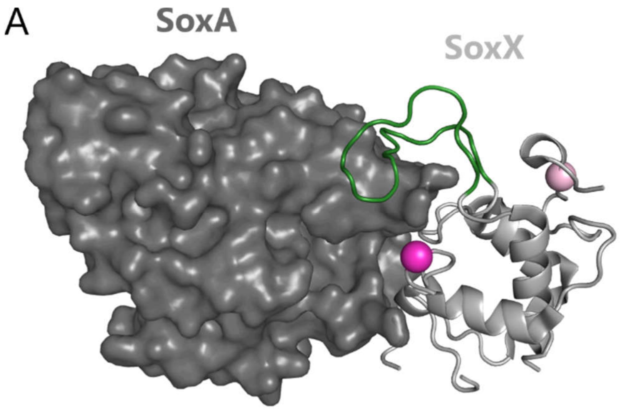
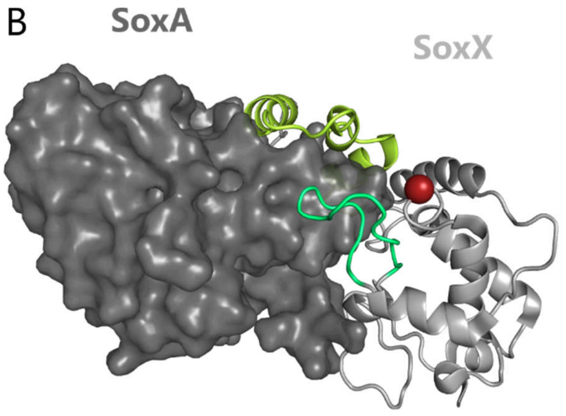
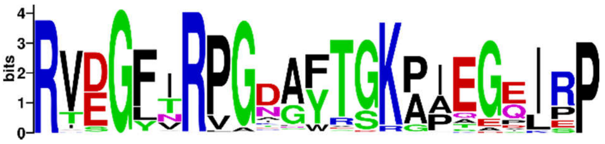
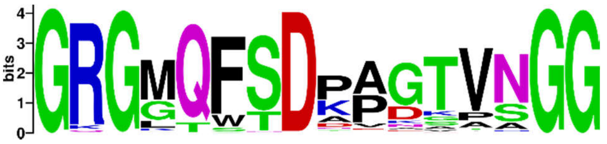
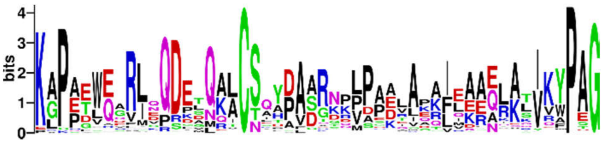
Table S1. ***Related to Figure 1*****: Statistics of the survey for interface add-ons in bacterial, heteromeric protein complexes.**

	Stoichiometries							
	AB	**A_2B_2**	**A_3B_3**	**A_4B_4**	**A_6B_6**	**ABC**	**$A_2B_2C_2$**	**Σ**
Heteromeric complex structures	705	424	112	68	89	197	144	1739
Representative complexes	426	209	47	36	30	102	68	918
Reference complexes	118	93	15	12	7	28	32	305
Survey of insertions in subunits of the reference complexes (*SU*)								
SU with insertions	42	43	5	2	5	8	12	117
insertions	*69*	*70*	*7*	*6*	*5*	*14*	*38*	*209*
SU with ≥ 1 IFR in insertion	30	37	2	2	4	8	12	95
insertions	*45*	*56*	*3*	*4*	*4*	*14*	*36*	*162*
SU with IFRs with $\Delta\Delta G_{IFR\rightarrow Ala}^{complex}$ < -2 kcal/mol	14	24	1	1	2	4	7	53
insertions	*16*	*28*	*1*	*1*	*2*	*7*	*11*	*66*
SU with interface-addons	5	13	0	1	1	2	3	26
interface add-ons	*7*	*15*	*0*	*1*	*1*	*2*	*4*	*30*
Survey of insertions in InterPro homologs (*H*) of *SU*								
H with insertions	75	68	12	10	7	22	18	212
insertions	*134*	*132*	*19*	*18*	*15*	*38*	*36*	*392*

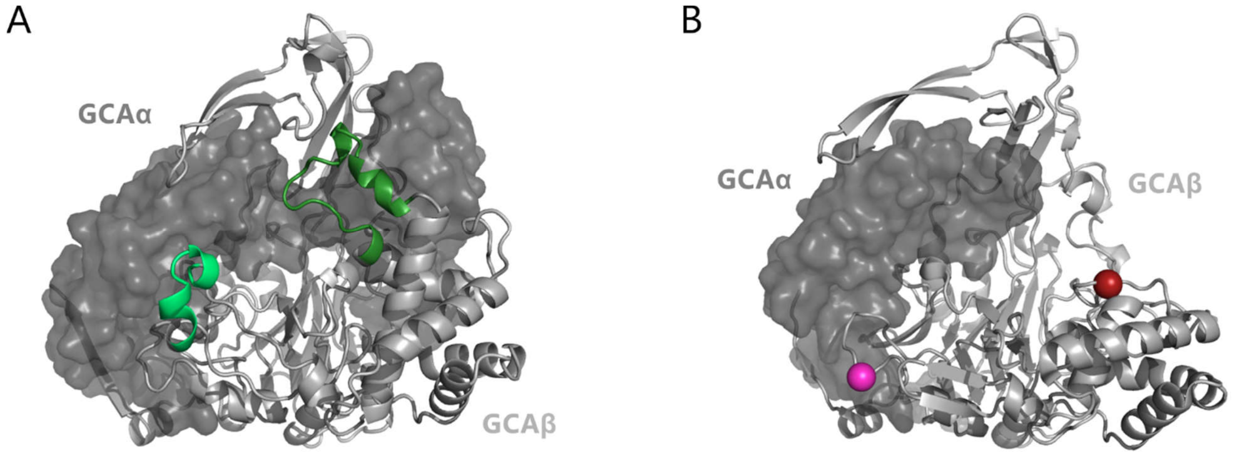
Table S2. ***Related to Figure 1*: List of reference structures with interface add-ons.** The 30 interface add-ons are grouped by subunit stoichiometry of the corresponding complex structure. For each interface add-on, the chain in which it is located and the start and end positions (according to the corresponding FASTA sequence) are given as well as the greatest mCSM-predicted protein-protein affinity change.

Stoichiometry	**PDB-ID**	**Chain**	**Location of interface add-on**	**lowest $\Delta\Delta G_{IFR}^{complex}$ / $\frac{kcal}{mol}$**
AB	1H32	B	97-114	-2.671
AB	3NY7	A	515-525	-3.545
AB	3OCD	B	29-70	-2.616
AB	3OCD	B	108-123	-2.718
AB	4HSR	B	353-371	-3.048
AB	4HSR	B	427-436	-2.828
AB	4YLF	A	246-267	-2.734
A_2B_2	1I1Q	A	71-121	-2.369
A_2B_2	1JJC	A	154-169	-2.063
A_2B_2	1WDK	C	186-207	-2.069
A_2B_2	3CDK	A	216-228	-2.936
A_2B_2	3NUH	B	566-730	-2.489
A_2B_2	3PVT	A	243-278	-2.044
A_2B_2	3RPF	A	122-133	-2.065
A_2B_2	3WWN	A	48-59	-2.138
A_2B_2	3WWN	A	67-79	-2.325
A_2B_2	4CHG	A	9-20	-3.575
A_2B_2	4CHG	A	43-49	-2.315
A_2B_2	4LW4	A	252-265	-2.136
A_2B_2	4ML0	A	51-81	-3.434
A_2B_2	4N6E	B	29-57	-3.132
A_2B_2	4P69	C	171-194	-2.473
A_4B_4	3TND	A	19-32	-2.727
A_6B_6	3J3R	A	405-463	-4.587
ABC	1GX7	E	50-71	-2.966
ABC	3IP4	A	299-374	-3.033
$A_2B_2C_2$	1E7P	A	114-133	-2.403
$A_2B_2C_2$	1E7P	C	103-124	-2.186
$A_2B_2C_2$	1EEX	B	177-202	-2.136
$A_2B_2C_2$	1MHY	B	55-90	-4.342

Table S3. ***Related to Figure 1*****: Examples of reference structures that contain interface add-ons (IAs).** The subunits possessing or lacking interface add-ons are shown in cartoon representation; these subunits are homologs and belong to the same InterPro family. Interface add-ons are colored in shades of green, the position of a missing interface add-on is indicated by a reddish sphere. The interaction partners are shown in surface representation. All structures and protein names are from the Protein Data Bank.

Reference structures 1h32, 3ocd	
	
1h32 *Rhodovulum sulfidophilum* SoxAX c-type cytochrome	3ocd *Starkeya novella* SoxAX c-type cytochrome
Heterodimeric c-type cytochrome complexes involved in the oxidation of thiosulfate (Friedrich et al., 2001).	
IA present in 18% of IPR030999 sequences (mainly Rhodobacterales) 	IA present in 20% of IPR030999 sequences (mainly α- and β-Proteobacteria) 
	IA present in 20% of IPR030999 sequences (mainly Proteobacteria) 
The IA (green) consitutes a major structural difference of SoxX to other members of the cytochrome c superfamily (Bamford et al., 2002). The IA is missing in SoxX from *Chlorobium tepidum*, *Aquifex aeolicus*, and *Rhodopseudomonas palustris*. These organisms do not contain permanent SoxAX complexes (Bamford et al., 2002). *R. sulfidophilum* SoxX lacks the IA found in the homolog from *Starkeya novella* (panel B), the corresponding position is indicated by a magenta sphere.	Misses the IA observed in 1h32 (position indicated by a dark red sphere) but contains an alternative IA at a different location (light green) that contributes to the interaction between SoxA and SoxX (Kilmartin et al., 2011). Also contains a IA at the N-terminus that wraps around the SoxA subunit.

Reference structure 4hsr

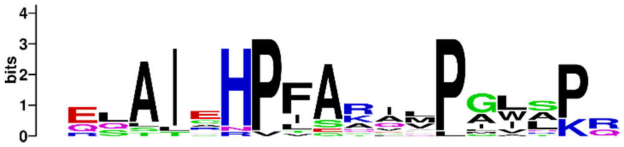


4hsr	1ghd
Pseudomonas N176	*Pseudomonas sp. 130*
Glutaryl-7-aminocephalosporanic acid acylase	Glutaryl-7-aminocephalosporanic acid acylase

Catalyzes the deacylation of glutaryl-7-aminocephalosporanic acid to 7-aminocephalosporanic acid (Kim et al., 1999).

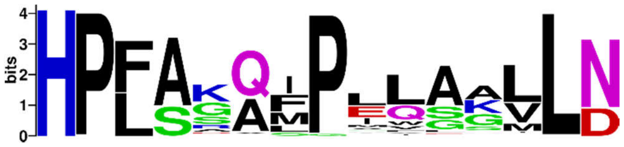
IA present in 0.4% of the IPR002692 sequences

(mainly β- and γ-Proteobacteria)



IA present in 1.7% of the IPR002692 sequences

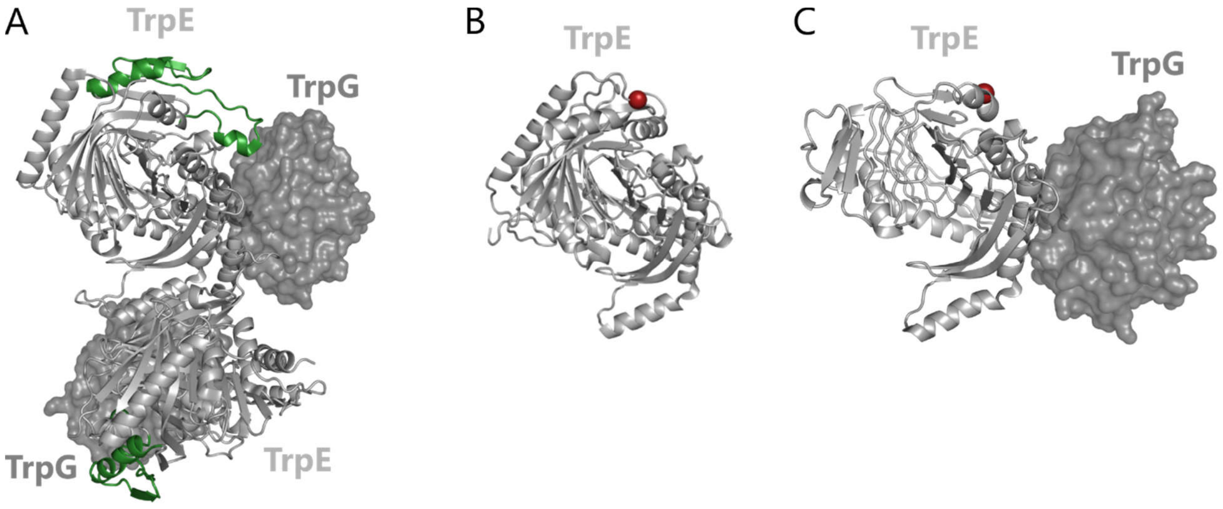
(mainly α- and γ-Proteobacteria)

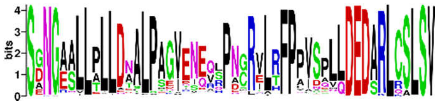


The class-III glutaryl-7-aminocephalosporanic acid acyclase (GCA) from *Pseudomonas N176* contains two IAs (green and light green) in its β-subunit that have been desribed as unique to class-III GCA enzymes (Golden et al., 2013).

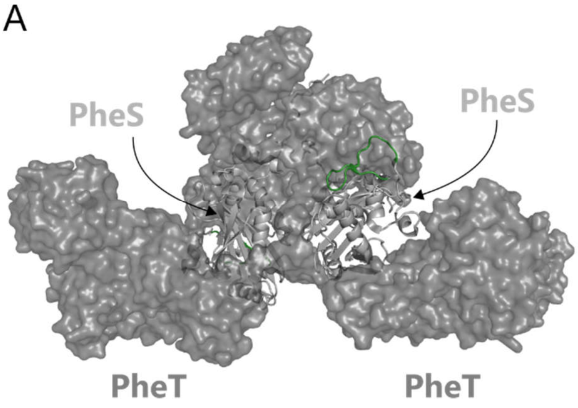
The class-I GCA from *Pseudomonas sp. 130* lacks the two IAs found in its class-III homolog (red and magenta spheres) and the interface between the α-and β-chains is less extended.

