Ancestral Sequence Reconstruction: Methods and Applications

Dissertation

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

vorgelegt von
Kristina Straub, geb. Heyn
aus Bad Kreuznach

Juni 2018

Das Promotionsgesuch wurde eingereicht am: 15.06.2018

Die Arbeit wurde angeleitet von: Prof. Dr. Rainer Merkl

Unterschrift: .....

Kristina Straub

Abstract

A major goal in the study of molecular evolution is to elucidate properties of ancestral proteins and to understand their adaption induced by changes in the environment. Due to the lack of macromolecular fossils, ancestral sequence reconstruction (ASR) is the only alternative to deduce sequences for evolutionary precursors of extant proteins. Within the last years, ancestral proteins were inferred spanning a time-period of more than 3 billion years. Ancestral proteins from eubacteria, archaea, yeast, and vertebrates could be reconstructed. Thus, ASR yielded insights into the early history of life and the evolution of proteins and of macromolecular complexes. Moreover, it turned out that ASR is an effiecient method of protein design, because the reconstructed sequences often possess favorable properties like an increased thermostability. The popularity and efficacy of ASR benefitted from improvements in DNA sequencing technology, the enormous rise of computer power and the refinements of algorithms for sequence and phylogenetic analyses to be seen during the last decades. Thus, elaborated ASR methods are at hand nowadays that can be applied to a variety of evolutionary problems. For an ASR application, the user has however to pick representatives from an overwhelming number of sequences, which is no trivial task. To advance ASR technology and to assist the user, the first part of this thesis focusses on the design of a standardized ASR protocol and the development of a novel filter aimed at facilitating sequence selection. In the second part, ASR is used as a method to elucidate properties of an ancestral enzyme complex and to identify protein-protein interaction hotspots.

References of Published Manuscripts

This thesis is composed of the following published or accepted manuscripts and one additional chapter, which contains unpublished data:

A **Straub, K.**, Merkl, R. (2018). Ancestral sequence reconstruction as a tool for the elucidation of a stepwise evolutionary adaptation. *In Computational Methods in Protein Evolution: Methods and Protocols*, Springer, New York. *In Press*

B Busch, F., Rajendran, C., **Heyn, K.**, Schlee, S., Merkl, R., & Sterner, R. (2016). Ancestral tryptophan synthase reveals functional sophistication of primordial enzyme complexes. *Cell chemical biology*, 23(6), 709-715.

C Holinski, A., **Heyn, K.**, Merkl, R., & 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*, 85(2), 312-321.

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

D Linde, M., **Heyn, K.**, Merkl, R., Sterner, R., & Babinger, P. (2018). Hexamerization of geranylgeranylglyceryl phosphate synthase ensures structural integrity and catalytic activity at high temperatures. *Biochemistry*, 57(16), 2335-2348.

E Kneuttinger, A.C., Winter, M., Simeth, N.A., **Heyn, K.**, Merkl, R., König, B., Sterner, R. (2018). Artificial light-regulation of an allosteric bi-enzyme complex by a photosensitive ligand. *ChemBioChem, published online*

F Plössl, K., Schmid, V., Ammon, M., **Straub, K.**, Merkl, R., Weber, B., Friedrich U. (2018). Pathomechanism of mutated and secreted retinoschisin in X-linked juvenile retinoschisis. *Submitted for Publication*

Personal Contributions

Publication A

Rainer Merkl and myself designed the protocol. Both authors wrote the manuscript and **Figure 2.1** was created by myself.

Publication B

The experiments were conducted by Florian Busch and Sandra Schlee. Chitra Rajendran performed crystallisation experiments. Rainer Merkl and myself performed ASR; I generated the figures and tables (**Figure 4.5, Figure 4.6, and Table 4.5**), analyzed 3D structures and created the corresponding pictures (**Figure 4.3, Figure 4.4, and Figure 4.9**). Florian Busch and I drafted the manuscript and I wrote the respective parts of the paper. Rainer Merkl and Reinhard Sterner supervised the research and all authors contributed to writing of the manuscript.

Publication C

The research was designed by all authors. Alexandra Holinski and I contributed equally to this publication: Biochemical experiments were performed by Alexandra Holinski and bioinformatic research was conducted by myself leading to all corresponding figures (**Figure 5.2, Figure 5.3, Figure 5.4, Figure 5.5, Figure 5.8, Table 5.2, Table 5.3, Table 5.4, and Table 5.5**). Rainer Merkl and Reinhard Sterner supervised the work; the manuscript was written by all authors.

Contents

Abstract v

References of Published Manuscripts vii

Personal Contributions ix

List of Figures xv

List of Tables xvii

1 General Introduction 1

1.1 Evolution in Biology . . . 1
1.2 Ancestral Sequence Reconstruction . . . 3
1.3 Aim and Scope of this Work . . . 7
1.4 Guide to the Following Chapters . . . 8

2 Ancestral Sequence Reconstruction as a Tool 11

Abstract . . . 11
2.1 Introduction . . . 12
2.2 Protocol . . . 14
2.2.1 Ancestral Sequence Reconstruction . . . 14
2.2.2 Identification of Specificity-determining Residues by Means of Intermediate Sequences . . . 16
2.3 Notes . . . 17

3 Sequence Selection by `FITSS4ASR` 21

3.1 Introduction . . . 21
3.2 Results . . . 23
3.2.1 Criteria Guiding Sequence Selection for ASR . . . 23
3.2.2 `FitSS4ASR`: Filtering Sequence Sets for ASR . . . 24
3.2.3 Choosing a Datasets for ASR . . . 27
3.2.4 Conventional Sequence Selection for ASR of GGGPS . . . 27
3.2.5 Sequence Selection by Means of `FitSS4ASR` for an ASR of GGGPS . . . 28

3.3 Discussion . . . 32
3.3.1 ASR Requires a Strong Phylogenetic Signal Necessitating a Rigorous Preselection of Sequences . . . 32
3.3.2 Future Directions . . . 33
3.4 Materials and Methods . . . 33
3.4.1 Conventional ASR Protocol . . . 33
3.4.2 `FitSS4ASR`, a Semi-supervised Protocol for Sequence Selection . . . 33
3.4.3 Indicators of ASR Suitability . . . 34
3.4.4 Ancestral Sequence Reconstruction . . . 35
3.5 Supplemental Figures and Tables . . . 36

4 The Ancient Nature of the Tryptophan Synthase Complex 37
Summary . . . 37
4.1 Introduction . . . 38
4.2 Results and Discussion . . . 39
4.2.1 Sequence Reconstruction of LBCA TS Subunits . . . 39
4.2.2 Stabilities of LBCA TS Subunits and Subunit Interaction . . . 40
4.2.3 Crystal Structure and Substrate Channeling of LBCA TS . . . 40
4.2.4 Impact of the β-subunit for the Catalytic Efficiency of the α-subunit . . . 42
4.2.5 Impact of the α-subunit for the Catalytic Efficiency of the β-subunit . . . 43
4.3 Significance . . . 44
4.4 Experimental Procedures . . . 45
4.4.1 Sequence Reconstruction . . . 45
4.4.2 Cloning and Expression . . . 45
4.4.3 Absorbance and Circular Dichroism (CD) Spectroscopy . . . 46
4.4.4 Differential Scanning Calorimetry (DSC) . . . 46
4.4.5 Analytical Size Exclusion Chromatography . . . 46
4.4.6 Fluorescence Titration . . . 47
4.4.7 Transient Kinetics . . . 47
4.4.8 Steady-state Kinetics . . . 47
4.4.9 Crystallization and Structure Determination . . . 48
4.5 Supplemental Figures and Tables . . . 49

5 Identification of a Protein Interface Hotspot 55
Abstract . . . 55
5.1 Introduction . . . 56
5.2 Materials and Methods . . . 59
5.2.1 Cloning and Mutagenesis of *his*F Genes . . . 59
5.2.2 Heterologous Expression and Purification of HisF Proteins and zmHisH . . . 60
5.2.3 Fluorescence Titration . . . 61

5.2.4 Far-UV CD-Spectroscopy 61
5.2.5 ASR of Intermediate Sequences 61
5.2.6 Interface Prediction 62
5.2.7 Homology Modelling 62
5.2.8 Calculating the Interaction Energy of Protein Complexes 62
5.2.9 Predicting Hotspots 62
5.3 Results 63
5.4 Discussion 68
5.5 Supplemental Figures and Tables 70

6 Comprehensive Summary, Discussion and Outlook **75**

Digital Supplemental Data **79**

Abbreviations **81**

References **85**

Acknowledgment **101**

List of Figures

1.1 Darwin's sketch of the tree of life . . . 1
1.2 Tree of life . . . 2
1.3 "Resurrection" of ancestral proteins based on ASR . . . 4
1.4 Calculation of a phylogenetic tree . . . 6

2.1 Identification of specificity-determining residue positions of the HisF:HisH interface by means of a vertical approach . . . 15

3.1 Criteria applied by `FitSS4ASR` to eliminate sequences . . . 24
3.2 Workflow of `FitSS4ASR` . . . 26
3.3 Phylogeny of the manually curated sequence set used for ASR of GGGPS predecessors . . . 29
3.4 The phylogeny of the sequence set generated by means of `FitSS4ASR` for ASR of GGGPS predecessors . . . 30

4.1 Reactions catalyzed by the α-subunit (α-reaction), the β-subunit (β-reaction), and the TS complex ($\alpha\beta$-reaction). . . . 38
4.2 Assembly of LBCA α- and β-subunits to the TS complex. . . . 41
4.3 Crystal structure of the LBCA TS complex . . . 42
4.4 Comparison of H-bonds between LBCA TS and stTS. . . . 43
4.5 Phylogenetic tree for the reconstruction of LBCA TS . . . 49
4.6 Amino acid sequences of LBCA TS subunits . . . 50
4.7 Thermal stability of LBCA α- and β-subunits . . . 50
4.8 Reaction course of two different nucleophiles at the LBCA β-subunit active site . 51
4.9 Hydrogen bond network at the α/β interfaces of LBCA TS and stTS . . . 52

5.1 Structure and reaction of the ImGP synthase (HisF:HisH complex) . . . 58
5.2 Phylogenetic tree based on 87 concatenated HisF and HisH sequences from seven phylogenetic clades . . . 63
5.3 Model of the LUCA-HisF:zmHisH complex . . . 65
5.4 Stepwise identification of a HisF hotspot for binding to zmHisH . . . 66
5.5 Identification of interface residues determining the affinity of LUCA-HisF and Anc1pa-HisF for zmHisH by means of *in silico* design . . . 66

5.6 Fluorescence titration experiments to determine dissociation constants for the interaction of zmHisH with various HisF subunits 70
5.7 Far-UV circular dichroism spectra of HisF proteins used for fluorescence titration with zmHisH 71
5.8 Phylogenetic tree used for reconstruction of ancestral HisF sequences after optimization with `FastML` 72

List of Tables

3.1 Comparing predecessors from manual and semi-automatic approach by their SeqId 31
3.2 MSA consisting of the 87 sequences of *GGGPS2_man* and reconstructed predecessors . . . 36
3.3 MSA consisting of the 61 sequences of *GGGPS2_auto* and reconstructed predecessors . . . 36
3.4 Phylogenetic tree deduced for *GGGPS2_man* . . . 36
3.5 Phylogenetic tree deduced for *GGGPS2_auto* . . . 36

4.1 Steady-state enzymatic parameters for the α-reaction of LBCA TS and ecTS . . 43
4.2 Steady-state enzymatic parameters for the β-reaction of LBCA TS and ecTS . . 44
4.4 Crystal structure of the LBCA TS: Data collection and refinement . . . 53
4.5 Multiple sequence alignment of concatenated α- and β-subunits of modern TS and sequences of LBCA α- and β-subunits . . . 53

5.1 Dissociation constants for the interaction of zmHisH with various HisF proteins . 64
5.2 Nucleotide and amino acid sequences for Anc1pa-HisF, Anc1pa-HisF*, Anc1tm-HisF, and Anc2tm-HisF . . . 73
5.3 Aligned sequences of modern HisF proteins used for phylogenetic analysis and of LUCA-HisF . . . 73
5.4 Log likelihood values and posterior probabilities of the reconstructed ancestral sequences at each position . . . 73
5.5 Hotspot prediction for HisF residues in ImGPS interfaces . . . 73

Chapter 1

General Introduction

1.1 Evolution in Biology

Since Darwin has postulated his theory of evolution (Darwin, 1859), it is generally accepted that today's living species evolved from a common origin. The diversity of life has been generated by millions of generations driven by natural selection. The idea of a common ancestor (CA) and the diversity of today's living species are best explained by a branching pattern of evolution, called an evolutionary tree. This concept is based on the principle of homology, which was defined by Darwin as the shared ancestry within a pair of structures (e. g. bones), or genes. Studying homologous structures from different animals in detail, Darwin could deduce a trend of adaptation to a specific habitat or function. Thus, Darwin was able to derive a first evolutionary tree (**Figure 1.1**) and since then a more and more sophisticated theory of evolution was developed that stimulated many fields of life science, e. g. the field of phylogenetic systematics (Hennig, 1965).