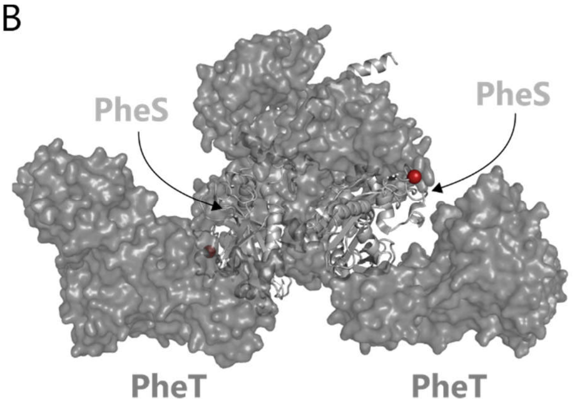
Reference Structure 1i1q



1i1q *Salmonella typhimurium* Anthranilate synthase	4pen *Mycobacterium tuberculosis* Anthranilate synthase	1qdl *Sulfolobus solfataricus* Anthranilate synthase
Catalyze the conversion of chorismate to anthranilate in the committed step of tryptophan biosynthesis (Zalkin, 1973).		
IA present in 6% of the IPR019999 sequences (mainly γ-Proteobacteria) 		
This large IA folds into α-helices, β-strands, and loops that are, for example, not present in anthranilate synthases from *M. tuberculosis* and *S. solfataricus*. It significantly extends the contact area between the TrpE and TrpG protomers.	Lacks the IA found in the anthranilate synthase from *S. typhimurium* (panel A).	Lacks the IA found in the anthranilate synthase from *S. typhimurium* (panel A).

Reference Structure 1jjc



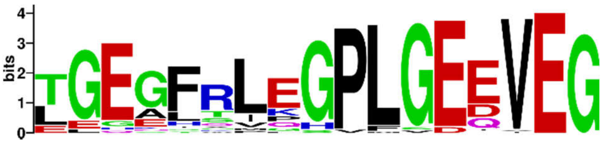


1jjc *Thermus thermophilus* Phenylalanyl-tRNA-synthetases	2rhq *Staphylococcus haemolyticus* Phenylalanyl-tRNA-synthetases

Class II tRNA-synthetases that catalyze the aminoacylation of tRNA^{Phe} (Meinnel et al., 1995).

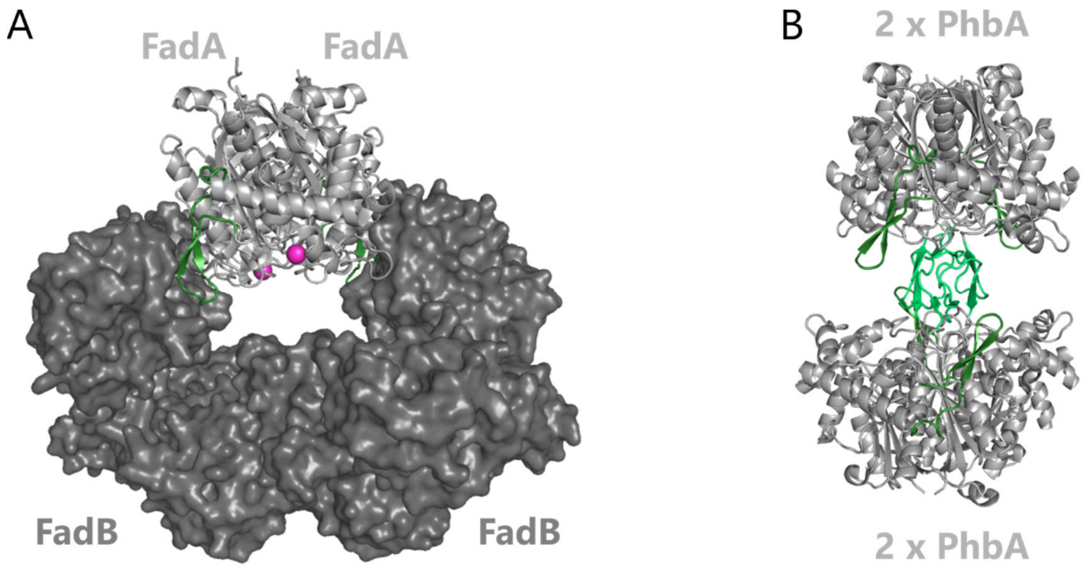
IA present in 1% of the IPR004529 sequences

(mainly the Deinococcus-Thermus group)



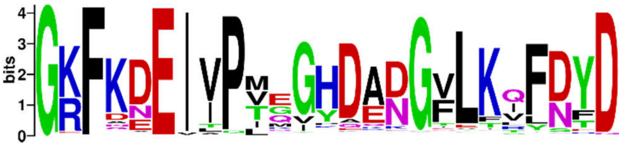
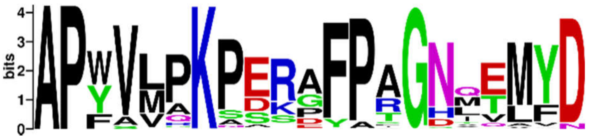
The IA is located at the interface between the PheS and PheT protomers.	The IA from *T. thermophilus* is missing in this homolog and also in related PheS enzymes from Gram-positive bacteria like *Staphylococcus haemolyticus*, *Streptococcus mutans*, *Moraxella catarrhalis*, and *Haemophilus influenza* (Evdokimov et al., 2008).

Reference Structure 1wdk

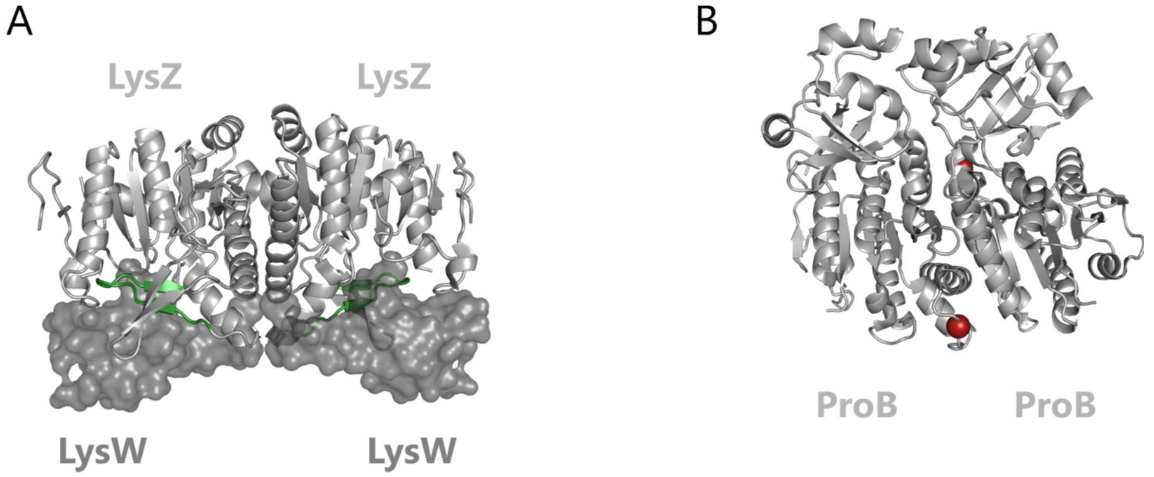


1wdk *Pseudomonas fragi* Fatty acid β-oxidation complex including 3-ketoacyl-CoA thiolase subunits (FadA)	1ulq *Thermus thermophilus* 3-ketoadipyl-CoA thiolase

Catalyze the degradation of 3-ketoacyl-CoA to acetyl-CoA and a shorter acyl-CoA as well as the reverse condensation reaction (Haapalainen et al., 2006).

IA present in 17% of the IPR002155 sequences (mainly γ-Proteobacteria)	**IA** present in 1% of the IPR002155 sequence (mainly α-, β-, and γ-Proteobacteria)
	
The IA is located at the interface between the 3-ketoacyl-CoA-thiolase subunit (FadA) and the FadB subunit which contains the 2-enoyl-CoA-hydratase/isomerase and L-3-hydroxyacyl-CoA-dehydrogenase activities. The IA is important for anchoring the FadB subunit to the FadA subunit (Ishikawa et al., 2004). The IA is for example missing in the FadA homolog “nonspecific lipid-transfer protein” from *Halorubrum distributum*. FadA does not contain the IA found in tetrameric thiolases (indicated by magenta spheres, cf. panel B).	Contains an IA (pale green), which is needed for tetramerization (Harijan et al., 2013). Also contains the IA (green) found in the thiolase from *P. fragi* (panel A).

Reference Structure 3wwn

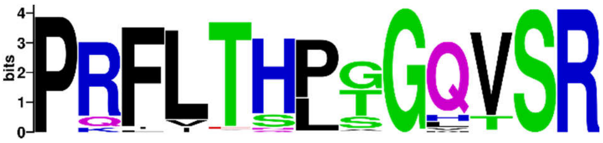


3wwn	2j5v
Thermus thermophilus	*Escherichia coli*
Acetyl-glutamate/acetyl-aminoadipate kinase	Glutamate-5-kinase

Phosphorylate acetyl-glutamate and acetyl-aminoadipate in arginine and lysine biosynthesis and glutamate in proline biosynthesis, respectively (Miyazaki et al., 2001).

IA present in 1% of the IPR001057 sequences

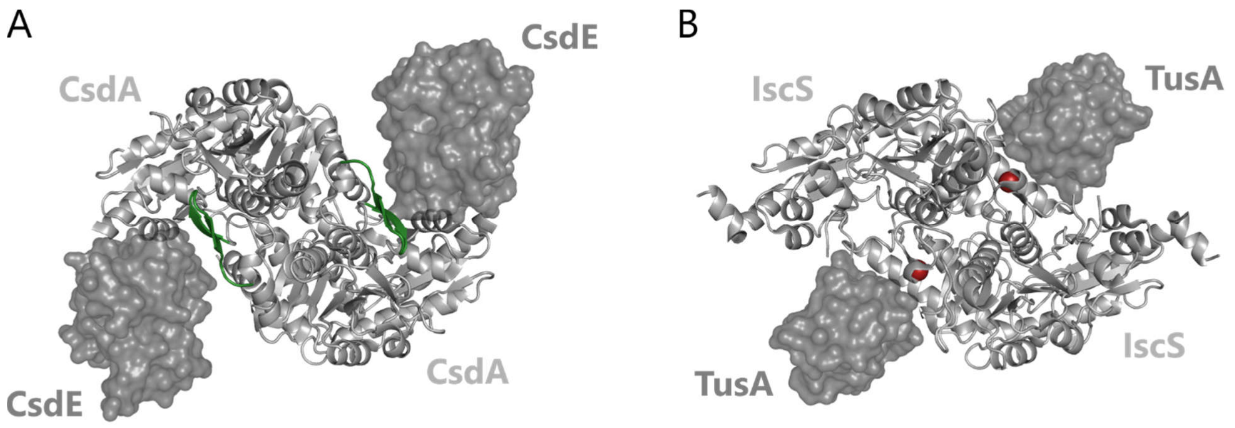
(mainly the Deinococcus-Thermus group)



The LysZ IA is part of its interface with LysW, which acts as a carrier protein for the phosphorylated LysZ products (Yoshida et al., 2015). The IA distinguishes LysZ from other kinase homologs; a LysZ-LysW system is mainly found in thermophilic bacteria and archaea (Yoshida et al., 2015).

The glutamate-5-kinase (ProB) comprises two or four identical protomers that lack the LysZ IA. Here the homodimer is shown. ProB is part of proline biosynthesis and phosphorylates only L-glutamate, whereas LysZ phosphorylates L-acetyl-glutamate and L-acetyl-aminoadipate. The LysZ IA (panel A) is located in the ProB interface (red sphere) required for the formation of homotetramers (Marco-Marin et al., 2007).

Reference Structure 4lw4

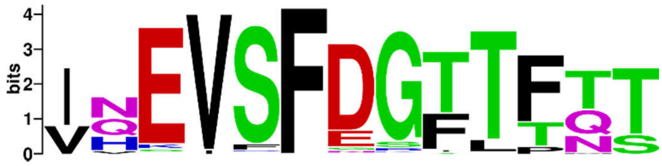


4lw4	3lvk
Escherichia coli	*Escherichia coli*
Cysteine desulfurase	Cysteine desulfurase

Catalyze the transfer of sulphur from cysteine to an acceptor substrate (Mihara and Esaki, 2002).

IA present in 1% of the IPR016454 sequences

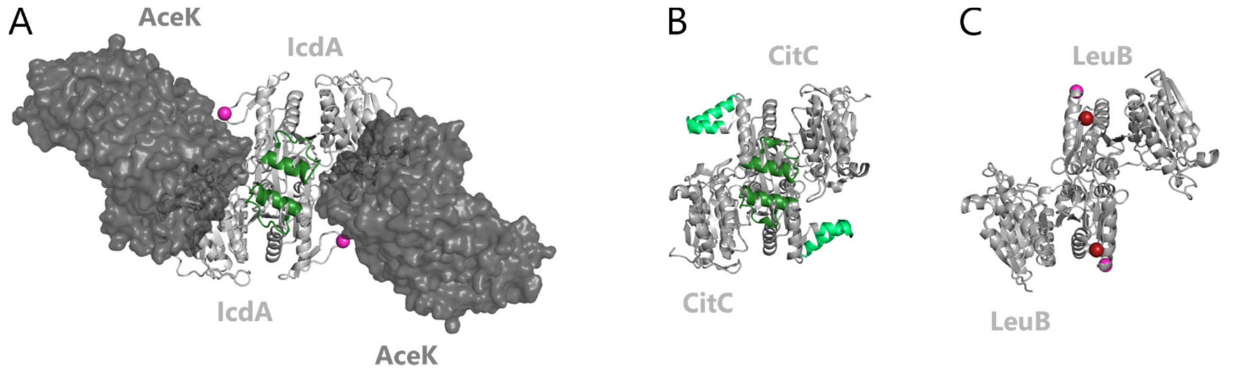
(mainly in Firmicutes)



E. coli CsdA contains an IA comprising two β-strands, which is not found in its homolog IscS from *E. coli* (panel B). The IA is located at the interface to the sulfur-acceptor subunit CsdE. This interface is different to the one utilized for interaction between IscS and TusA in the homologous complex (Kim and Park, 2013). CsdA accepts both L-cysteine and L-selenocystein as substrates and is part of the Csd (cysteine sulfinite desulfinase) system, whose implications are not completely clear yet (Loiseau et al., 2005).

E. coli IscS lacks the IA found in its CsdA homolog (panel A). IscS binds its interaction partner TusA in a different mode compared to the CsdA-CsdE interaction. The IscS-TusA complex is part of the Isc system, which is important for Fe-S-cluster biogenesis in *E. coli* (Mihara and Esaki, 2002).

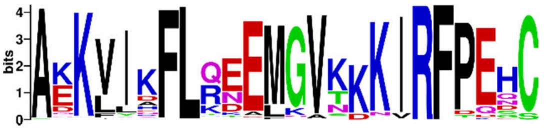
Reference Structure 4p69



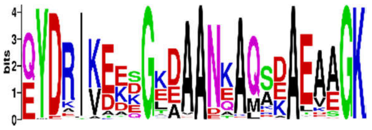
Reference structure 4p69	1hqs	1cm7
Escherichia coli	*Bacillus subtilis*	*Escherichia coli*
Isocitrate dehydrogenase	Isocitrate dehydrogenase	Isopropylmalate dehydrogenase

Isocitrate dehydrogenase catalyzes the converion of isocitrate to α-ketoglutarate in the Krebs cycle (Dean and Koshland, 1993). Isopropylmalate dehydrogenase catalyzes a similar reaction (decarboxylation of isopropylmalate to oxomethylvalerate) in leucine biosynthesis (Parsons and Burns, 1969).

IA present in 15% of the IPR001804 sequences (mainly in Proteobacteria, Firmicutes, and Bacteroidetes)



IA present in 5% of IPR001804 seq. (mainly in Firmicutes)



The isocitrate dehydrogenase from *E. coli* (IcdA) contains an IA that folds into a short α-helix (green). This helix and its surroundings have been desribed as a clasp domain that mediates interactions between the two IcdA monomers (Vinekar et al., 2012). It also contributes to the interface to the isocitrate dehydrogenase kinase/phosphatase AceK, which phoshporylates and dephosphorylates IcdA to regulate isocitrate flux between the Krebs cycle (LaPorte and Koshland, 1982) and the glyoxylate bypass (LaPorte, 1993). This helix is also present in the homologous enzyme CitC from *B. subtilis* (cf. panel B). Our analysis indicated the absence of the large insertion in IcdA that is found in CitC (magenta spheres; light green insertion in panel B). The absence of this insertion has been discussed as important for interactions with AceK (Yates et al., 2011).

The isocitrate dehydrogenase CitC from *B. subtilis* cotaints the same clasp-domain helix as IcdA (dark green). Additionally it contains a relatively large insertion comprising two helices (light green) that prevents interaction with AceK (Singh et al., 2002).

The isopropylmalate dehydrogenase from *E: coli* lacks both insertions found in IcdA (red spheres) and CitC (magenta spheres). No interactions with kinase-phosphates like AceK has been described for LeuB.

Table S4. ***Related to Figure 4A***: Fractions of TrpEx- and TrpE-species that possess a certain pattern of co-occurring synthases (TrpEx, TrpE, PabB) and glutaminases (TrpG and PabA).

TrpEx-species		
Fraction	**co-occurrences**	**Comment**
84%	TrpEx + TrpG PabB + PabA	Typical TrpEx-species with two separate glutaminases for the two synthases TrpEx and PabB.
16%	various patterns	These TrpEx-species mostly contain various combinations of multiple copies of one or more of the synthases and glutaminases. The situation is unclear for some species. This is presumably due to incompletely sequenced genomes or a symbiotic lifestyle, which lead to the loss of one or the other protein.
TrpE-species		
Fraction	**co-occurrences**	**Comment**
59%	TrpE/PabB + PabA	Typical TrpE-species with only a single PabA glutaminase for the two synthases TrpE and PabB.
23%	TrpE + PabA	TrpE-species that lack PabB. Among these species are many Archaea that lack the classical *pab* genes for folate biosynthesis. Incompletely sequenced TrpE-species may also contribute to this fraction.
18%	various patterns	These TrpE-species contain various combinations of TrpE- and PabA-homologs or lack one or the other proteins; again, most likely due to incomplete genome sequences. Some of these species contain additional copies of one or the other protein that may have developed specialized roles in supporting metabolism.

Table S5. ***Related to Figure 5A*****: SEC-SLS experiments for determining complex formation**. The apparent molecular weight ($\mathrm{MW}_{\mathrm{app}}$) was determined from refractive index and right-angle light scattering measurements. The calculated molecular weight ($\mathrm{MW}_{\mathrm{calc}}$) derives from the amino-acid sequence and was determined via Expasy ProtParam. Abbreviations: n.i.: peak(s) not identifiable, e. g. when proteins could not be separated with S75 or S200 SEC-columns. Table continues over two pages.

Synthase	**Glutaminase**	**$\mathrm{MW}_{\mathrm{app}}$ / kDa**	**$\mathrm{MW}_{\mathrm{calc}}$ / kDa**	**Percentage Error**	**Deduced oligomeric state**
stTrpEx	---	58.2	59.8	3%	monomer
ecTrpEx	---	58.9	58.6	1%	monomer
smTrpEx	---	60.9	58.2	5%	monomer
ppTrpE	---	117.1	110.5	6%	dimer
ssTrpE	---	48.9	48.7	<1%	monomer
ppPabB	---	55.9	50.2	11%	monomer
stPabB	---	56.1	52.0	8%	monomer
ecPabB	---	50.1	52.3	4%	monomer
bsPabB	---	59.8	54.3	10%	monomer
---	stTrpG	21.2	21.9	3%	monomer
---	ecTrpGD	58.5	57.9	1%	monomer
---	ppPabA	23.1	22.7	2%	monomer
---	smPabA	23.1	22.2	4%	monomer
---	ecPabA	23.5	21.8	8%	monomer
---	bsPabA	23.7	22.8	4%	monomer
stTrpEx	stTrpG	156.9	160.1	2%	Ex$_2$:G$_2$ Tetramer
stTrpEx	ecTrpGD	427.5	464.3	8%	Ex$_4$:GD$_4$ Octamer
stTrpEx	ppPabA	--- 59.4 22.1	--- 58.2 22.7	--- 2% 3%	no complex stTrpEx monomer ppPabA monomer
stTrpEx	smPabA	--- 57.4 22.1	--- 58.2 22.3	--- 1% 1%	no complex stTpEx monomer smPabA monomer
stTrpEx	ecPabA	--- 57.5 19.5	--- 58.2 21.8	--- 1% 11%	no complex stTrpEx monomer ecPabA monomer
stTrpEx	bsPabA	--- 56.6 20.8	--- 58.2 22.8	--- 3% 8%	no complex stTrpEx monomer bsPabA monomer
ecTrpEx	stTrpG	158.6	160.9	1%	Ex$_2$:G$_2$ Tetramer
ecTrpEx	ecTrpGD	424.8	466.0	9%	Ex$_4$:GD$_4$ Ocatmer
ecTrpEx	ppPabA	--- 55.8 20.7	--- 58.6 22.7	--- 5% 9%	no complex ecTrpEx monomer ppPabA monomer
ecTrpEx	smPabA	--- 58.8 22.2	--- 58.6 22.3	--- <1% 1%	no complex ecTpEx monomer smPabA monomer
ecTrpEx	ecPabA	--- 61.9 25.1	--- 58.6 21.8	--- 6% 15%	no complex ecTrpEx monomer ecPabA monomer
ecTrpEx	bsPabA	--- 60.9 23.3	--- 58.6 22.8	--- 3% 2%	no complex ecTrpEx monomer bsPabA monomer

Synthase	Glutaminase	MW$_{app}$ / kDa	MW$_{calc}$ / kDa	Percentage Error	Deduced oligomeric state
smTrpEx	stTrpG	155.1	160.2	3%	Ex$_2$:G$_2$ Tetramer
smTrpEx	ecTrpGD	238.6	232.3	3%	Ex$_2$:GD$_2$ Tetramer
smTrpEx	ppPabA	--- 61.8 23.2	--- 58.2 22.7	--- 6% 2%	no complex smTrpEx monomer ppPabA monomer
smTrpEx	smPabA	--- 64.2 23.4	--- 58.2 22.3	--- 10% 5%	no complex smTpEx monomer smPabA monomer
smTrpEx	ecPabA	--- 63.3 21.3	--- 58.2 21.8	--- 9% 3%	no complex smTrpEx monomer ecPabA monomer
smTrpEx	bsPabA	--- 64.2 26.2	--- 58.2 22.8	--- 10% 15%	no complex smTrpEx monomer bsPabA monomer
ppTrpE	stTrpG	--- 121.8 24.7	--- 110.8 21.9	--- 10% 13%	no complex ppTrpE dimer stTrpG monomer
ppTrpE	ecTrpGD	--- n.i.	--- ---	--- ---	no complex ---
ppTrpE	ppPabA	156.7	156.3	<1%	E$_2$:A$_2$ tetramer
ppTrpE	smPabA	157.6	155.4	1%	E$_2$:A$_2$ tetramer
ppTrpE	ecPabA	149.0	154.4	4%	E$_2$:A$_2$ tetramer
ppTrpE	bsPabA	143.6	156.3	8%	E$_2$:A$_2$ tetramer
ssTrpE	stTrpG	--- 54.9 42.8	--- 48.7 21.9	--- 13% 13%	no complex ssTrpE monomer stTrpG monomer
ssTrpE	ecTrpGD	--- 50.9 66.1	--- 48.7 57.9	--- 4% 14%	no complex ssTrpE monomer ecTrpGD monomer
ssTrpE	ppPabA	72.8	71.5	2%	E:A dimer
ssTrpE	smPabA	72.6	71.0	2%	E:A dimer
ssTrpE	ecPabA	71.1	70.5	1%	E:A dimer
ssTrpE	bsPabA	70.5	71.5	1%	E:A dimer
ppPabB	stTrpG	--- 50.1 21.0	--- 50.2 21.9	--- <1% 4%	no complex ppPabB monomer stTrpG monomer
ppPabB	ecTrpGD	--- 51.2 59.4	--- 50.2 57.9	--- 2% 2%	no complex ppPabB monomer ecTrpGD monomer
ppPabB	ppPabA	67.0	72.9	8%	B:A dimer
ppPabB	smPabA	71.9	72.5	1%	B:A dimer
ppPabB	ecPabA	74.1	72.0	3%	B:A dimer
ppPabB	bsPabA	60.5	73.0	17%	B:A dimer

Synthase	**Glutaminase**	**MW_{app} / kDa**	**MW_{calc} / kDa**	**Percentage Error**	**Deduced oligomeric state**
stPabB	stTrpG	---	---	---	no complex
		62.4	52.0	20%	stPabB monomer
		21.2	21.9	3%	stTrpG monomer
stPabB	ecTrpGD	---	---	---	no complex
		n.i.	---	---	---
stPabB	ppPabA	66.1	74.8	12%	B:A dimer
stPabB	smPabA	69.6	74.3	6%	B:A dimer
stPabB	ecPabA	67.6	73.8	8%	B:A dimer
stPabB	bsPabA	62.1	74.8	17%	B:A dimer
ecPabB	stTrpG	---	---	---	no complex
		51.7	52.0	1%	ecPabB monomer
		21.0	21.9	4%	stTrpG monomer
ecPabB	ecTrpGD	---	---	---	no complex
		52.5	52.0	1%	ecPabB monomer
		n.i.	---	---	---
ecPabB	ppPabA	59.4	74.8	21%	B:A dimer
ecPabB	smPabA	70.6	74.3	5%	B:A dimer
ecPabB	ecPabA	66.4	73.9	10%	B:A dimer
ecPabB	bsPabA	63.6	74.8	15%	B:A dimer
bsPabB	stTrpG	---	---	---	no complex
		54.1	54.3	<1%	bsPabB monomer
		21.0	21.9	4%	stTrpG monomer
bsPabB	ecTrpGD	---	---	---	no complex
		n.i.	---	---	---
bsPabB	ppPabA	79.5	77.1	4%	B:A dimer
bsPabB	smPabA	72.3	76.6	6%	B:A dimer
bsPabB	ecPabA	69.0	76.2	9%	B:A dimer
bsPabB	bsPabA	74.8	77.1	3%	B:A dimer

Table S6. ***Related to Figure 5A*****: Native mass spectrometry experiments for determining complex formation between various synthases and glutaminases**. The apparent molecular weight (MW_{app}) was determined from m/z values and is given with standard deviations. The calculated molecular weight (MW_{calc}) derives from the amino-acid sequence and was determined via Expasy ProtParam. Obtained molecular weights are consistent with the expected molecular weights.

Synthase	**Glutaminase**	**MW_{app} / Da**	**MW_{calc} / kDa**	**Percentage Error**	**Deduced oligomeric state**
stTrpEx	---	58235 ± 78	58153	1%	monomer
		116768 ± 102	116306	<1%	dimer
ecTrpEx	---	58956 ± 116	58559	<1%	monomer
		118655 ± 323	117118	1%	dimer
smTrpEx	---	58331 ± 61	58229	<1%	monomer
		116758 ± 59	116458	<1%	dimer
ppTrpE	---	112126 ± 36	110482	1%	dimer
ppPabB	---	50339 ± 15	50205	<1%	monomer
ecPabB	---	52470 ± 41	52035	<1%	monomer
---	stTrpG	21787 ± 0	21915	<1%	monomer
		43599 ± 21	43830	<1%	dimer
---	ecTrpGD	58041 ± 34	57935	<1%	monomer
---	ppPabA	22727 ± 0	22734	<1%	monomer
---	ecPabA	21866 ± 22	21837	<1%	monomer
		43904 ± 78	43674	<1%	dimer
stTrpEx	stTrpG	122026 ± 330	116306	5%	Ex_2 dimer
		160727 ± 29	160136	<1%	Ex_2:G_2 tetramer
stTrpEx	ecTrpGD	117527 ± 157	116306	1%	Ex_2 dimer
		176069 ± 376	174241	1%	Ex_2:GD trimer
		234252 ± 286	232176	1%	Ex_2:GD_2 tetramer
stTrpEx	ppPabA	---	---	---	no complex
ecTrpEx	ecTrpGD	118202 ± 75	117118	1%	Ex_2 dimer
		184066 ± 232	175053	5%	Ex_2:GD trimer
		235458 ± 36	232988	1%	Ex_2:GD_2 tetramer
smTrpEx	ecTrpGD	116888 ± 124	116458	<1%	Ex_2 dimer
		175649 ± 107	174393	<1%	Ex_2:GD trimer
		234467 ± 45	232916	<1%	Ex_2:GD_2 tetramer
ppTrpE	stTrpG	111771 ± 72	110482	1%	E_2 dimer
		133650 ± 36	132397	<1%	E_2:G trimer
ppTrpE	ppPabA	157468 ± 176	155950	<1%	E_2:A_2 tetramer
ppPabB	ppPabA	73482 ± 44	72939	<1%	B:A dimer
ecPabB	ecPabA	74615 ± 150	73872	1%	B:A dimer

Table S7. ***Related to Figure 5B*****: Kinetic parameters of synthases in complex with different glutaminases for the glutamine-dependent formation of anthranilate (for TrpEx and TrpE) and aminodeoxychorismate (for PabB)**. Values were determined at 25 °C and in the presence of 20 mM glutamine. Each combination was assayed at three different chorismate concentrations (90, 100, and 110 µM) and average values with standard deviations are listed. Combinations were assigned not active (n.a.), if the initial rates of anthranilate or aminodeoxychorismate formation were at least 240-fold lower (for TrpEx/TrpE) or 20-fold lower (for PabB) as the average initial rates of corresponding functional complexes. In these cases, no catalytic efficiency k_{cat}/K_M^{CH} could be determined (n.d.). Table continues to the next page.