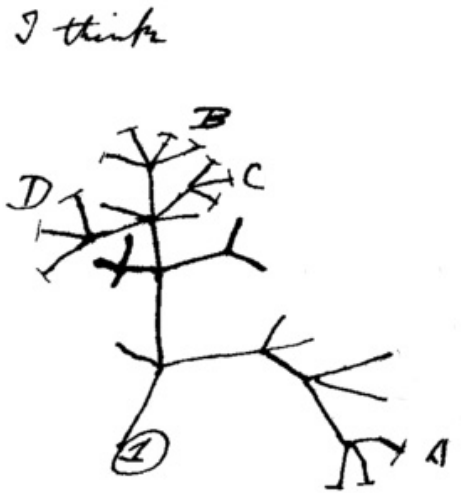

Figure 1.1: Darwin's sketch of the tree of life. A drawing from Darwin's notebook showing his first sketch of an evolutionary tree from around 1837. Adapted from Darwin (1837).

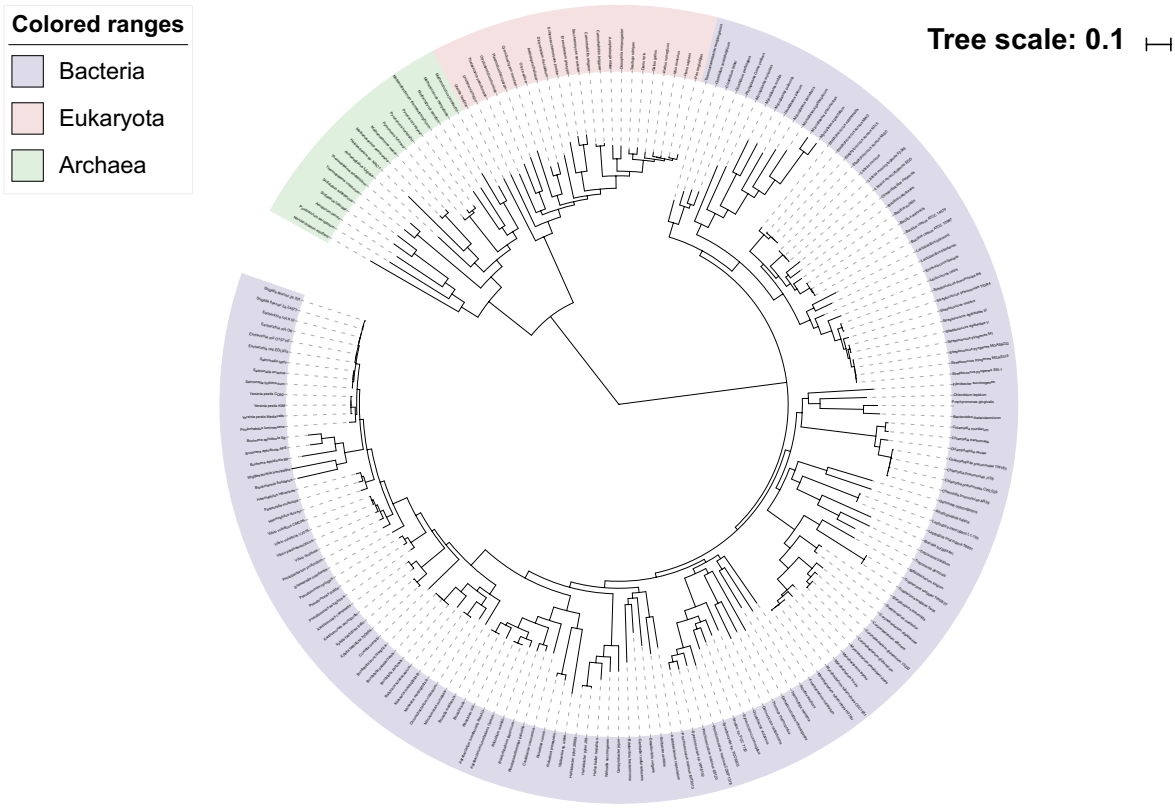

Figure 1.2: The tree of life representing the diversity of all living organisms. This tree is based on a phylogeny resulting from the analysis of 181 sequences. The tree supports the existence of three superkingdoms, namely Bacteria (blue), Eukaryota (red), and Archaea (green). Adapted from iTOL (Letunic and Bork, 2016).

Nowadays, evolution is studied on the molecular level albeit with the same concepts introduced by Darwin. With the advent of deoxyribonucleic acid (DNA) sequencing technology, genes are compared on their DNA sequences and termed homologous, if sequences share a certain level of similarity. Analogously, the homology of encoded proteins can be assessed by comparing the protein sequences (Needleman and Wunsch, 1970). Thus, the comparison of macroscopic traits like bones was replaced by the analysis of molecular features. Computational biology contributed a lot to evolutionary biology, for example with the development of phylogenetic models that describe mutational events on the level of DNA or proteins (Felsenstein, 1981). In contrast to mutations on the macroscopic level, it is uncomplicated to assess all kinds of alterations by means of probabilistic measures (Dayhoff et al., 1978). With an evolutionary model in hand, the computation of a phylogenetic tree is straightforward and can be formulated (for example) as an optimization problem. Thus, by choosing a proper set of genes or proteins, it is nowadays feasible to deduced a tree of life, which comprises representatives of all major clades that constitute the leaves (Letunic and Bork, 2016); (**Figure 1.2**). The root of the tree represents the CA according to Darwin's theory. The path from the CA represented by the root to present day organisms (outer circle) has been driven by natural selection and cannot be followed in detail

due to lacking intermediates.

However, in order to verify Darwin's theory and to understand evolution in detail, the desire to elucidate the appearance of ancestral traits has been immense. Oldest fossils date back to 635 million years ago (Gehling et al., 2000), thus the appearance of several animals like mammals or traits like feathers could be reconstructed. Unfortunately, microfossils that date back to 4.1 billion years ago (Bell et al., 2015) do not allow for the reconstruction of fragile organelles or individual macromolecules. On the other hand, Pauling and Zuckerkandl (1963) realized already in 1963 that molecules bear a signal of their history. After reliable algorithms had been designed (Felsenstein, 1981), an alternative to the analysis of fossils opened up, which is the reconstruction by means of phylogenetic methods. Nowadays, tremendous computer power is at hand and highly sophisticated sampling methods like Markov Chain Monte Carlo (MCMC) algorithms are used for Bayesian inference or maximum likelihood (ML) approaches. Thus, algorithms based on phylogenetic models are a common means for the computation of phylogenetic trees, which are subsequently used to reconstruct the sequences of extinct predecessors. Having these sequences at hand, a straightforward protocol makes it possible to express the proteins and to characterize them by means of all the experimental techniques of biochemistry and biophysics. Thus, this combination of computational and experimental biology has already been widely used (Liberles, 2007) to either test hypothesis of adaption (Frumhoff and Reeve, 1994), reconsider evolutionary relationships between the three superkingdoms (Gupta, 1998) or determine the origin of eukaryotes cell (López-García and Moreira, 2015; Eme et al., 2017). The fundamental results made it possible to understand adaptations, e. g. on climate conditions (Hoffmann and Sgrò, 2011) or interaction diversification (Plach et al., 2017) during evolution.

1.2 Ancestral Sequence Reconstruction

Since the 1980ies, novel computational methods allow the reconstruction of ancestral sequences and to travel back in time (Thornton, 2004; Hanson-Smith et al., 2010). This *in silico* technique, termed ancestral sequence reconstruction (ASR), requires four steps (Merkl and Sterner, 2016), which are depicted in (**Figure 1.3 A - G**).

Commonly, homologous sequences are retrieved from databases like UniProtKB (Apweiler et al., 2004) or with the help of `BLAST` (Altschul et al., 1990) to compile a set of extant sequences (**Figure 1.3 A**). The number of extant sequences required for an ASR depends on the protein-specific mutation rates and the time span of interest. Thus, between 11 (Yokoyama et al., 2008) and up to 200 or more sequences (Perez-Jimenez et al., 2011; Harms et al., 2013) were used for ASR. These extant sequences are then used to create a multiple sequence alignment (MSA) (**Figure 1.3 A**). During recent years, several algorithms showing comparable alignment quality have been introduced and were used to map residues to protein positions. Based on an MSA, a phylogenetic tree is deduced by means of state of the art methods like ML or with a Bayesian

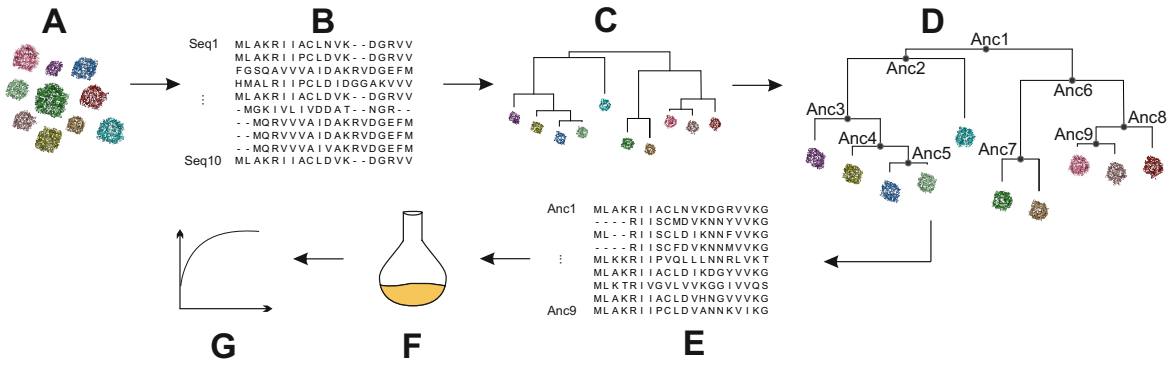

Figure 1.3: "Resurrection" of ancestral proteins based on ASR. The procedure consists of the steps illustrated in panels **A - G**. A set of homologous proteins **(A)** is chosen as the starting point. The protein sequences are aligned to an MSA **(B)** and a phylogenetic tree is derived **(C)**. By means of the phylogenetic tree, the sequence set, and a substitution model, the ancestral sequences related to the bifurcations of the tree are inferred **(D, E)**. Based on these sequences, proteins can be produced recombinantly, **(F)** and characterized by means of biophysical and biochemical methods **(G)**.

approach. There are several programs available, like the ML approach `RAxML` (Stamatakis, 2014) or the Bayesian approach `MrBayes` (Holder and Lewis, 2003). To select the best fitting model for the data set at hand, `ProtTest` (Abascal et al., 2005) can be used to identify the best generating evolutionary model. The validity of the derived phylogenetic model can be confirmed with bootstrapping in an ML analysis (Felsenstein, 1985) or with the help of multiple samples from the posterior distribution for Bayesian analyses (Rannala and Yang, 1996). The chosen extant sequences and the derived phylogenetic tree (**Figure 1.3 A, C**) combined with a substitution model form the basis for the computation of the ancestral sequences. In principle, ASR computes for each internal node a matrix indicating for each residue position the probability distribution of all amino acids. For the sake of simplicity, in most experiments the sequence with the highest likelihood has been considered for each internal node (**Figure 1.3 D, E**); see for example (Perica et al., 2014). Several programs, compared by Joy et al. (2016), are available for inferring ancestral sequences. An experimental characterization of the corresponding proteins requires the production of the protein in a recombinant form, expression of the protein in host cells and the characterization with biochemical experiments, e. g., activity assays (**Figure 1.3 E - F**).

Driven to extremes, ASR makes it possible to characterize ancestral proteins that date back to the Last Universal Common Ancestor (LUCA) that existed in the Paleoarchean era, i. e. at least 3.5 billion years ago (Nisbet and Sleep, 2001). These "resurrection" experiments have elucidated many aspects of the early life on Earth and the evolution of proteins and macromolecular complexes. For example, Wheeler et al. (2016) discussed several ancestral proteins, e. g. the ancestor of thioredoxin (Perez-Jimenez et al., 2011), which exhibit elevated thermostability. Busch et al. (2016) characterized an ancestral enzyme complex, namely the tryptophan synthase (TS). Regarding to functional properties at early stages of evolution, several ancestral

proteins exhibit broad substrate recognition, like the ancestor of the serine protease (Wouters et al., 2003).

A second reason for the great success is that ASR adds a further dimension to sequence analysis: From an evolutionary point of view, extant homologs represent variants observed for one point in time, thus the comparison of these proteins was termed "horizontal" approach. In contrast, ASR is a "vertical approach", as it takes into account the evolutionary history of the proteins under study. Considering the chronology of mutations is more straightforward to identify crucial but subtle amino acid differences (Harms and Thornton, 2010), because the sequences generated for internal nodes are similar to each other and contain fewer neutral mutations than many extant sequences. Thus, vertical approaches can drastically reduce experimental efforts to identify key residues.

For example, the vertical approach has been used to elucidate the linkage between protein structure and its function (Gumulya and Gillam, 2017). Additionally, Perica et al. (2014) showed that ancestral pyrimidine operon regulatory protein, PyrR, exhibit different oligomeric states and revealed 11 key mutations controlling this state. Ugalde et al. (2004) examined green flourescent protein (GFP)-like proteins from corals, where the ancestral genes illuminate in green, which turned to a red emission in the extant corals through a stepwise adaption. Moreover, ancestors of the sugar isomerase HisA from the histidine biosynthesis were examined to reveal the positions leading to promiscuity, i. e. a broad protein specificity (Plach et al., 2016).

Interestingly, it turned out that resurrected proteins are generally more stable and possess often a broader substrate specificity than the extant sequences used for reconstruction (Wheeler et al., 2016). It is a matter of debate, whether this higher thermostability is an artifact of the ASR protocol or a general feature of ancestral proteins (Williams et al., 2006). Protein design problems can profit from these properties as shown for the design of 3-isopropylmalate dehydrogenase (Watanabe et al., 2006) leading to designed enzymes with even higher thermostability. Zakas et al. (2017) designed a pharmaceutical important coagulation factor VIII that benefited from ASR with respect to biosynthetic efficiency, specific activity, stability, and immune reactivity. Cole et al. (2013) introduced a method that exploits a vertical approach as an additional source of information for altering or enhancing the function of the protein in protein engineering.