Synthase	Glutaminase	k_{cat} /s^{-1}	K_M^{CH} /µM	k_{cat}/K_M^{CH} /M^{-1}s^{-1}
stTrpEx	stTrpG	3.7 ± 0.3	11.3 ± 2.9	$3.5 \cdot 10^5$
stTrpEx	ecTrpGD	3.6 ± 0.5	10.5 ± 2.3	$3.6 \cdot 10^5$
stTrpEx	ppPabA	n.a.	n.a.	n.d.
stTrpEx	smPabA	n.a.	n.a.	n.d.
stTrpEx	ecPabA	n.a.	n.a.	n.d.
stTrpEx	bsPabA	n.a.	n.a.	n.d.
ecTrpEx	stTrpG	3.8 ± 0.8	12.0 ± 2.6	$3.3 \cdot 10^5$
ecTrpEx	ecTrpGD	3.7 ± 0.1	10.0 ± 2.8	$3.9 \cdot 10^5$
ecTrpEx	ppPabA	n.a.	n.a.	n.d.
ecTrpEx	smPabA	n.a.	n.a.	n.d.
ecTrpEx	ecPabA	n.a.	n.a.	n.d.
ecTrpEx	bsPabA	n.a.	n.a.	n.d.
smTrpEx	stTrpG	5.1 ± 0.2	28.0 ± 3.6	$1.9 \cdot 10^5$
smTrpEx	ecTrpGD	4.3 ± 0.2	20.0 ± 5.0	$2.2 \cdot 10^5$
smTrpEx	ppPabA	n.a.	n.a.	n.d.
smTrpEx	smPabA	n.a.	n.a.	n.d.
smTrpEx	ecPabA	n.a.	n.a.	n.d.
smTrpEx	bsPabA	n.a.	n.a.	n.d.
ppTrpE	stTrpG	n.a.	n.a.	n.d.
ppTrpE	ecTrpGD	n.a.	n.a.	n.d.
ppTrpE	ppPabA	3.4 ± 0.3	6.3 ± 1.5	$5.4 \cdot 10^5$
ppTrpE	smPabA	3.4 ± 0.8	7.3 ± 1.2	$4.6 \cdot 10^5$
ppTrpE	ecPabA	2.7 ± 0.03	2.1 ± 0.2	$1.3 \cdot 10^6$
ppTrpE	bsPabA	0.8 ± 0.1	5.7 ± 1.0	$1.4 \cdot 10^5$
ssTrpE	stTrpG	n.a.	n.a.	n.d.
ssTrpE	ecTrpGD	0.1 ± 0.01	16.0 ± 2.1	$4.4 \cdot 10^3$
ssTrpE	ppPabA	0.1 ± 0.01	11.6 ± 5.5	$8.7 \cdot 10^3$
ssTrpE	smPabA	0.2 ± 0.01	10.9 ± 0.5	$1.6 \cdot 10^4$
ssTrpE	ecPabA	0.2 ± 0.01	9.8 ± 2.3	$1.7 \cdot 10^4$
ssTrpE	bsPabA	0.3 ± 0.01	8.6 ± 2.7	$3.6 \cdot 10^4$
ppPabB	stTrpG	n.a.	n.a.	n.d.
ppPabB	ecTrpGD	n.a.	n.a.	n.d.
ppPabB	ppPabA	0.2 ± 0.02	33.3 ± 2.3	$6.7 \cdot 10^3$
ppPabB	smPabA	0.2 ± 0.04	35.3 ± 11.9	$5.7 \cdot 10^3$
ppPabB	ecPabA	0.1 ± 0.02	21.9 ± 3.9	$5.8 \cdot 10^3$
ppPabB	bsPabA	0.1 ± 0.01	76.7 ± 21.0	$9.9 \cdot 10^2$

Synthase	**Glutaminase**	**k_{cat} /s^{-1}**	**K_M^{CH} /µM**	**k_{cat}/K_M^{CH} /M^{-1}s^{-1}**
stPabB	stTrpG	n.a.	n.a.	n.d.
stPabB	ecTrpGD	n.a.	n.a.	n.d.
stPabB	ppPabA	0.6 ± 0.03	35.9 ± 1.5	$1.6 \cdot 10^4$
stPabB	smPabA	0.8 ± 0.1	44.5 ± 5.9	$1.9 \cdot 10^4$
stPabB	ecPabA	0.7 ± 0.1	34.3 ± 3.0	$2.0 \cdot 10^4$
stPabB	bsPabA	0.1 ± 0.01	43.4 ± 2.1	$3.0 \cdot 10^3$
ecPabB	stTrpG	n.a.	n.a.	n.d.
ecPabB	ecTrpGD	n.a.	n.a.	n.d.
ecPabB	ppPabA	0.7 ± 0.1	24.8 ± 3.4	$2.8 \cdot 10^4$
ecPabB	smPabA	0.7 ± 0.1	24.3 ± 2.8	$2.8 \cdot 10^4$
ecPabB	ecPabA	0.6 ± 0.1	20.9 ± 5.1	$2.7 \cdot 10^4$
ecPabB	bsPabA	0.1 ± 0.01	15.6 ± 3.4	$5.2 \cdot 10^3$
bsPabB	stTrpG	n.a.	n.a.	n.d.
bsPabB	ecTrpGD	n.a.	n.a.	n.d.
bsPabB	ppPabA	n.a.	n.a.	n.d.
bsPabB	smPabA	n.a.	n.a.	n.d.
bsPabB	ecPabA	n.a.	n.a.	n.d.
bsPabB	bsPabA	n.a.	n.a.	n.d.

Table S8. ***Related to Figure 5C*****: Apparent turnover rates (*k_{app}*) of various glutaminases for the hydrolysis of glutamine at 25 °C and in the presence of 4 mM glutamine**. Each glutaminase was assayed alone and in the presence of the listed synthases. The apparent turnover rates are average values of at least three independent measurements and are listed with standard deviations. The stimulation factors (f_{stim}) were calculated from k_{app} of a glutaminase in the presence of a given synthase and k_{app} in the absence of any synthase. Note that the latter was determined separately for each of the listed combinations. Abbreviations: n.a.: not active (k_{app} smaller than 0.001 s^{-1}); n.d.: not determined. Table continues to the next page.

Glutaminase	Synthase	k_{app} /s^{-1}	f_{stim}
stTrpG	---	n.a.	---
stTrpG	stTrpEx	0.065 ± 0.004	n.d.
stTrpG	ecTrpEx	0.073 ± 0.003	n.d.
stTrpG	smTrpEx	0.058 ± 0.005	n.d.
stTrpG	ppTrpE	0.024 ± 0.007	n.d.
stTrpG	ssTrpE	0.019 ± 0.001	n.d.
stTrpG	ppPabB	n.a.	n.d.
stTrpG	stPabB	0.002 ± 0.0001	n.d.
stTrpG	ecPabB	0.001 ± 0.0001	n.d.
stTrpG	bsPabB	n.a.	n.d.
ecTrpGD	---	n.a.	---
ecTrpGD	stTrpEx	0.065 ± 0.007	n.d.
ecTrpGD	ecTrpEx	0.061 ± 0.005	n.d.
ecTrpGD	smTrpEx	0.010 ± 0.002	n.d.
ecTrpGD	ppTrpE	0.005 ± 0.001	n.d.
ecTrpGD	ssTrpE	0.019 ± 0.001	n.d.
ecTrpGD	ppPabB	n.a.	n.d.
ecTrpGD	stPabB	n.a.	n.d.
ecTrpGD	ecPabB	n.a.	n.d.
ecTrpGD	bsPabB	n.a.	n.d.
ppPabA	---	0.015 ± 0.004	---
ppPabA	stTrpEx	0.012 ± 0.001	0.9 ± 0.1
ppPabA	ecTrpEx	0.008 ± 0.001	0.9 ± 0.1
ppPabA	smTrpEx	0.014 ± 0.001	1.0 ± 0.1
ppPabA	ppTrpE	0.048 ± 0.001	4.7 ± 0.8
ppPabA	ssTrpE	0.337 ± 0.012	20.8 ± 0.6
ppPabA	ppPabB	0.061 ± 0.005	5.0 ± 0.7
ppPabA	stPabB	0.077 ± 0.006	3.7 ± 0.6
ppPabA	ecPabB	0.124 ± 0.010	10.3 ± 0.2
ppPabA	bsPabB	0.129 ± 0.005	12.6 ± 1.4
smPabA	---	0.007 ± 0.002	---
smPabA	stTrpEx	0.007 ± 0.001	0.9 ± 0.1
smPabA	ecTrpEx	0.008 ± 0.001	0.8 ± 0.1
smPabA	smTrpEx	0.008 ± 0.001	0.9 ± 0.1
smPabA	ppTrpE	0.059 ± 0.003	10.0 ± 1.7
smPabA	ssTrpE	0.159 ± 0.015	16.6 ± 0.4
smPabA	ppPabB	0.054 ± 0.005	11.0 ± 2.1
smPabA	stPabB	0.117 ± 0.003	20.0 ± 4.9
smPabA	ecPabB	0.126 ± 0.006	24.4 ± 8.4
smPabA	bsPabB	0.199 ± 0.012	30.5 ± 4.1

Glutaminase	**Synthase**	k_{app} **/s^{-1}**	f_{Stim}
ecPabA	---	0.004 ± 0.001	---
ecPabA	stTrpEx	0.010 ± 0.001	3.6 ± 0.2
ecPabA	ecTrpEx	0.005 ± 0.001	1.8 ± 0.2
ecPabA	smTrpEx	0.007 ± 0.001	2.2 ± 0.1
ecPabA	ppTrpE	0.043 ± 0.007	17.5 ± 4.6
ecPabA	ssTrpE	0.176 ± 0.008	47.1 ± 5.6
ecPabA	ppPabB	0.042 ± 0.001	13.8 ± 2.4
ecPabA	stPabB	0.069 ± 0.013	14.2 ± 1.7
ecPabA	ecPabB	0.093 ± 0.003	41.1 ± 8.5
ecPabA	bsPabB	0.219 ± 0.022	55.9 ± 12.1
bsPabA	---	0.003 ± 0.001	---
bsPabA	stTrpEx	0.003 ± 0.001	1.3 ± 0.1
bsPabA	ecTrpEx	0.002 ± 0.001	1.1 ± 0.1
bsPabA	smTrpEx	0.003 ± 0.001	1.1 ± 0.1
bsPabA	ppTrpE	0.033 ± 0.002	10.1 ± 0.6
bsPabA	ssTrpE	0.217 ± 0.017	67.5 ± 2.5
bsPabA	ppPabB	0.016 ± 0.001	8.6 ± 0.5
bsPabA	stPabB	0.011 ± 0.002	3.8 ± 1.4
bsPabA	ecPabB	0.013 ± 0.003	3.8 ± 1.3
bsPabA	bsPabB	0.073 ± 0.005	20.9 ± 2.0

Table S9. ***Related to Table 1*****: Mass spectrometric characterization of interactions between stTrpEx/stTrpEx_Δ synthases and stTrpG/ppPabA glutaminases**. The apparent molecular weight (MW$_{app}$) was determined from m/z values and is given with standard deviations. The calculated molecular weight (MW$_{calc}$) derives from the amino-acid sequence and was determined via Expasy ProtParam. Obtained molecular weights are consistent with the expected molecular weights.

Synthase	**Glutaminase**	**MW$_{app}$ /kDa**	**MW$_{calc}$ /kDa**	**Percentage Error**	**Deduced Oligomeric State**
stTrpEx	stTrpG	160727 ± 29	160136	<1%	Ex$_2$:G$_2$ tetramer
		122026 ± 330	116306	5%	Ex$_2$ dimer
		58235 ± 78	58153	<1%	Ex monomer
stTrpEx	ppPabA	116768 ± 102	116306	<1%	Ex$_2$ dimer
		58235 ± 78	58153	<1%	Ex monomer
		22727 ± 0	22734	<1%	A monomer
stTrpEx_Δ	stTrpG	160866 ± 77	159112	<1%	Ex_Δ$_2$:G$_2$ tetramer
		58409 ± 130	57641	1.3%	Ex_Δ monomer
		43791 ± 32	43830	<1%	G$_2$ dimer
		21784 ± 2	21915	<1%	G monomer
stTrpEx_Δ	ppPabA	162033 ± 32	160750	<1%	Ex_Δ$_2$:A$_2$ tetramer
		81043 ± 22	80375	<1%	Ex_Δ: A dimer
		58210 ± 82	57641	<1%	Ex_Δ monomer
		22753 ± 12	22734	<1%	A monomer

Table S10. ***Related to Table 2*****: Mass spectrometric characterization of interactions between ppPabA* and TrpEx, TrpE, and PabB, respectively**. The apparent molecular weight (MW_{app}) was determined from m/z values and is given with standard deviations. The calculated molecular weight (MW_{calc}) derives from the amino-acid sequence and was determined via Expasy ProtParam. Obtained molecular weights are consistent with the expected molecular weights.

Synthase	Glutaminase	MW_{app} /kDa	MW_{calc} /kDa	Percentage Error	Deduced Oligomeric State
stTrpEx	ppPabA*	162612 ± 30	161794	<1%	Ex_2:A_2 tetramer
		81143 ± 51	80897	<1%	Ex:A dimer
		45708 ± 35	45488	<1%	A_2 dimer
		22843 ± 0	22744	<1%	A monomer
ecTrpEx	ppPabA*	163392 ± 51881	162606	<1%	Ex_2:A_2 tetramer
		530 ± 47454568	81303	<1%	Ex:A dimer
		8 ± 1	45488	<1%	A_2 dimer
		22843 ± 0	22744	<1%	A monomer
smTrpEx	ppPabA*	162751 ± 37881	161946	<1%	Ex_2:A_2 tetramer
		394 ± 31	80973	<1%	Ex:A dimer
		45688± 1	45488	<1%	A_2 dimer
		22843 ± 1	22744	<1%	A monomer
ppTrpE	ppPabA*	158485 ± 177	155950	1.6%	E_2:A_2 tetramer
		135377 ± 94	133226	1.5%	E_2:A trimer
		112410 ± 62	110482	1.7%	E_2 dimer
		45825 ± 80	45488	<1%	A_2 dimer
		22843 ± 0	22744	<1%	A monomer
ppPabB	ppPabA*	73365 ± 62	72949	<1%	A:B dimer
		50399 ± 67	50205	<1%	B monomer
		45750 ± 81	45488	<1%	A_2 dimer
		22843 ± 0	22744	<1%	A monomer