The application of ASR profited from the rapid progress of quite different life-science technologies: The outcome of sequencing projects led to an exponential growth of databases making a huge number of proteins available for ASR. Progress in gene-synthesis accompanied by a drastic reduction of costs turned resurrection experiments into a cost-effective tool to generate results in a timely manner. Ironically, the step to be expected least critical in resurrection experiments, namely ASR, became a bottleneck. As illustrated above, ASR can be divided into four steps, and some critical aspects will be highlighted in the following. The final outcome of ASR are the sequences of the internal nodes, whose composition depends on the phylogenetic tree computed beforehand for the chosen set of extant sequences and by applying an evolutionary

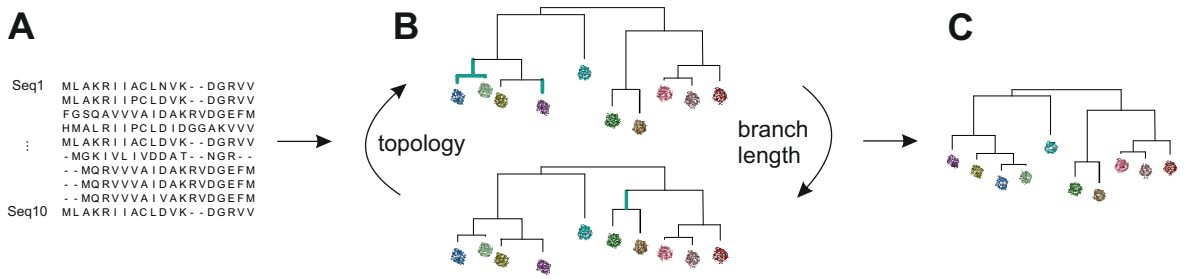

Figure 1.4: Calculation of a phylogenetic tree. The procedure consist of the steps illustrated in **A - C**. Based on an MSA consisting of extant sequences **(A)** a first phylogenetic tree is derived **(B)**. The topology and the branch lengths are consecutively optimized (changes are indicated in cyan) in order to increase the likelihood of the phylogenetic tree. These issues are solved as part of an optimization problem to obtain the final tree **(C)**, which is the most likely tree with respect to the input sequence set and the chosen phylogenetic model.

model. However, the user has to assess critically the phylogenetic tree prior to the reconstruction step in order to exclude errors that might rule out a valid reconstruction. Most critical are the length of all branches and the topology of the tree (Merkl and Sterner, 2016). For a reliable reconstruction, all branch lengths must be lower than one mutation per site to allow for a modelling of all mutations. The topology should be as unambiguous as possible to rule out alternative evolutionary scenarios. Even, if all sequences share a CA, i. e. are homologous, horizontal gene transfer (HGT) may cause topologies that are not compatible with the expected phylogeny. If the proteins under study are multi domain proteins, their composition has to be compared with great care to ensure that all proteins possess the same domains in the same order. A further problem that can impede reconstruction is the number of insertions and deletions that occurred during the genesis of the recent sequences. Only few algorithms can model some of these events in an evolutionary correct manner (Löytynoja and Goldman, 2008; Ashkenazy et al., 2012). Taken together, these constraints emphasize the judicious selection of the sequence set. This choice implies a sequence selection; however, their suitability for ASR is only confirmed after the computation of a tree. It follows that sequence selection is an iterative process, which requires to integrate a phylogenetic analysis.

It is the calculation of a phylogenetic tree (**Figure 1.4**) that turns ASR into a time-consuming process. As indicated above, the phylogenetic tree is derived from a given MSA of extant sequences (**Figure 1.4 A**). The calculation of the phylogenetic tree (**Figure 1.4 B**) can be viewed as an optimization problem: Topology and branch lengths are optimized consecutively (**Figure 1.4 C**, indicated in cyan) in order to increase the likelihood of the tree. After several rounds of optimization, the most likely phylogenetic tree regarding to the sequence data is obtained (**Figure 1.4 D**) and then the suitability of the tree for ASR can be assessed. Phylogenetic trees not suitable for ASR cannot be changed directly, as the appearance of the tree is determined by the sequence set. Thus, the sequence set has to be changed in order to support a tree suitable for ASR (Merkl and Sterner, 2016). However, alterations in the sequence set often

lead to unexpected changes in the topology, thus several rounds of alterations in the sequence set are necessary to obtain a suitable tree for ASR.

Since popularity and strength of ASR has increased during the last years, not only command line tools, but also simple-to-use webserver or programs are available that deduce a phylogenetic tree (Guindon et al., 2010; Stamatakis, 2014; Lartillot et al., 2009; Ronquist and Huelsenbeck, 2003). If a suitable data set is at hand, protocols that execute all steps of ASR can be applied (Tamura et al., 2011; Hanson-Smith and Johnson, 2016; Dereeper et al., 2008). However, a protocol for the compilation of a suitable sequence set leading to a reliable tree is not available. Moreover, all programs can only handle a relatively small number of sequences, which implies their deliberate selection from the enormous number of sequences deposited in databases like InterPro or UniProt (Li et al., 2008; Frickey and Lupas, 2004). Due to the design of the algorithms, between 150 and 200 sequences should be chosen for an ML approach and 30 to 80 present the limit for a Bayesian approach (Hanson-Smith and Johnson, 2016; Dereeper et al., 2008). So far, there exists no broadly applicable protocol for sequence selection; it is common practice to pick them manually with the help of an intuitive presentation (Hanson-Smith and Johnson, 2016; Dereeper et al., 2008). A few algorithms have been established to take over at least some part of the filtering procedure. Starting with sequences collected by means of a `BLAST` search, the algorithm implemented by Goremykin et al. (2010) excludes sequences based on their similarity and outputs sets of maximal 150 entries; a similar approach is `cd-hit` (Li and Godzik, 2006). Other programs, like `Gblocks` (Castresana, 2000) or `trimAl` (Capella-Gutiérrez et al., 2009) eliminate rows from the MSA that contain a large number of gaps in order to increase the quality of the phylogenetic signal. Thus, methods are available that solve some subtasks of sequence preparation; however, there exists no protocol that considers the above-mentioned criteria in a comprehensive manner.

1.3 Aim and Scope of this Work

During the last years, ASR turned from a method mastered by few specialists to a frequently used technology, although a generally accepted protocol is missing. In order to allow for the reliable reconstruction of proteins, a standard protocol was established within the scope of this thesis. It was used to reconstruct ancestors of the imidazole glycerol phosphate synthase (ImGPS) and the TS that were both characterized on their biochemically properties. Within the protocol, several features were used for sequence filtering, namely the length of the unaligned sequences, the amount of indels in the alignment, the length of the branches and the value of bootstrap values or posterior probabilities. In addition to this standard protocol a further protocol was developed to identify crucial positions with the help of a vertical approach, e. g. of complex formation. A combination of biochemical characterization and the *in silico* assessment of these proteins allowed us to narrow down several candidate positions to one crucial positions. Due to

the versatility of vertical approaches, the protocol can be adapted to different scientific problems.

Based on the standardized protocol, sequence selection was further improved by focusing on their rational selection in an automated manner. To perform this task, `FitSS4ASR` was developed that uses iteratively the above-defined features to evaluate sequence sets and phylogenetic trees and to remove sequences. The outcome are several alternative sets and the user can choose the most appropriate one. To support the user's decision, `FitSS4ASR` computes several scores assessing the phylogenetic variety of the sequence set and the robustness of the tree. Thus, `FitSS4ASR` makes it possible to find a suitable data set in a semi-automated manner.

As already mentioned, a standard protocol for ASR was established within the reconstruction of ancestors of ImGPS and TS. In order to reveal the level of specialization of an ancestral enzyme complex, the TS from the last bacterial common ancestor (LBCA) was reconstructed and experimentally characterized. It turned out that the reconstructed TS consists of two TrpA and two TrpB subunits as the TS from *Salmonella typhimurium* (stTS). Moreover, a comparison of the ancestral protein and the extant proteins made clear that TrpA and TrpB activate each other allosterically. A biochemical characterization showed a deactivation in the ancestral complex, whereas an activation occurs in the extant complex. Comparisons of the crystal structures of both complexes were conducted to link the differences in the activation process to differences on substructure or residue level; however, we were not able to pinpoint residues or structural parts responsible for the allosteric activation.

A second application of ASR has been performed on ImGPS, which consists of the synthase HisF and glutaminase HisH. To identify hotspots of complex formation, reconstructed HisF subunits were combined with the HisH subunit from *Zymomonas mobilis* (zmHisH). Interestingly, two ancestral HisF subunits had a differing binding behavior; thus, mutational experiments combined with *in silico* predictions were sufficient to narrow down the candidate positions to one hotspot. This application is an example indicating how a vertical approach allows for a specific property the rapid identification of a crucial position.

1.4 Guide to the Following Chapters

Each of the following four chapters corresponds to one manuscripts; two of them have been published and one is an accepted chapter of the book "Computational Methods in Protein Evolution". One chapter contains unpublished data.

The manuscript ***Ancestral Sequence Reconstruction as a Tool for the Elucidation of a Stepwise Evolutionary Adaptation*** describes our standard protocol of ASR and several pitfalls. Taking ImGPS as an example, it is also shown, how ASR can be used to identify hotspots in protein-protein interactions. ImGPS is a heterodimer consisting of the synthase subunit HisF and the glutaminase subunit HisH. By comparing the sequences of intermediate sequences leading from the LUCA-HisF to the extant HisF from *Pyrobaculum arsenaticum*

(paHisF) a neighbored pair of ancestral HisF subunits differing in the strength of complex formation to the extant zmHisH was identified. The candidate positions responsible for the different binding behavior are assessed by comparing the sequences. Furthermore, the approach is illustrated to narrow down few candidate positions with the help of structural and biochemical evaluation in combination with *in silico* predictions: Specifically, for the ancestral HisF subunits, it was demonstrated that one hotspot modulates protein-protein interaction. The *in silico* prediction was confirmed by an assessment of the complex consisting of HisF from *Thermotoga maritima* (tmHisF) and zmHisH. Furthermore, the transferability of the protocol to other scientific problems is shown.

The following chapter ***Sequence Selection by `FitSS4ASR` Alleviates Ancestral Sequence Reconstruction as Exemplified for Geranylgeranylglyceryl Phosphate Synthase*** contains unpublished data and describes the novel protocol `FitSS4ASR` that supports the user in selecting sequences for ASR (see also **chapter 2**). `FitSS4ASR` requires as input a sequence set that consist of several thousand homologs. This set is iteratively reduced with the help of sequence filters and by analyzing phylogenetic trees. The output of `FitSS4ASR` are several sequence sets of differing size, which are scored with respect to their suitability for ASR. The suitability of `FitSS4ASR` was made plausible by analyzing the trees deduced for the geranylgeranylglycerol phosphate synthase (GGGPS), which is an enzyme that forms taxon-specifically homodimers or homohexamers. The computed trees and inferred ancestors were compared to show the validity of `FitSS4ASR`.

The publication ***The Ancient Nature of Allostery and Substrate Channeling in the Tryptophan Synthase Complex*** reports on an application of ASR related to the TS from the LBCA. TS consists of the subunits TrpA and TrpB and the reconstructed sequences were the basis for a recombinant production and the subsequent experimental characterization. It turned out that the sophisticated allosteric activation observed between the two subunits of TS from *Salmonella typhimurium* existed already at an early phase of evolution. Comparison of crystal structures made clear that the structure of the subunits and their arrangement in the complex were not altered within 3.14 billion years.

The publication ***Combining Ancestral Sequence Reconstruction with Protein Design to Identify an Interface Hotspot in a Key Metabolic Enzyme Complex*** describes an application of a vertical approach used to identify binding hotspots of the protein-protein interface in ImGPS. The binding strength of reconstructed HisF enzymes to the zmHisH subunit was experimentally determined. Correlating these data with differences in the reconstructed interfaces, putative hotspots were predicted, which were further assessed by means of other *in silico* methods. We could show that one residue position is crucial for binding.

Chapter 2

Ancestral Sequence Reconstruction as a Tool for the Elucidation of a Stepwise Evolutionary Adaptation

Kristina Straub and Rainer Merkl

To appear as a book chapter of
Computational Methods in Protein Evolution: Methods and Protocols, Springer, New York. In Press, Editor: Tobias Sikosek

Key words ancestral sequence reconstruction, vertical analysis, evolutionary biochemistry, *in silico* mutagenesis, protein-protein interaction.

Abstract

Ancestral sequence reconstruction (ASR) is a powerful tool to infer primordial sequences from contemporary, i. e. extant ones. An essential element of ASR is the computation of a phylogenetic tree whose leaves are the chosen extant sequences. Most often, the reconstructed sequence related to the root of this tree is of greatest interest: It represents the common ancestor (CA) of the sequences under study. If this sequence encodes a protein, one can 'resurrect' the CA by means of gene synthesis technology and study biochemical properties of this extinct predecessor with the help of wet-lab experiments.