Supplementary Figures

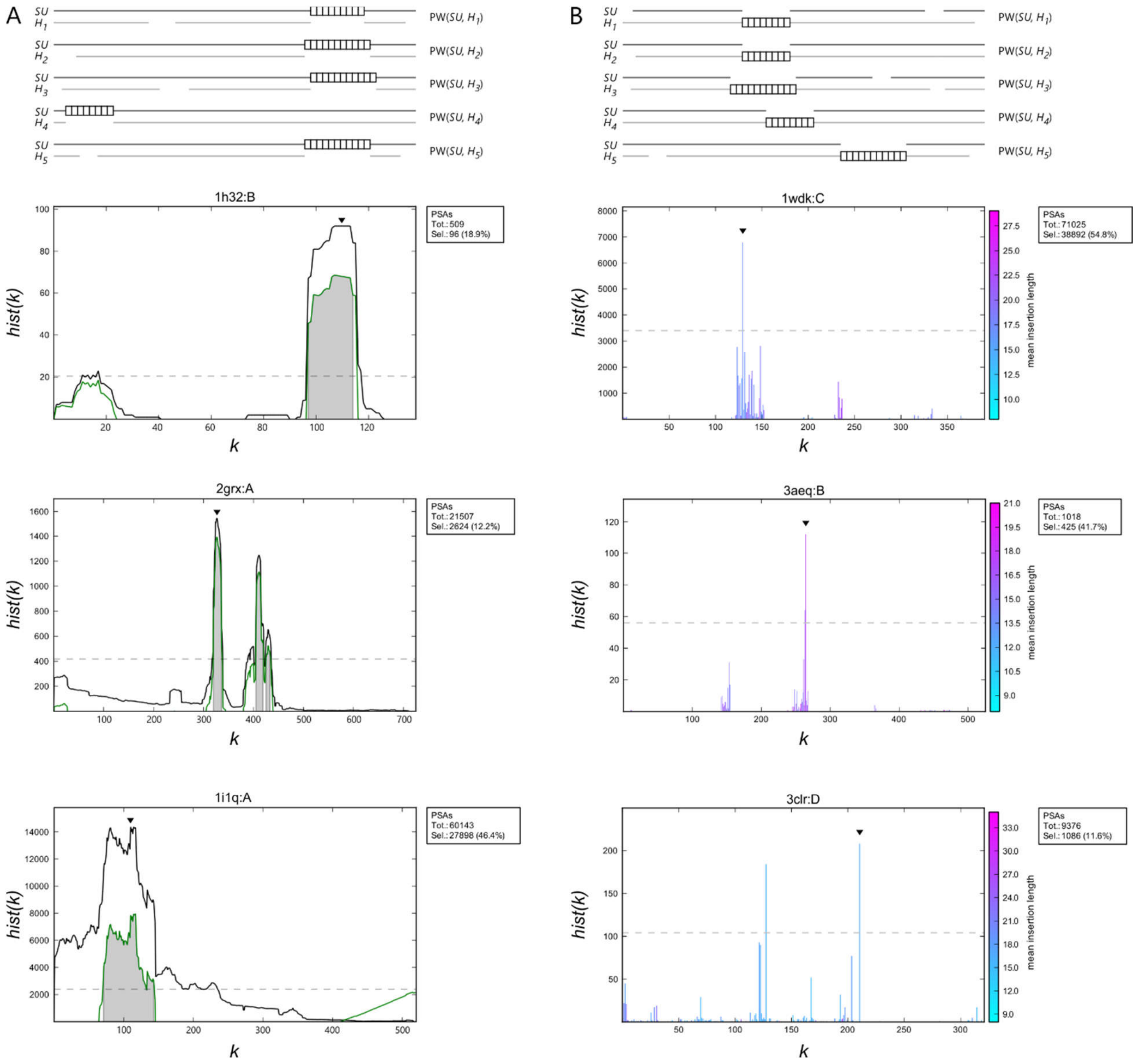


Figure S1. Related to Figure 1.

Example histograms resulting from the mapping of insertions in PW(*SU*, *H*).

A) Exemplary histograms used to identify insertions in subunits *SU* of the reference complexes with PDB IDs 1i1q (chain A), 1h32 (chain B), and 2grx (chain A). Above the first histogram, a schematic representation of five alignments PW(*SU*, $H_{1\text{-}5}$) with insertions (boxes) in *SU* is shown. The number of total alignments computed (Tot.) as well as the number and proportion of alignments selected for generating the histograms (Sel.) is given on the right of each panel. *hist*(*k*) specifies for each residue position *k* of *SU* how often it is part of an insertion with a length of at least eight residues. Black lines represent raw counts; green lines represent data corrected for potential noise. Note that corrected data can sometimes contain artifacts (compare panel 1i1q:A) that were not considered for further evaluation. The dashed horizontal line represents the detection cut-off of 0.2 · *max_hist*. **B)** Exemplary histograms used to identify insertion in InterPro homologs *H* of the subunits *SU* of the reference complexes with PDB IDs 3clr (chain D), 1wdk (chain C), and 3aeq (chain B). Above the first histogram, a schematic representation of five alignments PW(*SU*, $H_{1\text{-}5}$) with insertions (boxes) in H_1-H_5 is shown. The histograms were generated analogously to those described in panel A with the exception that *hist*(*k*) specifies for each residue position *k* of *SU* how often it is the start of an insertion with a length of at least eight residues in *H*. The mean length of the insertions is given by the cyan-magenta color gradient. The dashed horizontal line represents the detection cut-off of 0.5 · *max_hist*.

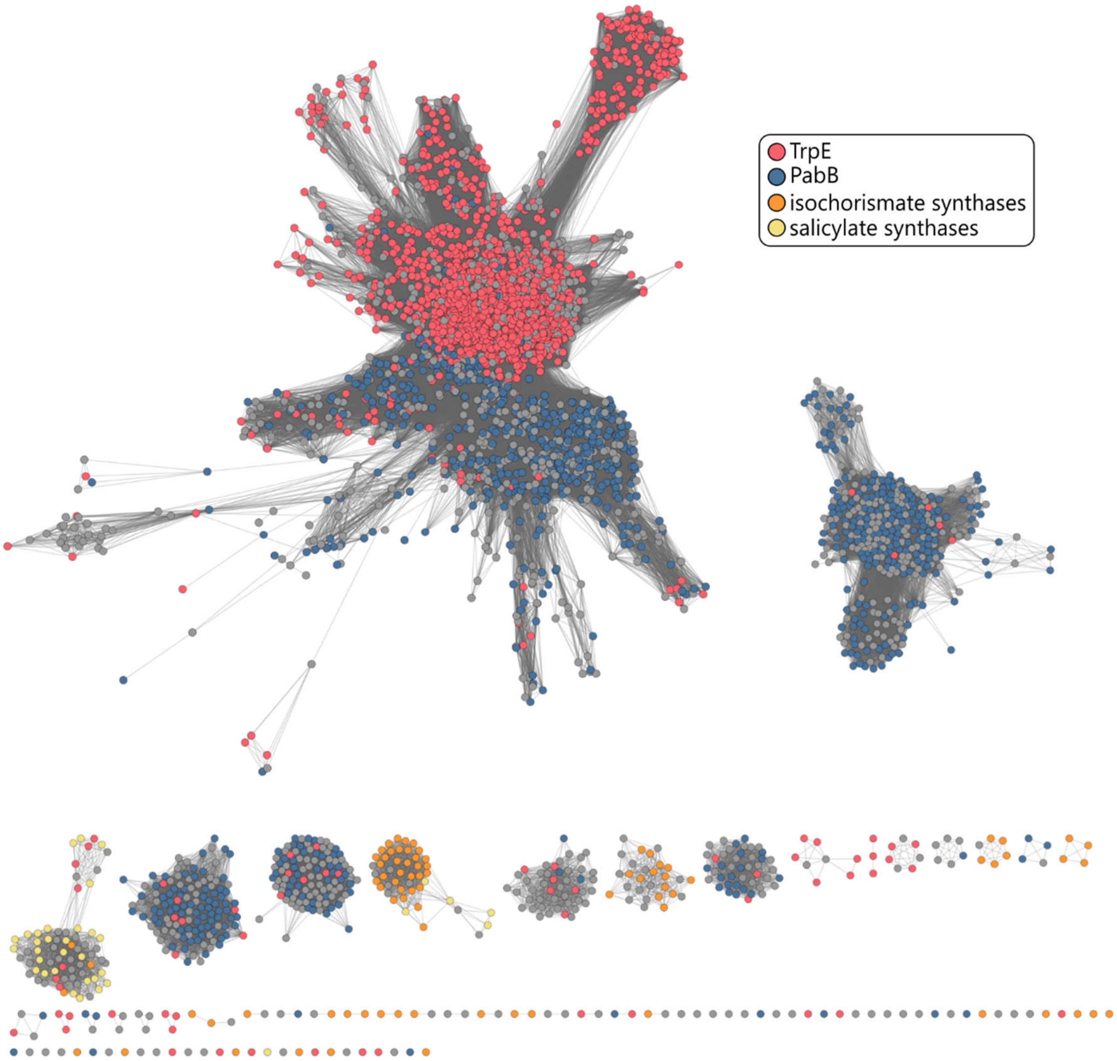


Figure S2. Related to Figure 3.

Sequence similarity network of the "anthranilate synthase component I-like" InterPro family (IPR019999) generated with an E-value cut-off of 1E-77.

Nodes are colored according to the functional annotation of InterPro. Grey nodes represent sequences with ambiguous annotation. Nodes represent a single sequence or groups of sequences with greater than 75% median sequence identity. Edges correspond to bi-directional BLAST hits with E-values lower than 1E-77 (median sequence identities greater than 37%). In addition to TrpE and PabB, this InterPro family includes few sequences that represent isochorismate synthases and salicylate synthases; both homologs of TrpE that use water instead of ammonia as a nucleophile and that are part of secondary metabolic biosynthesis pathways of iron-chelating siderophores and menaquinone (He et al., 2004; Plach et al., 2015)

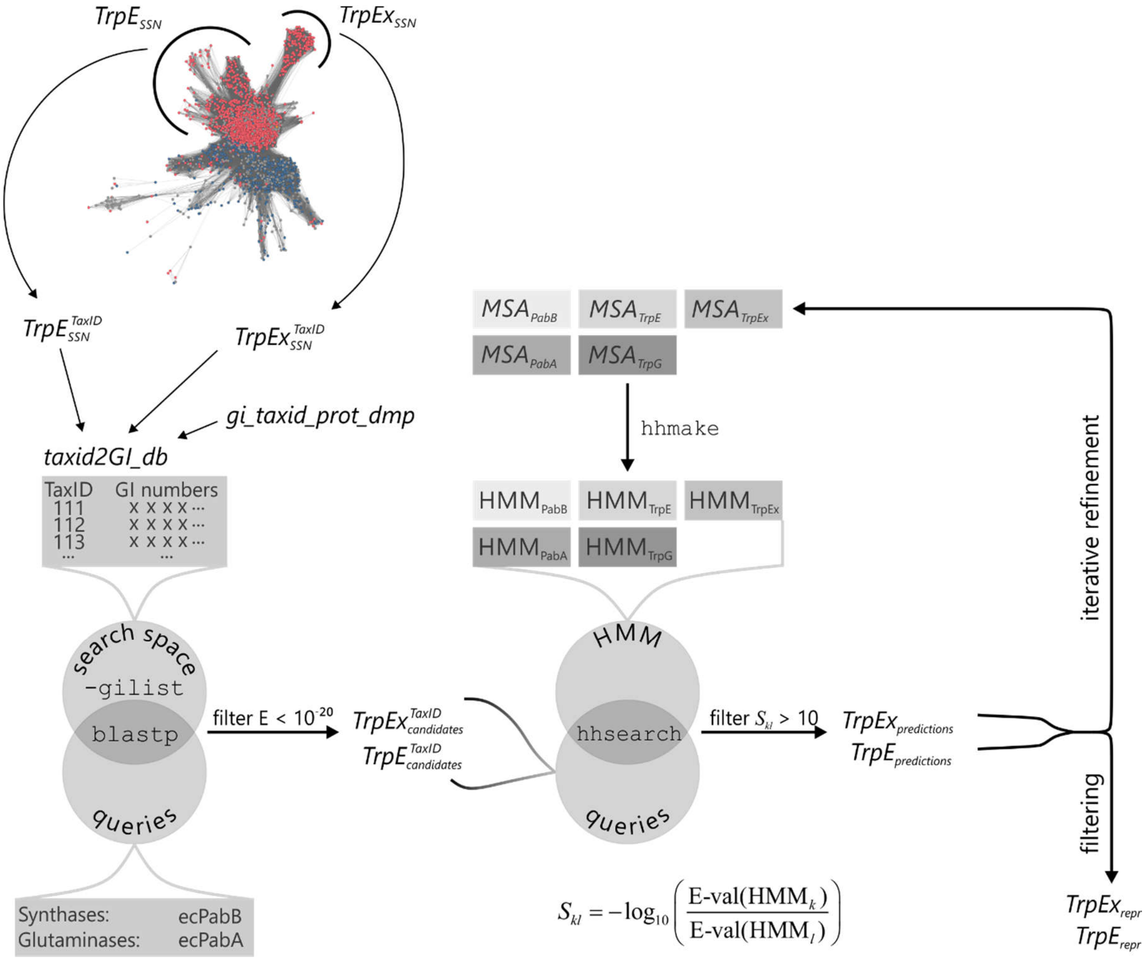


Figure S3. Related to Figure 4 and Experimental Procedures.

Schematic description of the computational routine used to determine phylogenetic distributions of the genes coding for TrpEx, TrpE, TrpG, PabB, and PabA.

The species that constitute $TrpEx_{SSN}$ and $TrpE_{SSN}$ were extracted from the SSN and stored in $TrpEx^{TaxID}_{SSN}$ and $TrpE^{TaxID}_{SSN}$ in the form of taxonomy identifiers (TaxIDs). For each species, an individual BLAST-search space was generated by linking one TaxID with all NCBI-listed, corresponding protein-related GI numbers, which were retrieved from the NCBI-provided *gi_taxid_prot_dmp* database. This information was stored in the local database *taxid2GI_db*. BLAST searches were conducted with `blastp`, using PabB and PabA from *E. coli* as query sequences for synthases and glutaminases, respectively. Default search parameters were used, except the argument `-gilist`, which was used to limit the search space to the species-wise TaxID-to-GI mappings stored in *taxid2GI_db*. TaxID-specific BLAST-hits with E-values lower than 1E-20 were stored in $TrpEx^{TaxID}_{candidates}$ and $TrpE^{TaxID}_{candidates}$. The sequences were annotated as either TrpEx, TrpE, PabB, TrpG, or PabA by comparison with Hidden-Markov-Models (HMMs) using `hhsearch` with default parameters. The five HMMs were parametrized with MSAs composed of verified TrpEx, TrpE, PabB, TrpG, and PabA sequences. The selectivity of the assignments was described by log-odds ratio scores S_{kl}. Assignments with scores $S_{kl} < 10$ were rejected, leading to $TrpEx_{predictions}$ and $TrpE_{predictions}$. The computational routine includes the option for iterative refinement of group-specific HMMs. Filtering out sequencing-bias and duplicate entries gave the final datasets $TrpEx_{repr}$ and $TrpE_{repr}$.