However, ASR deduces also sequences for all internal nodes of the tree and the well-considered analysis of these 'intermediates' can help to elucidate evolutionary processes. More-

over, one can identify key mutations that alter proteins or protein complexes and are responsible for the differing properties of extant proteins. As an illustrative example, we describe the protocol for the rapid identification of hotspots determining the binding of the two subunits within the heteromeric complex imidazole glycerol phosphate synthase.

2.1 Introduction

A major goal of life scientists is to understand the function of proteins on the residue level and often, computational biology contributes a lot to the finding of functionally or structurally important residues; for a review see Lee et al. (2007). For example, if the 3D structure of a protein is known, one can assess the contribution of individual residues to protein stability (Schymkowitz et al., 2005); additionally, one can predict catalytic sites (Janda et al., 2013) and protein interfaces (Zellner et al., 2012) by analyzing cavities or surface residues. Moreover, the comparison of results deduced for homologous proteins allows one to elucidate the evolution of specific protein functions (Plach et al., 2015). Similarly, protein sequences can be utilized; however, the predictive power of corresponding algorithms depends on the number of sequences that are at hand. In the post-genomic era, computational protein biology profits from the enormous number of known orthologs, i. e. sequences from different species that have the same ancestor and encode identical or similar functions. In order to identify residue positions that are crucial for a specific family, it is a common approach to generate a multiple sequence alignment (MSA), which is subsequently utilized to determine for each position in the protein the conservation level of each residue (Edgar and Batzoglou, 2006).

This and similar approaches are often named 'horizontal', because they are based on the analysis of a certain phase of evolution represented by the proteins found in extant species. Due to the enormous number of known sequences, these residue distributions can be determined quite precisely and the horizontal approach allows the identification of residues that are important for all members of a family. However, this method rarely identifies sets of residues that determine specificity in a family of functionally diverse proteins (Harms and Thornton, 2010). Thus, to study protein evolution, a more detailed analysis is needed, for example based on a clustering of sequences by means of neighbor joining (Saitou and Nei, 1987). A state-of-the-art method for the study of divergent evolution even in very large protein families is the usage of sequence similarity networks and genome neighborhood networks; for a recent review see Gerlt (2017). Such cluster algorithms are based on a simplified model of protein evolution; due to their computational complexity, models that are more elaborated are not applicable for the analysis of large datasets.

Although only applicable to a relatively small number of sequences, the implementation of highly reliable phylogenetic algorithms has added a further dimension to sequence analysis: It makes possible to trace back the evolution of a fair number of extant orthologs to common ancestors. If functional diversity is known for some of the extant orthologs, this 'vertical' approach

has great potential, because one can reconstruct the sequences of putative predecessors and identify those mutations that occurred along that branch of the family tree on which functional diversification occurred (Harms and Thornton, 2010).

The vertical approach is a specific application of ancestral sequence reconstruction (ASR), which became popular during the last decade, especially in combination with 'resurrection' experiments; for recent reviews see Merkl and Sterner (2016); Thornton (2004); Brooks and Gaucher (2007) or Hochberg and Thornton (2017). The typical protocol of each ASR consists of two steps: First, the user has to compute a phylogenetic tree tr_{phylo}. In all cases, the extant orthologs chosen by the user constitute the leaves, but the topology of tr_{phylo} is determined by sequence similarity, the selected evolutionary model, and the algorithm used for its computation. In contrast to a classical phylogenetic analysis, ASR requires a subsequent step that deduces for all internal nodes of tr_{phylo} sequences that represent predecessors. The composition of these sequences critically depends on the content of the leaves (extant orthologs) but also on the topology of tr_{phylo}. This is why tr_{phylo} has to fulfill certain quality criteria to guarantee proper sequence reconstruction. Nowadays, it is straightforward to supplement such an *in silico* reconstruction with wet-lab experiments: One can recombinantly resurrect proteins with the help of gene synthesis and characterize them with classical biochemical and biophysical methods (Thornton, 2004). Besides their relevance for answering evolutionary problems, resurrected proteins became increasingly important in protein engineering, because one can beneficially exploit their promiscuity (Bornscheuer et al., 2012) to tailor protein function (Romero-Romero et al., 2016).

In addition, the fact that ancestral proteins are frequently 'generalists' motivates their usage in vertical approaches. In the following, we detail a protocol for the identification of specificity-determining residues. The general strategy is to select a protein family of interest and a property to be evaluated. Then, one has to infer a phylogenetic tree and choose the branches of the family tree to be analyzed. The selection of branches may depend on *in silico* or wet-lab experiments aimed at finding branch-determining leaves, i. e. extant proteins with differing functions. The final task is to reconstruct the sequences of predecessors with the help of ASR (*see* **2.2.1**) and to identify specificity-determining residues by comparing the sequences of ancestral sequences within the chosen branches (*see* **2.2.2**). Again, the assessment of these residues may comprise *in silico* and/or wet-lab analyses.

We used this strategy to study the stepwise adaptation of the protein-protein interface (PPI) from the heterodimeric imidazole glycerol phosphate synthase (ImGPS). This enzyme mediates the incorporation of nitrogen into PRFAR by catalyzing the transfer of the amido nitrogen of glutamine to an acceptor substrate (Massiere and Badet-Denisot, 1998; Zalkin and Smith, 1998). In bacteria and archaea, ImGPS consists of the cyclase subunit HisF and the glutaminase subunit HisH, which assemble with high affinity to a bi-enzyme complex (Beismann-Driemeyer and Sterner, 2001). Despite detailed biochemical and structural studies (List et al.,

2012), the specific residue positions responsible for HisF:HisH complex formation were unknown. This is why we identified key residue positions of this PPI by means of a vertical approach (Reisinger et al., 2014b; Holinski et al., 2017), which is illustrated in **Figure 2.1**.

2.2 Protocol

2.2.1 Ancestral Sequence Reconstruction

- Collect a large number of orthologs. Start with a specific sequence of interest and use `BLAST` (Altschul et al., 1990) to deduce orthologs from the *nr* or *refseq_protein* databases of the NCBI (Pruitt et al., 2009) or the EBI database *UniProt* (UniProt, 2013); alternatively select the corresponding *InterPro* family (Hunter et al., 2012) (*see* **Note 1**). Choose a *bona fide* protein as a reference sequence and, if possible, several sequences that can serve as an outgroup. Additionally, include the sequences of those proteins ($prot_i$) that possess differing properties, whose determinants shall be elucidated by the subsequent analysis.
- Create an MSA. According to our experience, `MAFFT` (Katoh and Standley, 2013) is a highly versatile and robust method that can cope with large sequence sets (*see* **Note 2**).
- Eliminate redundant sequences and obvious outliers like those that are much shorter or longer than the reference sequence. Additionally, eliminate sequences that induce conspicuously large indels in the MSA (*see* **Note 3**). A versatile tool supporting these tasks is `Jalview` (*see* **Note 4**).
- Repeat steps 2 and 3 until the MSA consists of a homogeneous set of sequences.
- If the protein under study is part of a larger complex, perform MSA generation for each subunit. Afterwards, concatenate the sequences in a species-specific manner (*see* **Note 5**) and create an MSA consisting of the concatenated sequences.
- Optionally, replace the database identifiers with more informative names for the sequences (*see* **Note 6**). Remove less informative residue positions from the MSA. Apply `Gblocks` (Castresana, 2000) to eliminate all columns containing more than 50 % gaps. Use the resulting MSA for the inference of the phylogenetic tree, but not for the subsequent sequence reconstruction, which is based on the full MSA. Compute a phylogenetic tree tr_{phylo} with a method of choice. We prefer `PhyloBayes` (Lartillot et al., 2009) and start eight independent MCMC samplings in parallel with a maximal length of 50,000 samples to guarantee congruence (*see* **Note 7**). If congruence is reached, we deduce the consensus tree computed by `readpb` from the samples following the burn-in phase of the MCMC computation. The number of samples that have to be excluded (burn-in) can be determined with `VMCMC` (Ali et al., 2017); often, the first 25 % of the samples are considered as burn-in and discarded. Alternatively, use other state-of-the-art probabilistic methods like `MrBayes` (Ronquist and

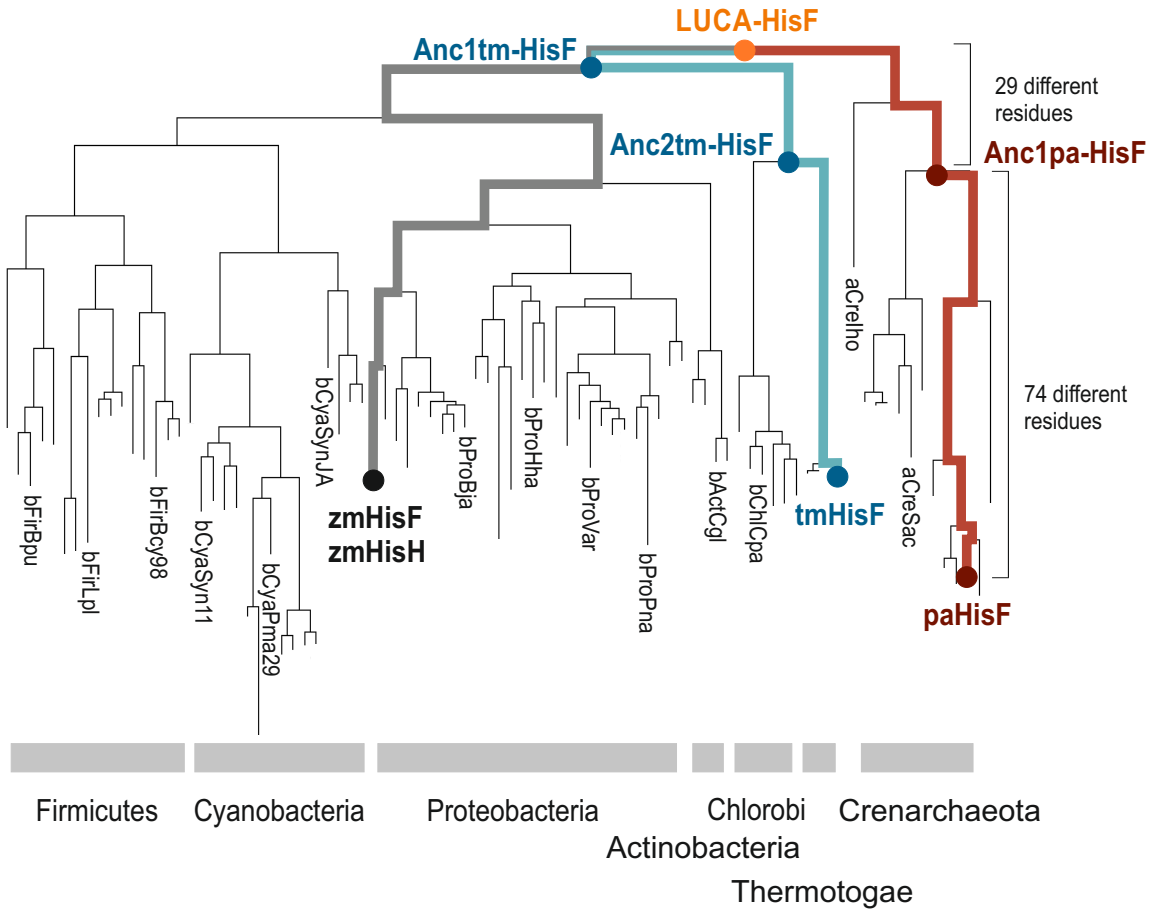

Figure 2.1: Identification of specificity-determining residue positions of the HisF:HisH interface by means of a vertical approach. Initial binding studies had shown that subunits from phylogenetically unrelated species are not compatible: The HisF subunit from the Crenarchaeon *Pyrobaculum arsenaticum* (paHisF) did not bind HisH from the Proteobacterium *Zymomonas mobilis* (zmHisH). For the rapid identification of crucial residue positions within the HisF interface, 87 HisF sequences from seven phyla were chosen for a vertical analysis. Thus, we deduced ancestral sequences linking the native interaction partner of zmHisH, namely zmHisF (the leaf of the grey branch) and the distant paHisF (the leaf of the brown branch). Ancestral proteins were resurrected and their binding to zmHisH was characterized experimentally. HisF corresponding to the Last Universal Common Ancestor (LUCA-HisF) bound zmHisH. In contrast, the first intermediate (Anc1pa-HisF) on the branch leading to paHisF that differed markedly from LUCA-HisF did not bind zmHisH. Anc1pa-HisF deviates from LUCA-HisF by not more than 29 residues, but from paHisF by 74 residues. A subsequent *in silico* analysis focusing on the PPI of HisF allowed us to narrow down the number of putative key residue positions to two. Their role was assessed by experimental binding studies; one was identified as an interface hotspot. To trace the species-specific evolution of PPIs in more detail, the two predecessors (Anc1tm-HisF and Anc2tm-HisF) on the path (shown in blue) leading to HisF from *Thermotoga maritima* (tmHisF) were resurrected as well. Both intermediates bound zmHisH, but tmHisF was a poor binder. The mutual exchange of residues from the latter three sequences at corresponding positions confirmed their hotspot quality; for details see Holinski et al. (2017) or **chapter 5**. Note that these residues are located at the rim of the PPI and only moderately conserved, which explains why they have not been discovered previously. To avoid overloading the graph, only a few of the extant sequences are shown with their `Key2Ann` annotation indicating the phylogenetic lineage, i. e. the superkingdom (first character), the phylum (following three characters) and the species name (last three characters).

Huelsenbeck, 2003) or `BEAST` (Bouckaert et al., 2014) to compute the phylogenetic tree (*see* **Note 8**). For a given MSA of amino acid sequences, one can utilize `ProtTest` (Abascal et al., 2005) to determine the best fitting evolutionary model prior to MCMC sampling.

- Visualize tr_{phylo} by means of `NJplot` (Perriere and Gouy, 1996) or `FigTree` (Rambaut, 2012) and assess the length of the individual edges and their posterior probabilities. All edge lengths must indicate mutation rates $<<1$ mutation per site and the posterior probabilities of relevant internal nodes must exceed the value of 0.75. Furthermore, make sure that the resulting phylogenetic hierarchy of the chosen sequences (species) is plausible: For example, compare the topology of tr_{phylo} with the relationships of the sequences (species), determined for the iTOL project (Ciccarelli et al., 2006) or the 'nearly universal tree' of life (Puigbo et al., 2009). This comparison allows one to eliminate cases of horizontal gene transfer and to avoid long-branch attraction. If tree topology is not plausible, consider to choose a different set of sequences and repeat the procedure (*see* **Note 9**).
- If the sequence set does not contain an outgroup, use `NJplot` (Perriere and Gouy, 1996) or an alternative algorithm to root tr_{phylo} for subsequent sequence reconstruction. Positioning the root is critical for the computation of the CA sequence. Choose the location of the root according to a plausible hierarchy to be determined by one of the methods described in the previous step. If an outgroup was used for rooting, we recommend to eliminate the corresponding sequences during sequence reconstruction to prevent undesired effects on residue composition.
- Use the rooted tree prepared in the last step and the full MSA to reconstruct the ancestral sequences related to internal nodes. Methods of choice are `PAML` (Yang, 2007) or `FastML` (Ashkenazy et al., 2012), which can handle indels (*see* **Note 10**). If possible, choose the same substitution model as used for tree construction. ASR programs compute for each residue position posterior probabilities for all 20 amino acids. If alternative predictions with relatively high posterior probabilities exist, a near-ancestor sequence ensemble can be calculated for each node; for details see Bar-Rogovsky et al. (2015). If one sequence per internal node is of interest, select for each position the residue possessing the highest posterior probability.

2.2.2 Identification of Specificity-determining Residues by Means of Intermediate Sequences

- In analogy to **Figure 2.1**, determine the branches of tr_{phylo} that interconnect the two or more recent proteins $prot_i$ under study, i. e. those that possess diversified properties.
- Compile an initial set *anc_prot*, consisting of ancestral proteins that differ most likely from the extant proteins $prot_i$ and support an efficient characterization. For example, one can pairwise compare all ancestral sequences to choose several intermediates, i. e. ancestral

sequences that span the sequence differences between the $prot_i$ in approximately similar proportions. We recommend the usage of `Jalview` for sequence selection (*see* **Note 11**). The finding that primordial proteins are often generalists suggests to add the CA sequence to *anc_prot* and to characterize the corresponding protein with high preference.

- Optional step: If the 3D structure of a $prot_i$ is known, compute homology models of all *anc_prot* (*see* **Note 12**) and try to minimize further the number of candidate residues to be studied in the following steps. If protein function is of interest, use the compiled annotations of `PDBsum` (*www.ebi.ac.uk/pdbsum/*) or an alternative database to assess the position of the differing residues with respect to a catalytic center or a binding site. If complex formation is under study, consider a webserver like `PISA` (*www.ebi.ac.uk/pdbe/pisa/*) that details characteristics of residues located in PPIs. One can also predict the contribution of residues to protein or complex stability by utilizing force fields to calculate differences in free energy (see **Note 13**). For the example presented in **Figure 2.1**, we could reduce the number of putative key residue positions to two by combining *in silico* approaches.
- Optional step, if experimental characterization is intended: Choose protein sequences for the resurrection experiments and design their gene sequences. Produce the proteins recombinantly and characterize them according to the specific problem. The choice of suitable wet-lab experiments depends on the characteristics under assessment and may contain tests of enzyme activity or complex stability. Additionally, it is advisable to confirm proper protein folding by means of far-UV CD spectroscopy.
- Associate the determined effects with the introduced mutations to deduce the stepwise evolutionary adaptation towards the properties of recent proteins. In case of ambiguous results, repeat steps **StepChooseIntermediates** - **StepExpCharacterization** of the protocol given in **section 2.2** and extend the analyses to additional intermediates and/or single point mutations.

2.3 Notes

1. Compiling an appropriate sequence set for ASR is more an art than an artisanal activity and sequence selection is an iterative process that requires several rounds of user interaction. This is why the initial number of sequences should be as high as possible. Choose sequences that are most likely orthologs and avoid the addition of paralogous sequences by comparing gene duplicates. If a Bayesian approach is used to infer the phylogenetic tree, running time is an issue that currently limits the finally selected number of recent sequences to ≈ 200. Make sure that the chosen sequences originate from phyla needed to deduce the intended set of predecessors. If one wants to represent the last universal common ancestor, the chosen sequences must at least come from several bacterial and archaeal

clades. Store sequence sets in Multi-FASTA file format, which is accepted by most tools required for ASR.

2. Use a `MAFFT` method that is accuracy-oriented, i. e. one of the `'INS'` modes. This selection depends on the size of the MSA; for details see the `MAFFT` manual. For the initial generation of large MSAs, the option `--auto` is also appropriate.

3. Modelling the history of insertions and deletions on an evolutionary time scale is difficult and requires for most ASR algorithms the manual adjustment of primordial sequences. One can minimize errors by choosing a set of sequences of relatively uniform length.

4. `Jalview` is an excellent tool for the preparation of sequence sets used in ASR. The `Jalview` command `Edit\Remove redundancy` allows the selection of a percentage identity threshold and initiates the subsequent comparison of all sequence pairs. If the similarity of any two sequences exceeds this cutoff, the shorter sequence is discarded. A cutoff of 95 % or lower is useful to remove redundant sequences and to avoid highly articulated subtrees. The command `Calculate\Sort by length` makes it possible to identify easily sequences that are much shorter or longer than the reference sequence. These sequences and those introducing strikingly long indels can be erased by clicking their name and the delete button. The command `Web Service\Alignment` offers several alternatives for MSA creation, among them is `MAFFT`.

5. Concatenation helps to deduce a robust tree due to the stronger phylogenetic signal spread over a larger set of residue positions. Make sure that the sequences originate from the same species by using for their linkage the *Tax-Id* assigned by the `taxonomy browser` of the NCBI. Note that concatenation is only valid for sequences that co-evolve and share the same evolutionary history for the entire period under study.

6. For the visual inspection of trees, it is helpful to replace the hard to interpret database identifiers with names that indicate the function of the proteins and/or the phylogenetic position of the species contributing the sequences. We use our in-house tool `Key2Ann` (Pürzer et al., 2011) to denote the phylogenetic lineage; *see* **Figure 2.1** for an example.

7. A detailed description of all the programs and their options belonging to the software suite `PhyloBayes` can be found at *www.phylobayes.org*. For the reconstruction of amino acid sequences, we use the `CAT` or `JTT` model and specify a minimal effective sample size of 100. Congruence can be tested by calculating the *maximum difference of posterior probabilities of tree bipartitions* (*maxdiff*) by using the `PhyloBayes` tool `bpcomp`; the *maxdiff* value should be below 0.3 (Lartillot et al., 2009). Computation time can be reduced by using the multi-core version `PhyloBayes-MPI`. Note that an MCMC calculation may take several weeks, if a large number of recent sequences were chosen.

8. A detailed description of the `BEAST` functionality can be found at *www.beast2.org*. The `BEAST` tool `LogCombiner` can be used to discard the burn-in samples and `Tracer` allows

one to determine the effective sample size. `TreeAnnotator` assists the user in summarizing information from a sample of trees onto a consensus tree. Computation time of `BEAST` can be reduced by incorporating the `BEAGLE` library for parallel processing.

9. Long branches (> 1.0 mutations per site) and low posterior probabilities (< 0.75) prevent the reliable computation of ancestral states. The same is true, if divergence of the sequence set is too small or if the tree is highly articulated. To overcome these problems, the content of the sequence set has to be altered. For example, one can exclude sequence sets amendable to long branches and erase some sequences in highly articulated subtrees.

10. According to our experience with MSAs containing a small number of indels, `FastML` performs well in ASR. If the MSA contains a larger number of indels, one can try several values of the advanced option `probability cutoff to prefer ancestral indel over character` and compare the results. For further processing, choose the sequences computed as a `marginal reconstruction`. Note that `FastML` does not offer all evolutionary models implemented for `PhyloBayes` or `BEAST`. Alternatively, one can use `PRANK` (Löytynoja and Goldman, 2008) or `Historian` (Holmes, 2017) that are based on alternative models of indel evolution. Due to the method used for indel reconstruction, the lengths of reconstructed sequences may deviate from the mean length of extant sequences as N- and C-termini are of higher variability than the rest of the sequence. Thus, it might be necessary to trim the reconstructed sequences.

11. For a set of sequences, the similarity of all pairs can easily be determined by executing the `Jalview` command `Calculate\Pairwise Alignment`.

12. Several alternatives are available to compute homology models of subunits and protein complexes, among them are `YASARA` (Krieger et al., 2009), `I-Tasser` (Zhang, 2008), or `HHSearch` (Söding, 2005) in combination with `Modeller` (Webb and Sali, 2014). For ASR experiments, one can expect reliable models, because the sequence similarity between the template (a $prot_i$) and the target (an *anc_prot*) is usually high.

13. The effect of a mutation on protein or complex stability can be assessed *in silico* by utilizing programs like `FoldX` (Guerois et al., 2002), which is a stand-alone application, but also integrated into `YASARA`. To predict the contribution of point mutations on protein stability, assess the corresponding $\Delta\Delta G$ values. To estimate the effect on complex stability, compute the $\Delta\Delta G$ value indicating the binding energy difference between a 'wild-type complex' and a complex with a mutated PPI. $|\Delta\Delta G$ values$| > 2$ kcal/mol are considered a significant contribution of one residue to complex stability. For this FoldX analysis, three functions have to be executed subsequently, namely `RepairPDB`, `BuildModel`, and `AnalyseComplex`. For this specific application of `FoldX`, the 'wild-type complexes' may consist of $prot_i$ or *anc_prot* sequences, which can differ in their length. In order to identify the corresponding residues, create an MSA containing $prot_i$ and *anc_prot* sequences to coordinate their positions.

Acknowledgement

This work was supported by the Deutsche Forschungsgemeinschaft (ME2259/2-1). Calculations were facilitated by using advanced computational infrastructure provided by the Leibniz Supercomputing Center of the Bavarian Academy of Sciences and Humanities (www.lrz.de) under grant pr48fu. We thank Samuel Blanquart for continuous support, many helpful hints, and fruitful discussions.

Chapter 3

Sequence Selection by `FitSS4ASR` Alleviates Ancestral Sequence Reconstruction as Exemplified for Geranylgeranylglyceryl Phosphate Synthase

This chapter contains unpublished data.

3.1 Introduction

During the last forty years, ancestral sequence reconstruction (ASR) became a very successful means of computational biology. Its usage has elucidated completely different aspects of protein evolution, which are intractable with other methods; for recent reviews see Joy et al. (2016); Merkl and Sterner (2016); Wheeler et al. (2016); Gumulya and Gillam (2017) or Hochberg and Thornton (2017). ASR algorithms compute for a given set of extant homologs a phylogenetic tree and deduce for all internal nodes the most likely sequences (Liberles, 2007). Their composition depends on the chosen phylogenetic model (Ashkenazy et al., 2012) and the extant homologous sequences that specify the leaves of the tree. Driven to extremes, the most ancient sequences that can be reconstructed are related to the LUCA that existed in the Paleoarchean era, i. e. at least 3.5 billion years ago (Nisbet and Sleep, 2001). The *in silico* and biochemical characterization of "resurrected" proteins from these early phases of evolution were key to characterize primordial proteins (Thornton et al., 2003; Hobbs et al., 2012) and the corresponding habitats (Perez-Jimenez et al., 2011). Due to the lack of macromolecular fossils, ASR is the only informative means to gain insight into the intricacy of ancient proteins (Reisinger et al., 2014a)

and to reproduce adaptations of extinct species to climatic, ecological and physiological changes (Boussau et al., 2008; Akanuma et al., 2013).

A second reason for the great success is that ASR adds a further dimension to sequence analysis: From an evolutionary point of view, extant homologs represent variants observed for one point in time, thus the comparison of these proteins was termed "horizontal" approach. In contrast, ASR is a "vertical approach", as it takes into account the evolutionary history of the proteins under study. Considering the chronology of mutations is more straightforward to identify crucial but subtle amino acid differences (Harms and Thornton, 2010), because the sequences generated for internal nodes are similar to each other and contain fewer neutral mutations than many extant sequences. Thus, vertical approaches can drastically reduce experimental efforts to identify key residues as demonstrated for the specificity of hormone receptors (Harms and Thornton, 2010), the fluorescence properties of GFP variants (Field and Matz, 2010), or the specificity of protein-protein interfaces (Holinski et al., 2017). Moreover, the insight that ancestral proteins are generally more robust and often more versatile, i. e. promiscuous, than their modern successors (Wouters et al., 2003; Wheeler et al., 2016) has opened new fields for the usage of reconstructed predecessors in protein design (Gumulya and Gillam, 2017).