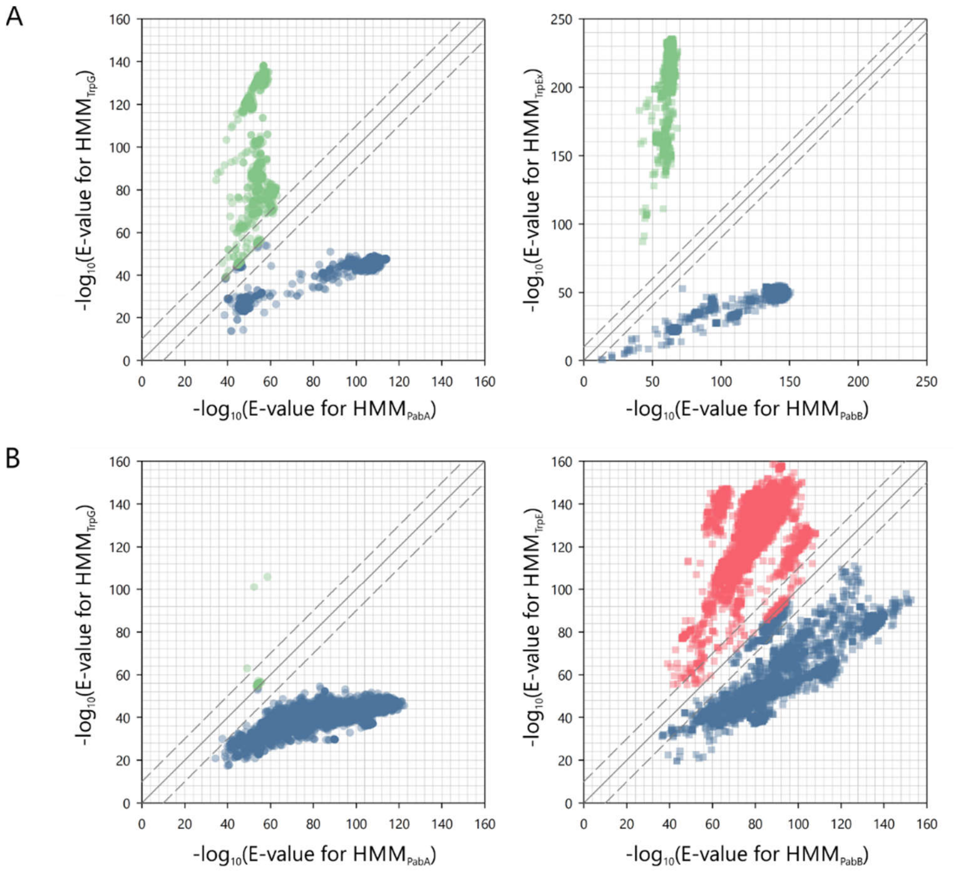


Figure S4. Related to Figure 4.

Distribution of E-values in sequence-to-HMM comparisons for species in *TrpEx*$_{SSN}$ (A) and *TrpE*$_{SSN}$ (B).

Sequences retrieved from BLAST-searches with the glutaminase PabA from *E. coli* as a query were compared to HMMs representing PabA and TrpG (left panels). Sequences retrieved from BLAST-searches with the synthase PabB from *E. coli* as a query were compared to HMMs representing TrpEx and PabB or TrpE and PabB, respectively (right panels). The dashed diagonals indicate the area of assignments with scores $S_{kl} < 10$. These assignments were rejected.

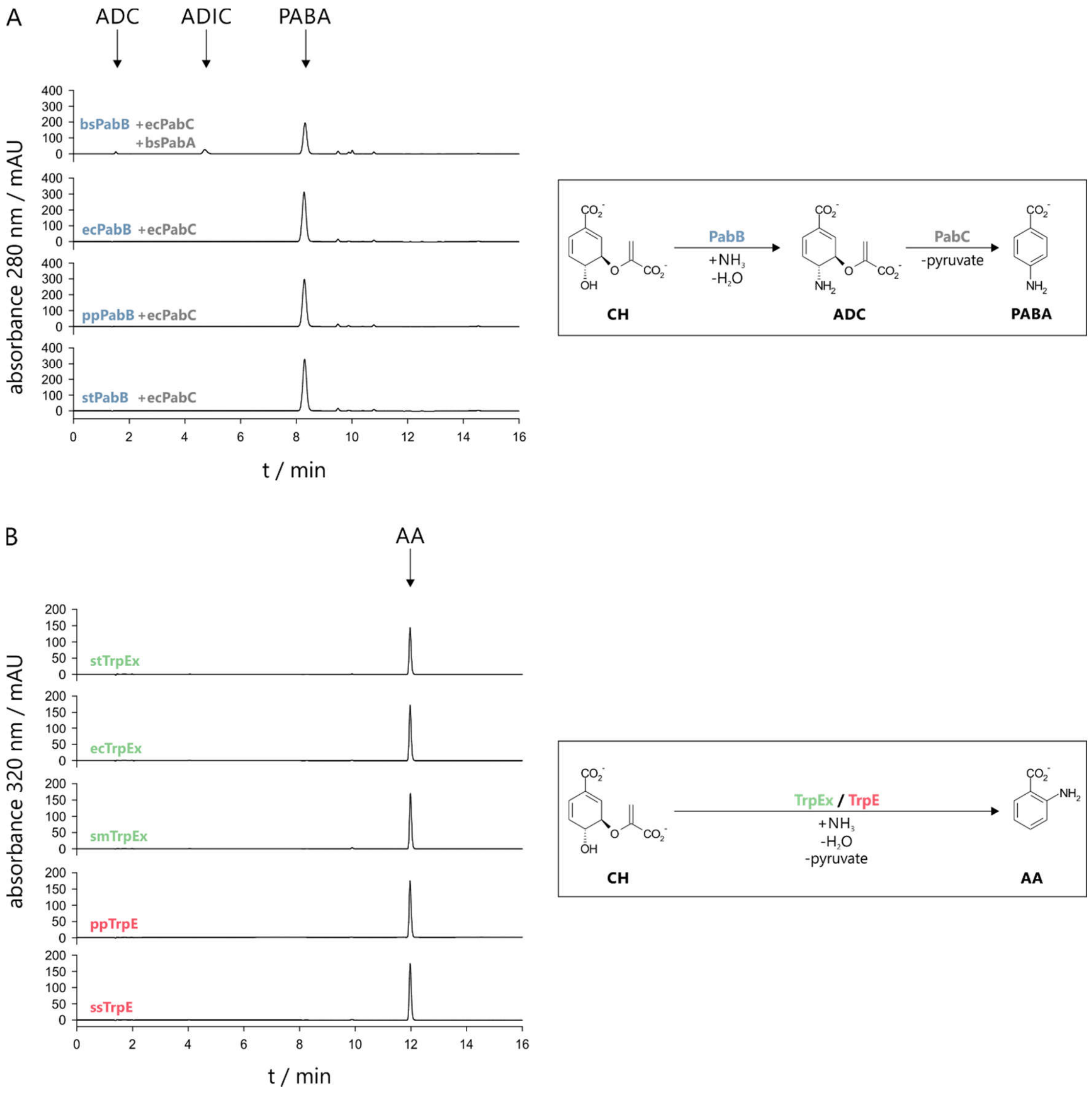


Figure S5. Related to Results.

HPLC chromatograms of the reaction mixtures of PabB (A) and TrpEx/TrpE (B) with chorismate (CH) and ammonium chloride.

Reaction schemes are given on the right. The aminodeoxychorismate-lyase PabC from *E. coli* was added to the reaction mixtures of the PabB enzymes to convert the PabB product aminodeoxychorismate (ADC) to p-aminobenzoate (PABA), which could be identified by comparison with a chemical standard. Notably, bsPabB was only active when supplemented with its associated glutaminase bsPabA. Aminodeoxyisochorismate (ADIC) is an intermediate specific for the bsPabB-catalyzed conversion of chorismate to ADC and was identified based on previously reported HPLC analyses of chorismate derivatives (He et al., 2004; He and Toney, 2006). Anthranilate (AA) was identified by comparison with a chemical standard.

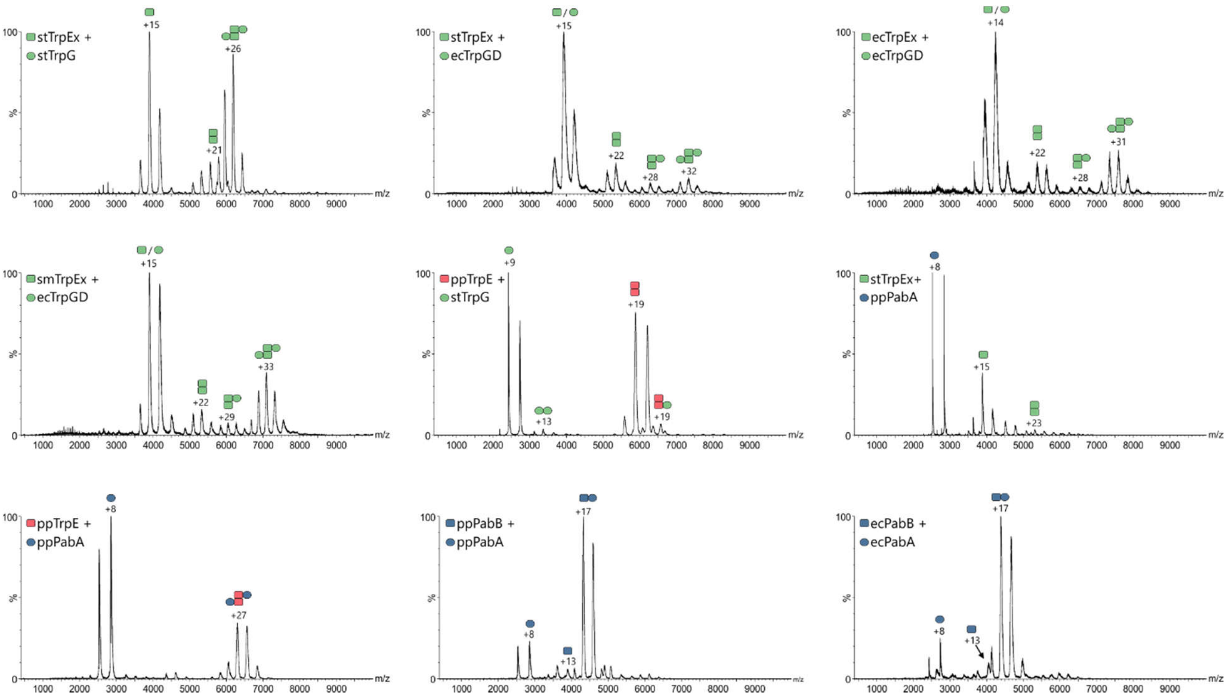


Figure S6. Related to Figure 5A.

Mass spectrometric characterization of interactions between selected TrpEx, TrpE, and PabB homologs and selected TrpG and PabA glutaminases.

For each combination, a representative mass spectrum and pictograms of complexes and sub-complexes are shown. Charges of the most populated charge species are included. The assembly of the TrpEx$_2$:TrpG$_2$ and ppTrpE$_2$:ppPabA$_2$ tetramers most likely proceeds via dimeric and trimeric intermediates. As it was shown that biophysically characterized assembly pathways strongly reflect evolutionary histories of protein complexes (Levy et al., 2008; Marsh et al., 2013), we assume that these intermediates are also physiologically relevant.

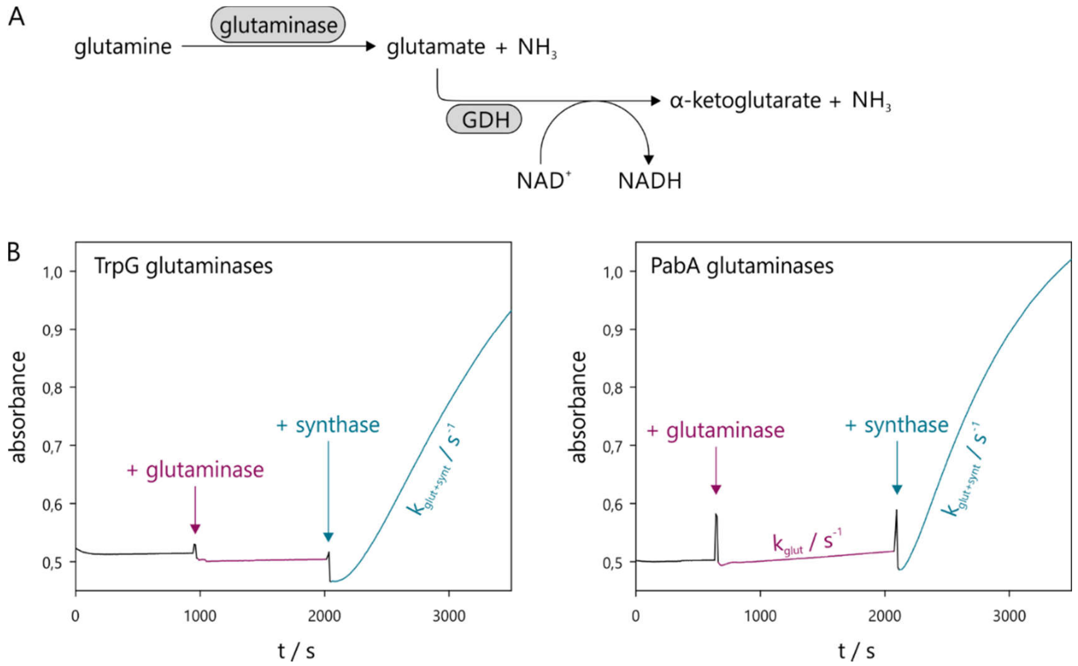


Figure S7. Related to Figure 5C.

Glutaminase stimulation assay.

A) Schematic representation of the steady-state glutaminase activity assay. Glutamine is hydrolyzed to glutamate and ammonia by the glutaminases. Glutamate, in turn, is deaminated to α-ketoglutarate by glutamate dehydrogenase (GDH) with concomitant reduction of NAD$^+$ to NADH. **B)** Schematic representation of progress curves for the assay with TrpG glutaminases (left) and PabA glutaminases (right). After preincubation, glutaminases are added to the assay mixture and the progress curve is recorded for at least 15 minutes. Then, synthases are added and the progress curve is again monitored for at least 15 minutes. The stimulation factor for a pair of synthase and glutaminase is calculated by dividing the apparent turnover rates of the glutaminase-synthase complex ($k_{glut+synt}$) with that of the glutaminase alone (k_{glut}). TrpG glutaminases display no basal glutaminase activity (only $k_{glut+synt}$ available). Therefore, no stimulation factor can be calculated.

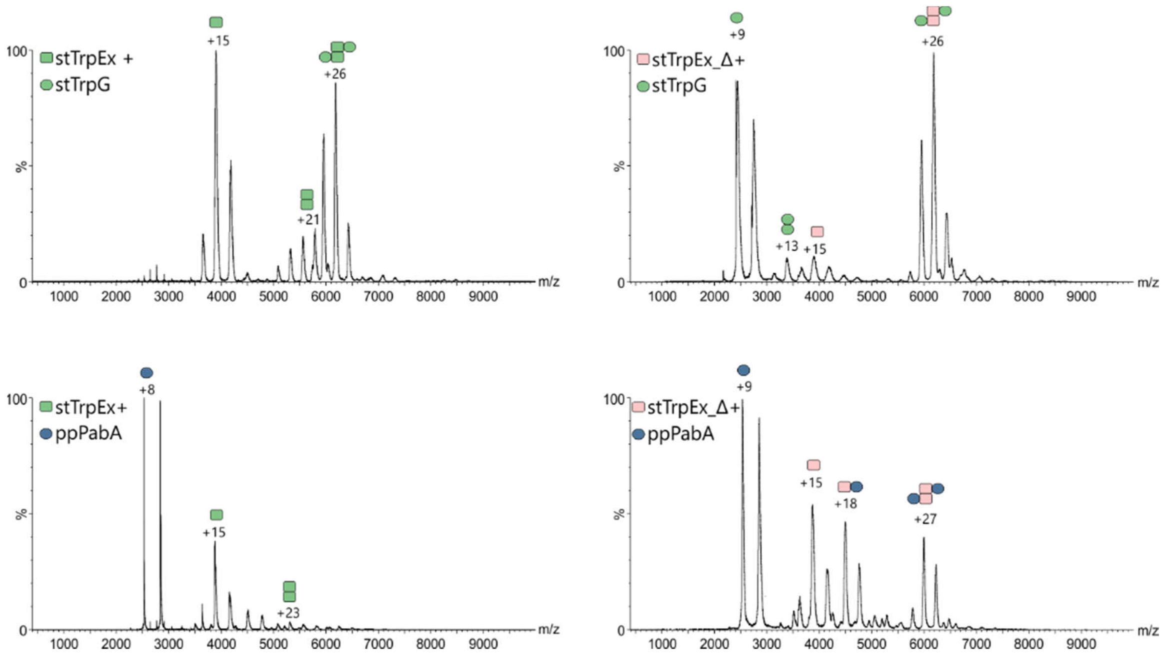


Figure S8. Related to Figure 6 and Table 1.