In all these applications, a crucial prerequisite for a successful usage of ASR is the reliability of the reconstructed sequences. The protocols implemented for ASR are based on proven algorithms and for each step of the reconstruction process, probability measures allow for the assessment of their outcome; see e. g. (Merkl and Sterner, 2016). For the convenience of the user, specialized servers have been implemented that execute an ASR protocol in a fully automated manner for a given set of sequences (Dereeper et al., 2008; Kumar et al., 2012; Hanson-Smith and Johnson, 2016). As huge numbers of homologs do not necessarily improve the reconstruction of ancestral states (Li et al., 2008), not more than 150 to 200 input sequences are commonly picked by the user. However, the current databases offer for functionally important proteins several thousand homologous sequences, which urges the user to choose a drastically reduced subset. This selection process is an important and difficult phase of ASR, because additional constraints like sequence length or the phylogenetic origin of the chosen sequences have to be considered concurrently. Moreover, sequence selection greatly affects the quality of the phylogenetic tree, which must meet high standards for ASR (Pagel et al., 2004). Thus, one has to choose for the protein under study the homologs from those species that support a highly robust tree. In contrast, if one is interested to determine the phylogeny of a given set of species, it is a common approach to select proteins with strong phylogenetic signals (Salichos and Rokas, 2013). Consequently, due to the specific evolutionary history of individual species and proteins, a careful selection of the input is a critical step of most phylogenetic analyses.

Often, users create for ASR initially a large set of homologs by means of a `BLAST` search and pick sequences with the help of tools like `cd-hit` (Li and Godzik, 2006) or more specialized ones (Frickey and Lupas, 2004; Fuellen et al., 2005; Dereeper et al., 2008; Tamura et al., 2011).

In order to combine several orthogonal methods of sequence selection, we designed `FitSS4ASR`. This tool draws upon well-proven concepts applied to iteratively refined sequence sets and can be used in a semi-automatic manner with minimal user interaction. Our tool selects a small set of sequences representing a wide phylogenetic range that constitute a highly robust tree topology. We confirmed the validity of sequence selection *in silico* with the help of geranylgeranylglyceryl phosphate synthases (GGGPS). This enzyme is involved in the biosynthesis of membrane lipids and catalyzes the formation of an ether bond between glycerol 1-phosphate (G1P) and polyprenyl diphosphate (Tamura et al., 2011). We used the same ASR protocol, but two different set of recent GGGPS homologs to compute predecessors. The first set of sequences was compiled in an elaborate and time-consuming manner requiring extensive manual curation. The second set was created by applying `FitSS4ASR` that reduced user-intervention drastically. For both sets, phylogenetic trees and predecessors were computed.

3.2 Results

3.2.1 Criteria Guiding Sequence Selection for ASR

Commonly, the first step of sequence selection is the generation of a superset by means of `BLAST` (Altschul et al., 1997) or the choice of a precompiled dataset as offered by InterPro (Mitchell et al., 2014) or similar databases. Owing to the success of sequencing projects, these initial sets contain much more homologous sequences than practically useful. Thus, the aim in developing `FitSS4ASR` was not to support the user in constructing a tree for a given set of sequences, but to find a set of representatives that allow for the reliable reconstruction of predecessors.

One major constraint of sequence selection is the phylogenetic origin of the candidates that must represent a sufficiently wide phylogenetic diversity. For example, to reconstruct LUCA sequences, an optimal sequence set represents typically six dominating bacterial, two archeal and some eukaryotic clades (Hug et al., 2016). Usually, it is easy to provide for a given protein a broad phylogenetic representation due to the wide coverage of extant sequences deposited in databases. Thus, the crucial task of sequence selection is rigorous but specific filtering and the appropriate combination of filters might advantageously be exploited to increase the robustness of the ASR process. First, a single representative can be chosen for each subset of highly similar sequences. In order to eliminate flawed sequences caused by misassembly or gene-prediction errors, non-canonical outliers whose length differs significantly, i. e. by more than 2σ from the mean can be eliminated as well (**Figure 3.1 A**). The evolutionary correct modelling of indels is still difficult, thus it is appropriate to ignore also sequences with internal insertions as indicated by a sequence alignment (MSA, **Figure 3.1 B**) (Dereeper et al., 2008).

Other filter criteria (Merkl and Sterner, 2016) are only applicable after a phylogenetic tree was deduced from the input. Horizontal gene transfer (HGT, **Figure 3.1 C**) is a frequent

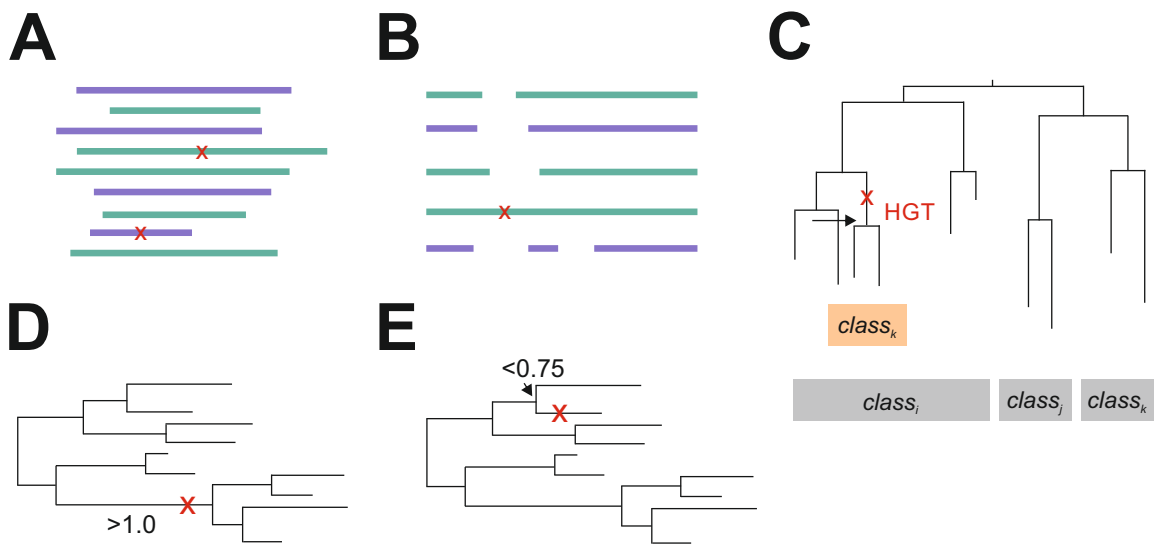

Figure 3.1: Criteria applied by FitSS4ASR to eliminate sequences. Striking elements (sequences or branches) are indicated by a red x. **(A)** Sequences that deviate in length significantly from the mean. **(B)** Sequences that possess internal insertions. **(C)** Sequences that were most likely transferred between the genomes of phylogenetically unrelated species by means of horizontal gene transfer (HGT) as exemplified for species from three phylogenetic classes $class_i$, $class_j$, and $class_k$. **(D)** Sequences inducing subtrees with long branches. **(E)** Sequences causing a weakly supported subtree topology.

phenomenon in bacterial genomes (Ochman et al., 2000), which complicates ASR due to non-constant mutation rates. To exclude the results of apparent HGT events, sequences that cause an aberrant phylogeny incompatible with a monophyletic origin have to be removed. Moreover, to eliminate long branch attraction (Bergsten, 2005) and to ensure a reliable reconstruction of ancestral states, the length of all branches has to indicate mutation rates below 1.0 mutations per site (**Figure 3.1 D**). Finally, bootstrap values/posterior probabilities corresponding to nodes of interest have to exceed the value of 0.75 (**Figure 3.1 E**), which is considered as an indicator of sufficient support (Soltis and Soltis, 2003). By removing sequences constituting an isolated subtree or by adding additional sequences, the user can modulate the topology and subdivide long branches (Wiens, 2005). However, the effects caused by an altered input are often unpredictable, which compels the testing of many alternative combinations. Thus, an interactive sequence selection may turn into a tedious and time-consuming torture.

3.2.2 FitSS4ASR: Filtering Sequence Sets for ASR

In order to support sequence selection in a comprehensive manner, FitSS4ASR consists of a series of methods that iteratively filter sequence sets and perform phylogenetic analyses to eliminate non-canonical sequences as described above (**Figure 3.2 A**). However, an elimination of HGT was not possible to execute automatically. Therefore, HGT has to be assessed and removed manually after running FitSS4ASR. To begin with, representatives are chosen based

on the outcome of `cd-hit` (Li and Godzik, 2006) that clusters sequences on their similarity. Subsequently, sequences that significantly deviate in length from the mean or introduce internal gaps are eliminated. The remaining sequences constitute the set $SEQ_{k=1}$, which is subjected to an analysis of tree topology. `FitSS4ASR` offers two alternatives for phylogenetic analysis, namely the maximum likelihood approach `RAxML` (Stamatakis, 2006) and the Bayesian approach `MrBayes` (Ronquist and Huelsenbeck, 2003). We parametrized both programs for the computation of a series of trees tr_k^i and a consensus tree tr_k . For subsequent analysis of tree robustness, `FitSS4ASR` saves during each iteration k the dataset $Iter_k = \{tr_k, SEQ_k\}$ consisting of the tree tr_k and the sequences SEQ_k under study.

Until the main loop of `FitSS4ASR` terminates, the series of trees tr_k^i is used to identify sequences with an ambiguous or insufficient phylogenetic signal (Sanderson and Shaffer, 2002) causing in the trees an unstable phylogenetic position based on two different criteria: `RogueNaRok` (Aberer et al., 2012) identifies "rogue" sequences that possess different sister sequences in trees generated during a phylogenetic analysis. The program eliminates sequences based on the *relative bipartition information criterion* (RBIC) that increases support of a tree and stops if RBIC cannot be further improved by pruning more sequences. However, this optimality criterion does not identify all unstable taxa (Wilkinson and Crotti, 2017), thus we implemented a more rigorous alternative that identifies "solitary" sequences. Our algorithm identifies sequences seq_k^r of a given set SEQ_k that do not possess any of the sisters s in at least 75 % of the k-specific trees tr_k^i ; see Methods. `FitSS4ASR` allows the user to choose one of three alternatives for sequence elimination, namely the removal of i) rogue, ii) solitary, iii) sequences that are rogue or solitary. The elimination of these sequences may create branches of undesired length. Thus, all other sequences inducing branches longer than 1 mutation/site are eliminated as well and the remaining sequences are subjected to further rounds of refinement, until one of two stopping criteria is reached: `FitSS4ASR` ends sequence elimination if SEQ_k contains not more than 60 sequences or if no sequences are eliminated during the last 10 iteration steps. Thus, `FitSS4ASR` generates a series of iteratively reduced sets and the output of the last iteration $Iter_{last} = \{tr_{last}, SEQ_{last}\}$ contains $u = |SEQ_{last}|$ sequences.

Upon completion of sequence elimination, `FitSS4ASR` assesses the robustness of the generated datasets to offer alternatives from which the user can choose (**Figure 3.2 B**). To begin with, up to 15 datasets $Iter_k^+$ that contain approximately evenly distributed between u and maximally 500 sequences are taken from the last rounds of sequence selection. Due to the nested hierarchy of the sequence sets SEQ_k^+ , we expect a consistent core topology of the trees tr_k^+ and deviations in individual trees are indicative of less suitable sequence sets. To eliminate such sets, `FitSS4ASR` deduces a supertree and discards trees tr_k^+ that are not compatible with this topology; see Methods. The m remaining sets $Alt_{s,s=1..m} = \{tr_s, SEQ_s\}$ are further subjected to a perturbation test, which we devised as a final assessment of tree robustness: We consider a sequence set SEQ_s "phylogenetically robust", if the addition of randomly chosen sequences has only a minor effect on tree topology. For a broad sampling, `FitSS4ASR` generates