Mass spectrometric characterization of interactions between stTrpEx/stTrpEx_Δ synthases and stTrpG/ppPabA glutaminases.

For each combination, a representative spectrum and pictograms of complexes and sub-complexes are shown. Charges of the most populated charge species are included. Synthases and glutaminases were mixed equimolarily prior to analysis (20 µM each). Apparent molecular weights are provided in **Table S9**.

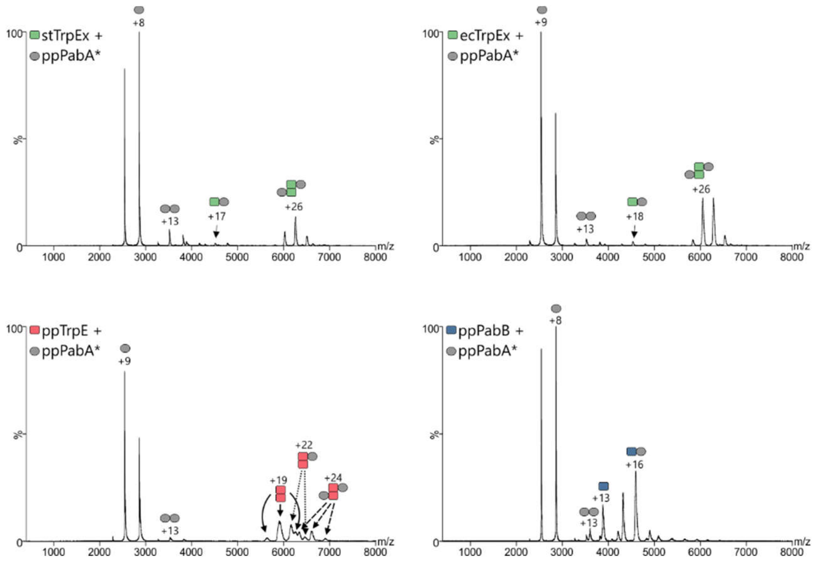


Figure S9. Related to Table 2.

Mass spectrometric characterization of interactions between ppPabA* and stTrpEx, ecTrpEx, smTrpEx, ppTrpE, and ppPabB.

For each combination, a representative spectrum and pictograms of complexes and sub-complexes are shown. Charges of the most populated charge species are included. 30 µM ppPabA* were mixed with 20 µM of the other proteins prior to analysis. Apparent molecular weights are provided in **Table S10**.

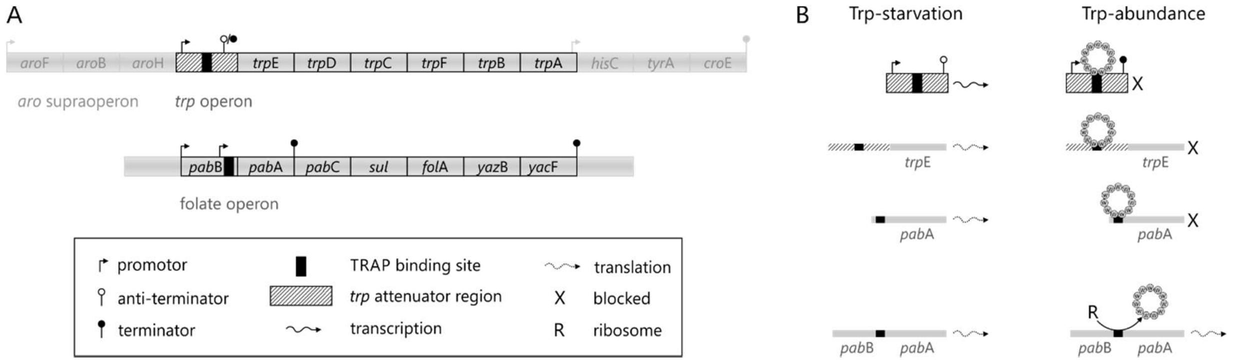


Figure S10. Related to Results and Figure 8.

Differential regulation of tryptophan and folate biosynthesis in *B. subtilis*.

A) Organization of the tryptophan and folate biosynthetic genes within the *trp*- and folate-operons. The genes coding for the components of the tryptophan biosynthetic machinery are grouped in the *trp*-operon, which itself is part of the larger *aro*-supraoperon (Henner et al., 1985). Notably, the *trp*-operon does not contain a gene for a glutaminase; instead the single PabA glutaminase of *B. subtilis* is encoded together with PabB in the folate operon that has an unusual two-promoter structure (Slock et al., 1990; Yakhnin et al., 2007). **B)** Mechanisms for the differential regulation of tryptophan and folate in *B. subtilis* and their response to tryptophan starvation or abundance. The central player in regulating tryptophan and folate biosynthesis in *B. subtilis* is the tryptophan-sensing protein TRAP (Babitzke, 1997), which binds excessive tryptophan and exercises transcriptional attenuation and translational control both on the tryptophan and the folate operon. TRAP is represented as a wheel of 11 identical subunits. In a situation where the cell is in need for tryptophan (left-hand side), neither the transcription of the *trp*-operon, nor the translation of *trp*- and folate-operon transcripts is inhibited by TRAP. In a situation of excess tryptophan (right-hand side), activated TRAP binds to a region downstream of the *trp*-operon promotor which results in the formation of a terminator structure and blocks transcription. In addition, the binding of TRAP to the Shine-Dalgarno region of *trp*-operon transcripts sequesters this region, thereby blocking translation (Merino et al., 1995). Activated TRAP also reduces the levels of PabA, the glutaminase required for tryptophan biosynthesis, by directly blocking ribosome binding to the Shine-Dalgarno sequence of *pab*A transcripts (Babitzke et al., 1994; Du et al., 1997). If biosynthesis of folate is required while cellular levels of tryptophan are high, production of the single available glutaminase PabA required for folate biosynthesis is unfavorably blocked by TRAP. This situation is resolved by two mechanisms: First, evolutionary fine-tuning of binding affinities has led to the circumstance that the binding of TRAP to the *pab*A-transcript only reduces translation about 12-fold, whereas translation of *trp*-operon transcripts is down-regulated by about 900-fold (Yakhnin et al., 2007). Second, the additional promoter of the folate-operon, which is located upstream of *pab*B leads to transcripts that contain the coding regions for PabB and PabA (right-hand side, lower part). Translation-mediated displacement of TRAP by the ribosome now enables the synthesis of PabB and the glutaminase PabA, resulting in the full ADCS complex required for folate biosynthesis (Yakhnin et al., 2007).

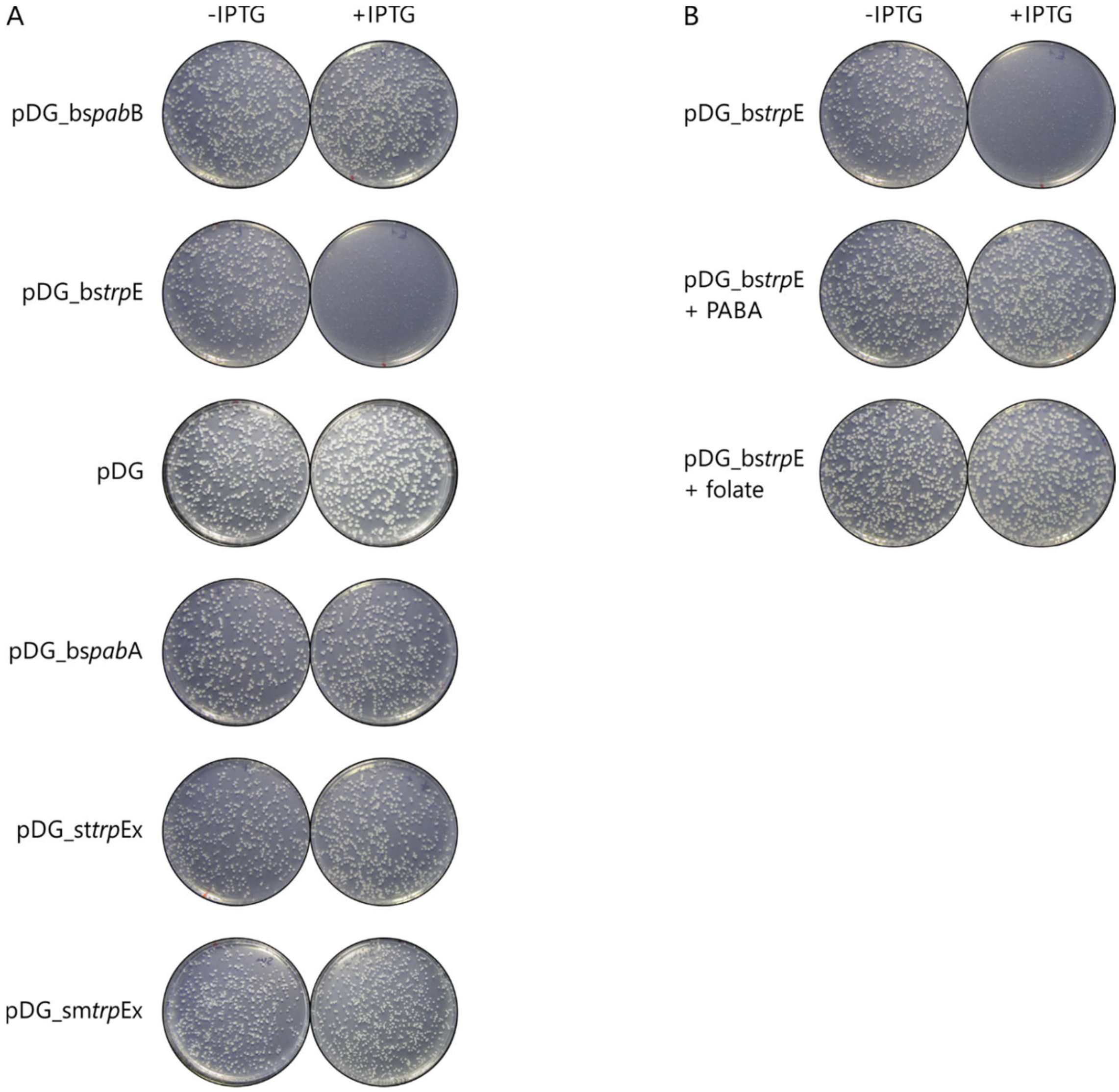


Figure S11. Related to Figure 8.

Minimal medium agar plates with colonies grown from different *B. subtilis* transformants in the absence and presence of IPTG.

B. subtilis cells were transformed with pDG148 plasmids containing the indicated genes and grown at 37 °C for 48 hours. **A)** Representative plates with colonies grown in the absence (left column) and presence (right column) of 2 mM IPTG. **B)** Representative plates showing the offsetting effect of 10 µg/mL *p*-aminobenzoic acid (PABA) and 10 µg/mL folate on bs*trp*E overexpression. In the presence of PABA or folate colonies grow to the same size as in the absence of IPTG.

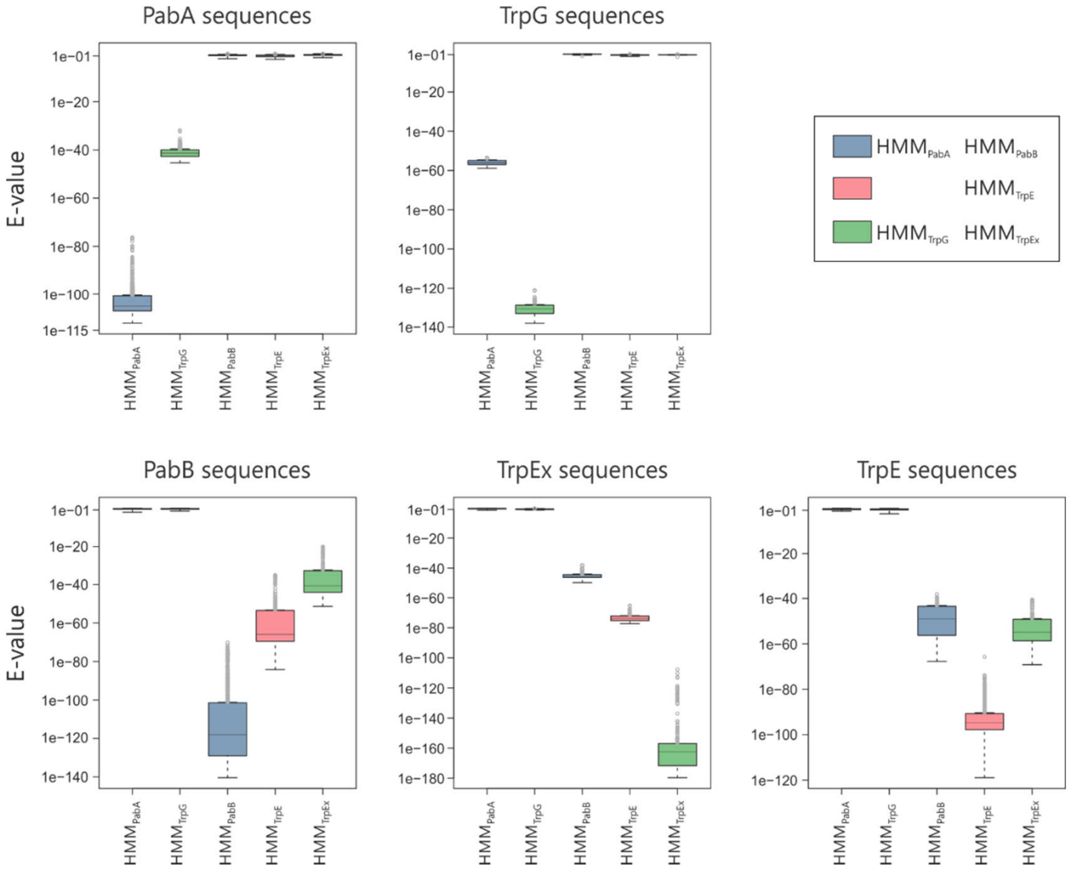


Figure S12. Related to Experimental Procedures.

Cross-validation of HMM_{enz} .