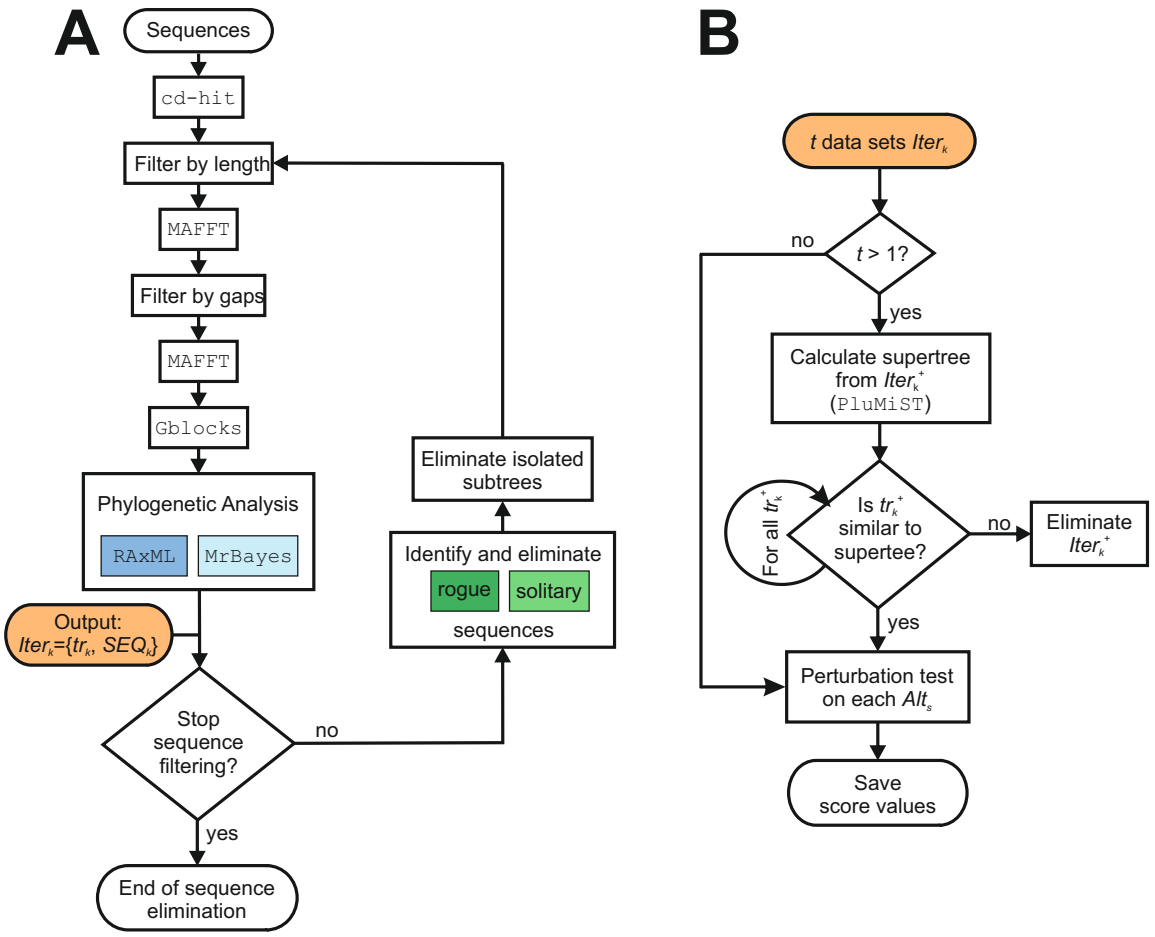

Figure 3.2: Workflow of FitSS4ASR. (A) Iterative sequence elimination. Initially, highly similar sequences identified by `cd-hit` and sequences that deviate significantly from the mean length or introduce internal gaps are removed. The remaining sequences constitute a sequence set, which is iteratively reduced until one of two stopping criteria is fulfilled. During each iteration, `FitSS4ASR` performs a phylogenetic analysis by means of `RAxML` or `MrBayes` based on an MSA created by means of `MAFFT` and `Gblocks`. During each iteration k, the output $Iter_k = \{tr_k, SEQ_k\}$ consisting of a consensus tree and a sequence set is stored for subsequent analysis. The topologies of the trees generated for each iteration k are further analyzed to identify "rogue" or/and "solitary" sequences, whose localization varies among the individual trees. These and sequences causing isolated subtrees with branches longer than 1 mutation / site are eliminated. **(B)** Assessing the robustness of tree topologies. Taking up to 15 representative datasets $Iter_k^+$, `FitSS4ASR` computes a supertree and eliminates all datasets with deviating trees. The remaining trees tr_k^* are subjected to a robustness analysis based on extended sequence sets, which contain additional sequences taken from the initial dataset SEQ_1 . Scores rating the robustness of the trees are saved for the final assessment by the user.

100 sequence sets SEQ_s^* , each of which consist of SEQ_s plus 10 randomly picked sequences chosen from the initial set SEQ_1 and computes the corresponding trees. For the subsequent tree comparison, our algorithm prunes the 100 trees to the sequences SEQ_s and uses the trees for the computation of a consensus tree tr_s^* . If the comparison of the tree topologies tr_s^* and tr_s indicates only minor differences, we consider tr_s robust to perturbations and SEQ_s suitable for ASR. As noted, `FitSS4ASR` utilizes two different methods for phylogenetic analysis and three for sequence elimination; thus, the final sets Alt_s may originate from any of these six combinations and characteristics are needed to assist a selection.

3.2.3 Choosing a Datasets for ASR

After program termination, the user has to choose from the m alternative datasets $Alt_s = \{tr_s, SEQ_s\}$ according to his needs. `FitSS4ASR` calculates five scores to support the user with his decision: Based on the score $tax\_num(Alt_s)$ (**Equation (3.1)**), the user can survey the phylogenetic coverage of the sequence set SEQ_s . Two scores assess the quality of tr_s : $branch\_distr(Alt_s)$ (**Equation (3.2)**) is a measure for the existence of exceedingly long branches longer than 1 mutation per site and $pp\_distr(Alt_s)$ (**Equation (3.3)**) indicates the "reliability" of branches near the root of the tree. $tr\_rob(Alt_s)$ (**Equation (3.4)**), summarizes the phylogenetic robustness of tr_s with respect to perturbations and $tr\_mf(Alt_s)$ (**Equation (3.5)**) penalizes the existence of multifurcations. We consider a dataset SEQ_s a good choice for ASR, it the phylogenetic coverage is sufficient and if all other scores are close to 1.0. For a first orientation, the user can compare the $ASR\_score(Alt_s)$ values, which are for each dataset the product of the latter four scores (**Equation (3.6)**).

3.2.4 Conventional Sequence Selection for ASR of GGGPS

In order to confirm further the efficacy of our approach, we performed in parallel a conventional and a `FitSS4ASR` assisted ASR of the enzyme GGGPS. Based on an extensive phylogenetic analysis, extant GGGPS sequences have been divided into group I and group II enzymes that differ with respect to phylogenetic origin and oligomerization states (Peterhoff et al., 2014). The reconstruction of a common ancestor of both groups is not feasible due to the length of the edge (> 4 mutations per site) that interconnects the nodes representing the ancestors of group I (AncGGGPS1) and group II (AncGGGPS2) enzymes. Moreover, whereas all group I enzymes oligomerize to dimers, group II enzymes form dimers or hexamers in a phylogeny dependent manner (Peterhoff et al., 2014). As we were interested to elucidate critical parameters of complex formation, we focused our analysis on group II enzymes. We started sequence selection with the analysis of a comprehensive and precompiled set $GGGPS_{initial}$, which consisted of 217 entries from InterPro family IPR008205 (version 67.0). To generate this set, the above-mentioned filters were applied; additionally, all clades represented by just one sequence were eliminated

by comparing `Key2ann` (Pürzer et al., 2011) annotations. This program replaces each sequence identifier with a human-readable annotation representing the phylogenetic lineage of the contributing species. We used these annotations to determine the phylogenetic diversity of this sequence set.

An MSA was computed by means of `MAFFT` (Katoh and Standley, 2013) and all columns containing more than 50 % gaps were eliminated by utilizing `Gblocks` (Castresana, 2000). For the resulting MSA, a first phylogenetic tree was deduced with `PhyloBayes` (Lartillot et al., 2009). Applying the above mentioned quality standards, we assessed the robustness and suitability of this and subsequently generated trees for ASR. It seems a simple task to pick a robust subset from not more than 217 sequences. However, nine rounds of optimization requiring the manual adaptation of the sequence set were needed. The resulting MSA *GGGPS2_man* consisting of 87 sequences gave rise to a phylogenetic tree that fulfilled all our robustness criteria. This tree was deduced from two MCMC chains and their maximum difference of posterior probabilities of tree bipartitions was 0.000243, which indicated high convergence. The MSA is listed in **Table 3.2** and the resulting phylogenetic tree is shown in **Figure 3.3** and **Table 3.4**.

3.2.5 Sequence Selection by Means of `FitSS4ASR` for an ASR of GGGPS

The sequences $GGGPS_{initial}$ used above for a conventional sequence selection were subjected to `FitSS4ASR`. The program converged after two rounds of iteration and the $ASR\_score(Alt_s)$ of the alternatives suggested to consider a specific set $Alt_s = \{tr_s, SEQ_s\}$. However, the recommended specific set Alt_2 contained not more than 58 sequences and just 1 crenarchaeal sequence, as `FitSS4ASR` does not preserve the phylogenetic diversity of the input. Thus, the crenarchaeal subset taken from $GGGPS_{initial}$ was added to SEQ_s and a further `FitSS4ASR` run was performed, which resulted in a final set *GGGPS2_auto* consisting of the 61 sequences listed in **Table 3.3**. An assessment of the resulting tree (**Figure 3.4** and **Table 3.5**), which was generated with minimal user interaction, confirms that it fulfills all criteria for ASR.

Moreover, a comparison of **Figures 3.3** and **3.4** makes clear that both trees possess a highly similar topology, which is consistent with the hierarchy of the phyla determined for the iTOL project (Ciccarelli et al., 2006). This finding testifies to the strong phylogenetic signal within the two sequence sets *GGGPS2_man* and *GGGPS2_auto*; interestingly, the two sets overlap by not more than 34 sequences. To sum up, `FitSS4ASR` was able to deduce a suitable phylogenetic tree for ASR highly comparable to the manual approach, which testifies the new approach.

Reconstruction of Ancestral GGGPS Sequences

GGGPS group II enzymes form dimers or hexamers in a phylogeny dependent manner and it is unknown when these oligomerization states arose. We wanted to follow the advent of these states for the full evolutionary interval dating back to the last ancestor AncGGGPS2.

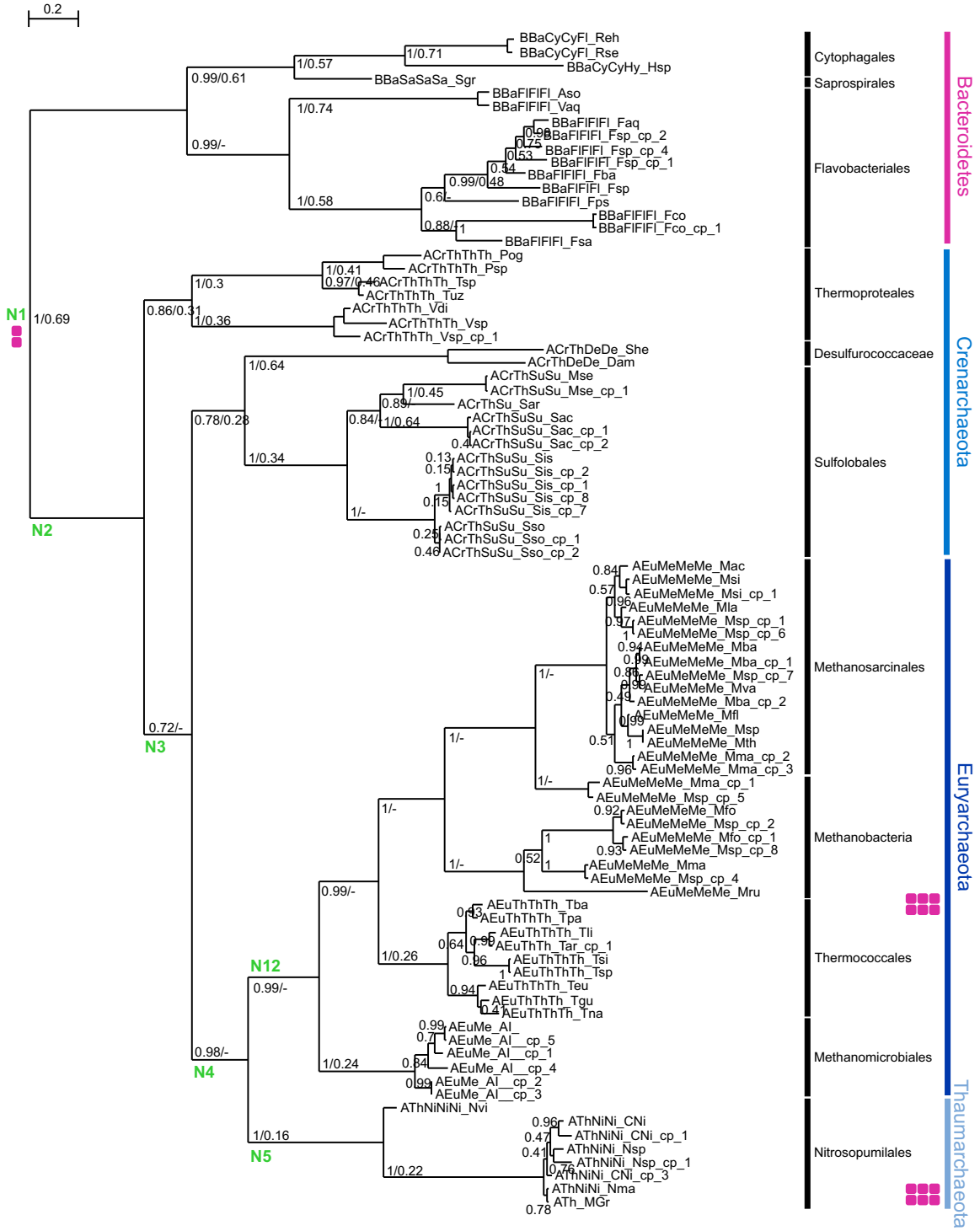

Figure 3.3: Phylogeny of the manually curated sequence set used for ASR of GGGPS predecessors. The 87 sequences of *GGGPS2_man* represent four major microbial phyla, namely Bacteroidetes, Crenarchaeota, Euryarchaeota, and Thaumarchaeota. All sequences were annotated by means of `Key2ann` (Pürzer et al., 2011) to indicate for each species the phylogenetic lineage detailing superkingdom, phylum, class, order, family, and species. For example, the label "ACrThDeDe_She" represents the species *Staphylothermus hellenicus* "_She", which is from the superkingdom Archaea "A", the phylum Crenarchaeota "Cr", the class Thermoprotei "Th", the order Desulfurococcales "De", and the family Desulfurococcaceae "De". The tree was computed by means of `PhyloBayes` and for central nodes, the posterior probability and a bipartition score is given, if available. The length of the horizontal bar corresponds to 0.2 substitutions per site. The ancestor AncGGGPS2_man is indicated by N1.