Comparison of randomly selected PabA-sequences from $\mathrm{HMM}_{\mathrm{PabA}}$ results in the lowest E-values for $\mathrm{HMM}_{\mathrm{PabA}}$ (top left panel). The distance to the next "best" HMM, in this case $\mathrm{HMM}_{\mathrm{TrpG}}$, is around 60 orders of magnitude. The same is true for the other four cross-validations, although the E-value distances between the genuine and the improper sequence-to-HMM comparisons are somewhat lower for the cases of PabB, TrpEx, and TrpE. This is the result of higher sequence identities between these enzymes compared to the sequences identities between PabA and TrpG.

Supplemental References

Ashkenazy, H., Erez, E., Martz, E., Pupko, T., and Ben-Tal, N. (2010). ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. Nucleic Acids Res. *38*, W529-533.

Atkinson, H.J., Morris, J.H., Ferrin, T.E., and Babbitt, P.C. (2009). Using sequence similarity networks for visualization of relationships across diverse protein superfamilies. PLoS One *4*, e4345.

Babitzke, P. (1997). Regulation of tryptophan biosynthesis: Trp-ing the TRAP or how *Bacillus subtilis* reinvented the wheel. Mol. Microbiol. *26*, 1-9.

Babitzke, P., Stults, J.T., Shire, S.J., and Yanofsky, C. (1994). TRAP, the trp RNA-binding attenuation protein of *Bacillus subtilis*, is a multisubunit complex that appears to recognize G/UAG repeats in the *trp*EDCFBA and *trp*G transcripts. J. Biol. Chem. *269*, 16597-16604.

Bamford, V.A., Bruno, S., Rasmussen, T., Appia-Ayme, C., Cheesman, M.R., Berks, B.C., and Hemmings, A.M. (2002). Structural basis for the oxidation of thiosulfate by a sulfur cycle enzyme. EMBO J. *21*, 5599-5610.

Ciccarelli, F.D., Doerks, T., von Mering, C., Creevey, C.J., Snel, B., and Bork, P. (2006). Toward automatic reconstruction of a highly resolved tree of life. Science *311*, 1283-1287.

Dean, A.M., and Koshland, D.E., Jr. (1993). Kinetic mechanism of *Escherichia coli* isocitrate dehydrogenase. Biochemistry *32*, 9302-9309.

Du, H., Tarpey, R., and Babitzke, P. (1997). The trp RNA-binding attenuation protein regulates TrpG synthesis by binding to the *trp*G ribosome binding site of *Bacillus subtilis*. J. Bacteriol. *179*, 2582-2586.

Evdokimov, A.G., Mekel, M., Hutchings, K., Narasimhan, L., Holler, T., McGrath, T., Beattie, B., Fauman, E., Yan, C., Heaslet, H.*, et al.* (2008). Rational protein engineering in action: the first crystal structure of a phenylalanine tRNA synthetase from *Staphylococcus haemolyticus*. J. Struct. Biol. *162*, 152-169.

Friedrich, C.G., Rother, D., Bardischewsky, F., Quentmeier, A., and Fischer, J. (2001). Oxidation of reduced inorganic sulfur compounds by bacteria: emergence of a common mechanism? Appl. Environ. Microbiol. *67*, 2873-2882.

Gerlt, J.A., Bouvier, J.T., Davidson, D.B., Imker, H.J., Sadkhin, B., Slater, D.R., and Whalen, K.L. (2015). Enzyme Function Initiative-Enzyme Similarity Tool (EFI-EST): A web tool for generating protein sequence similarity networks. Biochim. Biophys. Acta *1854*, 1019-1037.

Gille, C., and Frommel, C. (2001). STRAP: editor for STRuctural Alignments of Proteins. Bioinformatics *17*, 377-378.

Golden, E., Paterson, R., Tie, W.J., Anandan, A., Flematti, G., Molla, G., Rosini, E., Pollegioni, L., and Vrielink, A. (2013). Structure of a class III engineered cephalosporin acylase: comparisons with class I acylase and implications for differences in substrate specificity and catalytic activity. The Biochemical journal *451*, 217-226.

Haapalainen, A.M., Merilainen, G., and Wierenga, R.K. (2006). The thiolase superfamily: condensing enzymes with diverse reaction specificities. Trends Biochem. Sci. *31*, 64-71.

Harijan, R.K., Kiema, T.R., Karjalainen, M.P., Janardan, N., Murthy, M.R., Weiss, M.S., Michels, P.A., and Wierenga, R.K. (2013). Crystal structures of SCP2-thiolases of Trypanosomatidae, human pathogens causing widespread tropical diseases: the importance for catalysis of the cysteine of the unique HDCF loop. Biochem. J. *455*, 119-130.

He, Z., Stigers Lavoie, K.D., Bartlett, P.A., and Toney, M.D. (2004). Conservation of mechanism in three chorismate-utilizing enzymes. J. Am. Chem. Soc. *126*, 2378-2385.

He, Z., and Toney, M.D. (2006). Direct detection and kinetic analysis of covalent intermediate formation in the 4-amino-4-deoxychorismate synthase catalyzed reaction. Biochemistry *45*, 5019-5028.

Henner, D.J., Band, L., and Shimotsu, H. (1985). Nucleotide sequence of the *Bacillus subtilis* tryptophan operon. Gene *34*, 169-177.

Ishikawa, M., Tsuchiya, D., Oyama, T., Tsunaka, Y., and Morikawa, K. (2004). Structural basis for channelling mechanism of a fatty acid beta-oxidation multienzyme complex. EMBO J. *23*, 2745-2754.

Joseph, P., Fantino, J.R., Herbaud, M.L., and Denizot, F. (2001). Rapid orientated cloning in a shuttle vector allowing modulated gene expression in *Bacillus subtilis*. FEMS Microbiol. Lett. *205*, 91-97.

Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. *30*, 772-780.

Kilmartin, J.R., Maher, M.J., Krusong, K., Noble, C.J., Hanson, G.R., Bernhardt, P.V., Riley, M.J., and Kappler, U. (2011). Insights into structure and function of the active site of SoxAX cytochromes. J. Biol. Chem. *286*, 24872-24881.

Kim, D.W., Kang, S.M., and Yoon, K.H. (1999). Isolation of Novel Pseudomonas diminuta KAC-1 Strain Producing Glutaryl 7-Aminocephalosporanic Acid Acylase. The Journal of Microbiology *37*, 200-205.

Kim, S., and Park, S. (2013). Structural changes during cysteine desulfurase CsdA and sulfur acceptor CsdE interactions provide insight into the trans-persulfuration. J. Biol. Chem. *288*, 27172-27180.

LaPorte, D.C. (1993). The isocitrate dehydrogenase phosphorylation cycle: regulation and enzymology. J. Cell. Biochem. *51*, 14-18.

LaPorte, D.C., and Koshland, D.E., Jr. (1982). A protein with kinase and phosphatase activities involved in regulation of tricarboxylic acid cycle. Nature *300*, 458-460.

Letunic, I., and Bork, P. (2007). Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics *23*, 127-128.

Levy, E.D., Boeri Erba, E., Robinson, C.V., and Teichmann, S.A. (2008). Assembly reflects evolution of protein complexes. Nature *453*, 1262-1265.

Loiseau, L., Ollagnier-de Choudens, S., Lascoux, D., Forest, E., Fontecave, M., and Barras, F. (2005). Analysis of the heteromeric CsdA-CsdE cysteine desulfurase, assisting Fe-S cluster biogenesis in *Escherichia coli*. J. Biol. Chem. *280*, 26760-26769.

Marco-Marin, C., Gil-Ortiz, F., Perez-Arellano, I., Cervera, J., Fita, I., and Rubio, V. (2007). A novel two-domain architecture within the amino acid kinase enzyme family revealed by the crystal structure of *Escherichia coli* glutamate 5-kinase. J. Mol. Biol. *367*, 1431-1446.

Marsh, J.A., Hernandez, H., Hall, Z., Ahnert, S.E., Perica, T., Robinson, C.V., and Teichmann, S.A. (2013). Protein complexes are under evolutionary selection to assemble via ordered pathways. Cell *153*, 461-470.

Meinnel, T., Mechulam, Y., and Blanquet, S. (1995). Aminoacyl-tRNA synthetases: occurrence, structure, and function. In tRNA, D. Söll, and U.L. RajBhandary, eds. (Washington, DC: ASM Press), pp. 251-292.

Merino, E., Babitzke, P., and Yanofsky, C. (1995). trp RNA-binding attenuation protein (TRAP)-trp leader RNA interactions mediate translational as well as transcriptional regulation of the *Bacillus subtilis* trp operon. J. Bacteriol. *177*, 6362-6370.

Mihara, H., and Esaki, N. (2002). Bacterial cysteine desulfurases: their function and mechanisms. Appl. Microbiol. Biotechnol. *60*, 12-23.

Miyazaki, J., Kobashi, N., Nishiyama, M., and Yamane, H. (2001). Functional and evolutionary relationship between arginine biosynthesis and prokaryotic lysine biosynthesis through alpha-aminoadipate. J. Bacteriol. *183*, 5067-5073.

Parsons, S.J., and Burns, R.O. (1969). Purification and properties of beta-isopropylmalate dehydrogenase. J. Biol. Chem. *244*, 996-1003.

Plach, M.G., Löffler, P., Merkl, R., and Sterner, R. (2015). Conversion of anthranilate synthase into isochorismate synthase: implications for the evolution of chorismate-utilizing enzymes. Angew. Chem. Int. Ed. *54*, 11270-11274.

Pürzer, A., Grassmann, F., Birzer, D., and Merkl, R. (2011). Key2Ann: a tool to process sequence sets by replacing database identifiers with a human-readable annotation. J. Integr. Bioinform. *8*, 153.

Schneider, C.A., Rasband, W.S., and Eliceiri, K.W. (2012). NIH Image to ImageJ: 25 years of image analysis. Nat. Methods *9*, 671-675.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. *13*, 2498-2504.

Singh, S.K., Miller, S.P., Dean, A., Banaszak, L.J., and LaPorte, D.C. (2002). *Bacillus subtilis* isocitrate dehydrogenase. A substrate analogue for *Escherichia coli* isocitrate dehydrogenase kinase/phosphatase. J. Biol. Chem. *277*, 7567-7573.

Slock, J., Stahly, D.P., Han, C.Y., Six, E.W., and Crawford, I.P. (1990). An apparent *Bacillus subtilis* folic acid biosynthetic operon containing *pab*, an amphibolic *trp*G gene, a third gene required for synthesis of para-aminobenzoic acid, and the dihydropteroate synthase gene. J. Bacteriol. *172*, 7211-7226.

Smoot, M.E., Ono, K., Ruscheinski, J., Wang, P.L., and Ideker, T. (2011). Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics *27*, 431-432.

Söding, J. (2005). Protein homology detection by HMM-HMM comparison. Bioinformatics *21*, 951-960.

Vinekar, R., Verma, C., and Ghosh, I. (2012). Functional relevance of dynamic properties of Dimeric NADP-dependent Isocitrate Dehydrogenases. BMC Bioinformatics *13 Suppl 17*, S2.

Wang, W., and Malcolm, B.A. (1999). Two-stage PCR protocol allowing introduction of multiple mutations, deletions and insertions using QuikChange Site-Directed Mutagenesis. BioTechniques *26*, 680-682.

Wegkamp, A., Teusink, B., de Vos, W.M., and Smid, E.J. (2010). Development of a minimal growth medium for Lactobacillus plantarum. Lett. Appl. Microbiol. *50*, 57-64.

Yakhnin, H., Yakhnin, A.V., and Babitzke, P. (2007). Translation control of *trp*G from transcripts originating from the folate operon promoter of *Bacillus subtilis* is influenced by translation-mediated displacement of bound TRAP, while translation control of transcripts originating from a newly identified *trp*G promoter is not. J. Bacteriol. *189*, 872-879.

Yates, S.P., Edwards, T.E., Bryan, C.M., Stein, A.J., Van Voorhis, W.C., Myler, P.J., Stewart, L.J., Zheng, J., and Jia, Z. (2011). Structural basis of the substrate specificity of bifunctional isocitrate dehydrogenase kinase/phosphatase. Biochemistry *50*, 8103-8106.

Yoshida, A., Tomita, T., Fujimura, T., Nishiyama, C., Kuzuyama, T., and Nishiyama, M. (2015). Structural insight into amino group-carrier protein-mediated lysine biosynthesis: crystal structure of the LysZ.LysW complex from *Thermus thermophilus*. J. Biol. Chem. *290*, 435-447.

Zalkin, H. (1973). Anthranilate synthetase. Adv. Enzymol. Relat. Areas Mol. Biol. *38*, 1-39.

Zhang, Y., and Skolnick, J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res. *33*, 2302-2309.

List of figures

List of tables

Acknowledgements

First and foremost, I would like to thank my supervisor Prof. Dr. Reinhard Sterner for his valuable support, guidance, and advice throughout my work. I deeply appreciate his encouragement to do research in a self-sufficient way. He always had a sympathetic ear and did everything to provide an excellent work atmosphere on the professional as well as on the personal level.

I am no less grateful to Prof. Dr. Rainer Merkl. I profoundly appreciate the plenty of discussions and conversations on all the different topics of my work that helped to shape the stories of this thesis and my publications. He sparked my interest in computational science and coding and I highly value the opportunity he gave me to participate in a variety of bioinformatic research.

I would like to thank Prof. Dr. Christine Ziegler for giving me access to key laboratory equipment. In addition, I am grateful to Prof. Dr. Ralf Wagner, who mentored my work in the context of the Regensburg International Graduate School of Life Sciences.

Financial support by the *Fonds der Chemischen Industrie* is gratefully acknowledged.

I sincerely thank my collaboration partners for their valuable contributions to this work: Patrick Löffler, who performed molecular dynamics simulations for many different projects and for giving plenty of computational advice; Dr. Bernd Reisinger for the joint work on the ancestral HisA variants and the many useful discussions on all kinds of experimental and scientific questions; Prof. John Gerlt and his staff from the *Enzyme Function Initiative* for always being ready to help during the computation of sequence similarity networks; Florian Busch for the mass spectromeric characterization of proteins essential for publication C and his supervisor Prof. Vicky Wysocki for making available state-of-the-art mass spectrometers; Markus Busch for his perseverance and commitment in the development of the computational tools for genomic profiling; Leonhard Heizinger for his integral work on the computational survey for interface add-ons; and last but not least Kristina Heyn for her help with all kinds of phylogenetic questions.

A huge thank you goes to Christiane Endres, Sonja Fuchs, and Jeannette Ueckert for their invaluable technical assistance and support in the lab and for making every days work always easy and fun. I would also like to thank Klaus-Jürgen Tiefenbach and Claudia Pauer for IT and administrative support.

My heartfelt thanks go to all current and former members of the Sterner and Merkl groups. I enjoyed the excellent working atmosphere and the mutual support and I would like to thank all of them for the exciting and fun time I had. Thank you for the countless matches of table soccer and darts, although not enough for a 180.

A very special thank you goes to Florian Semmelmann, who has never ceased to support me with his manpower, his enthusiasm, and his friendship during the years. A big part of this thesis would not be as it is without him.

Above all, I am deeply grateful to my family for their unlimited and unconditional encouragement and support not only during these last four years. Dear Teresa, my love and my best friend: Thank you so much for putting up with me and for everything that goes beyond my work and this thesis!