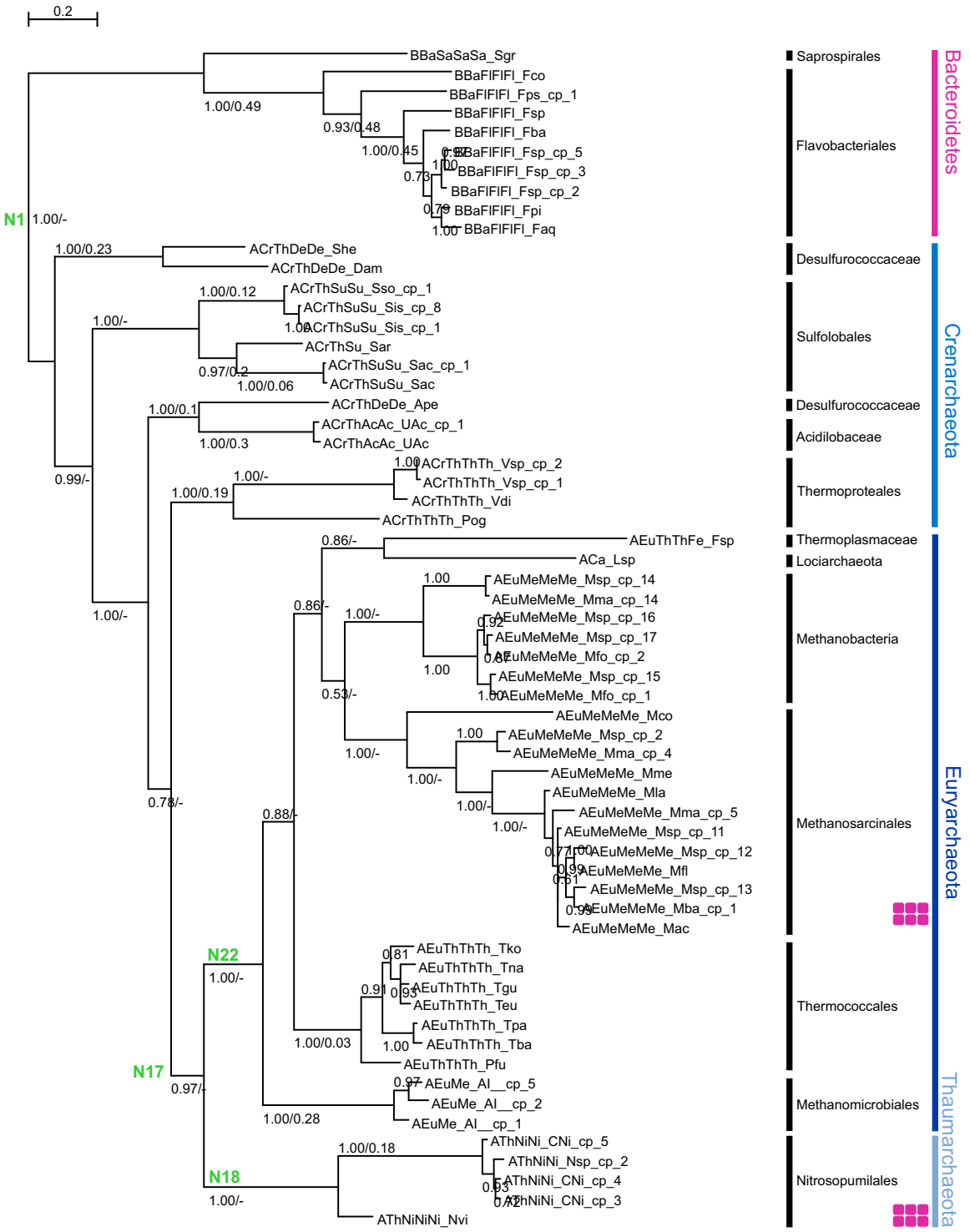

Figure 3.4: The phylogeny of the sequence set generated by means of `FitSS4ASR` for ASR of GGGPS predecessors. The 61 sequences of *GGGPS2_auto* represent four major microbial phyla, namely Bacteroidetes, Crenarchaeota, Euryarchaeota, and Thaumarchaeota. All sequences were annotated by means of `Key2ann` (Pürzer et al., 2011); see legend of **Figures 3.3**. The tree was computed by means of `MrBayes` and for central nodes, the posterior probability and a bipartition score is given, if available. The length of the horizontal bar corresponds to 0.2 substitutions per site. The ancestor AncGGGPS2_auto is indicated by N1.

Predecessors manual	semi-automatic	sequence identity (SeqId)
AncGGGPS2_N1_man	AncGGGPS2_N1_auto	75 %
AncGGGPS2_N4_man	AncGGGPS2_N17_auto	86 %
AncGGGPS2_N5_man	AncGGGPS2_N18_auto	93 %
AncGGGPS2_N12_man	AncGGGPS2_N22_auto	91 %

Table 3.1: Comparing predecessors from manual and semi-automatic approach by their SeqId. The SeqId of the most primordial sequences, AncGGGPS2_N1_man and AncGGGPS2_N1_auto is not more than 75 %. Sequence identity increases for intermediates with their distance to the AncGGGPS2 sequence.

Linde et al. (2018) showed that GGGPS from Euryarchaeota form hexamers and Peterhoff et al. (2014) predicted a hexameric form for GGGPS from Thaumarchaeota. In addition, experimental studies revealed that the reconstructed AncGGGPS2_N1_man forms a dimer (to be published). Thus, sequences representing the sequence of the last ancestor and intermediates AncGGGPS2-N* were reconstructed by means of `FastML` (Ashkenazy et al., 2012) constituting a path from the ancestor of AncGGGPS2* to the bifurcation into Euryarchaeota and Thaumarchaeota. To assess the robustness of our protocol, we utilized in parallel the sets *GGGPS2_man* and *GGGPS2_auto* for ASR (see Methods) and identified corresponding intermediates indicated in **Figures 3.3** and **3.4**. Due to small invariances between the two topologies, only four corresponding intermediates were identified on the path.

First, we compared the sequences of the predecessors. The ancestor AncGGGPS2_man consists of 246 amino acids and shares 65 % sequence identity (SeqId) with the most similar extant sequence ACrThThTh_Tuz, which is from *Thermoproteus uzoniensis*. The ancestor AncGGGPS2_auto possesses 242 amino acids and shares 72 % SeqId with the most similar extant sequence ACrThDeDe_She from *Staphylothermus hellenicus*. AncGGGPS2_man and AncGGGPS2_auto share 75 % identical residues and of the 45 differences, 36 were exchanges of similar residues. In addition, the three further identified pairs of ancestral sequences were compared (**Table 3.1**). Each of these sequence pairs show high SeqId between 86 % and 93 %. A trend of slightly decreasing SeqId can be identified when travelling back in time. The overall high similarity between these pairs reflects the congruence between the topologies from both approaches.

3.3 Discussion

3.3.1 ASR Requires a Strong Phylogenetic Signal Necessitating a Rigorous Preselection of Sequences

Often, the analysis of large datasets is regarded as valuable for the recovery of statistically well-supported and "true" phylogenies. However, it is known that the analysis of large datasets under optimal models of sequence evolution does not guarantee robust phylogenetic inference (Ho and Jermiin, 2004; Rodríguez-Ezpeleta et al., 2007; Salichos and Rokas, 2013). Moreover, the misleading effects of certain biases are correlated with the size of a dataset (Lartillot and Philippe, 2004). One notoriously observed bias is long-branch attraction (LBA), which leads to a clustering of taxa with high evolutionary rates (long branches) regardless of the phylogenetic relatedness. LBA is caused by strong violations of phylogenetic model assumptions due to highly heterogeneous evolutionary rates within some lineages. To overcome this problem, it was proposed to eliminate fast-evolving taxa (Stefanović et al., 2004; Rivera-Rivera and Montoya-Burgos, 2016) or fast-evolving genes from multi-gene datasets (Brinkmann et al., 2005) and algorithms like `Phylo-MCOA` can detect outlier genes and species by comparing the topologies produced by individual genes (de Vienne et al., 2012). However, these methods often necessitate the parallel analysis of several datasets. To reach highest flexibility, we focused on elimination methods that need for outlier detection not more than the dataset and trees required for the intended ASR. The inspection of suboptimal trees provides insight into the interplay among conflicting phylogenetic signals (Swofford et al., 1996) and to reduce them we integrated the elimination of rogue and solitary sequences. The scores determined by `FitSS4ASR` for the assessment of carefully compiled datasets support the user in his decision, which should be more than the blind reliance on optimality criteria and should also consider contradictory factors adequately (Ho and Jermiin, 2004). For example, as we demonstrated for the GGGPS reconstruction, it might be necessary to add manually sequences to broaden the phylogenetic basis.

Interestingly, the two sequence sets *GGGPS2_man* and *GGGPS2_auto* overlap by not more than 34 sequences and the two ancestors AncGGGPS2_man and AncGGGPS2_auto share not more than 75 % identical residues. In contrast to the two most primordial ancestors, the sequences of the predecessors along the path to the extant Euryarchaeota and Thaumarchaeota are highly similar. Thus, identical phenotypes are expected for corresponding pairs. Since the oligomerization state of GGGPS is a delicate property that changes between dimeric and hexameric in a phylogeny-dependent manner (Peterhoff et al., 2014), the pairwise assessment of the oligomerization states will indicate the suitability of `FitSS4ASR` in more detail. These experiments are not finished, but the accordance of phylogenetic trees and reconstructed sequences argue in favor of highly similar phenotypes.

3.3.2 Future Directions

We restricted the function of `FitSS4ASR` to sequence selection and implemented two filters for the elimination of sequences with a weak phylogenetic signal. The integration of further methods for sequence elimination is feasible, which could be based on the length of individual branches or novel methods (Lemoine et al., 2018) assessing their robustness. One limitation of `FitSS4ASR` is its blindness against the phylogenetic diversity of the chosen sequences, which must be extended manually in case of a too narrow phylogenetic representation. This interaction is required, because we have so far no clue for an appropriate algorithm.

We consider the integration of the full ASR protocol into `FitSS4ASR` less useful, due to the various demands of individual reconstruction projects. For example, different ASR methods and models are in use and just recently, `SubRecon` was introduced for the investigation of substitutions on a single branch of interest (Goldstein and Kelso, 2018). However, irrespective of the ASR problem to be solved, a sequence set has to be selected beforehand and thus the current implementation of `FitSS4ASR` helps to increase the effectiveness of any ASR protocol.

3.4 Materials and Methods

3.4.1 Conventional ASR Protocol

`Jalview` (Waterhouse et al., 2009) was used for sequence comparison, redundancy filtering and MSA generation. Phylogenetic trees were computed with the help of `PhyloBayes3.0` (Lartillot et al., 2009) utilizing a time-homogenous `CAT` model and by launching 4 independent MCMC samplings of length 50,000 to ensure convergence. For the final dataset, the consensus tree was deduced by concatenating two chains after a burn-in of 6000 trees.

3.4.2 `FitSS4ASR`, a Semi-supervised Protocol for Sequence Selection

Sequence elimination: The protocol implemented with `FitSS4ASR` reduced successively the content of a given set $SEQ = \{seq_1, ..., seq_n\}$ of n input sequences and ended after t rounds of iterations, if one of two stopping criteria was fulfilled; compare **Figure 3.2**. To begin with, the phylogenetic origin of the n input sequences was determined by means of `Key2ann` (Pürzer et al., 2011), which replaced each sequence identifier with a human-readable annotation representing the phylogenetic lineage of the contributing species. `cd-hit` (Li and Godzik, 2006) was used to reduce sequence redundancy and to eliminate identical sequences. Sequences deviating in length by more than 2σ from the mean were removed. For the resulting sequences, an initial MSA was computed with default parameters by means of `MAFFT` (Katoh and Standley, 2013), which exhibited best performance in ASR applications (Vialle et al., 2018). Sequences introducing

internal gaps with a minimal length of 5 in at least 90 % of the sequences were removed; the resulting sequence set $SEQ_{k=1}$ contained $m = |SEQ_k|$ sequences. For subsequent analysis, the sequences were realigned and columns containing more than 50 % gap symbols were eliminated by utilizing `Gblocks` (Castresana, 2000). Two `MAFFT` phylogenetic analyses were performed by means of `RAxML` or `MrBayes`. The `RAxML` option `-f a` and the substitution model `PROTGAMMAAUTO` were used to compute 100 trees and a consensus tree. The `MrBayes` option `sumt` and the `gtr` and `invgamma` model were used to generate 1,000,000 trees and a consensus tree by means of two chains. During each iteration, the dataset $Iter_k = \{tr_k, SEQ_k\}$ consisting of the consensus tree and the sequences was stored for the subsequent analysis. Computation ended, if the sequence elimination step described below did not further reduce the content of SEQ_k , i. e. $m = |SEQ_k|$ was constant for 10 rounds or if m was ≤ 60. Otherwise, sequence elimination was initiated.

For sequence elimination, the tree list generated during each iteration k was further analyzed by means of `RogueNaRok` (Aberer et al., 2012) to identify "rogue" sequences. Additionally, for each seq_k^r all sister sequences seq_k^s were identified in all trees tr_k^i of the tree list. These occurrences were summed up in a matrix $sis_k[r, s]$ and normalized for each r to identify "solitary" sequences for which holds $sis_k[r, s] < 0.75 \, \forall s$. Depending on the chosen selection parameter, either one of the elimination methods or a combination of both was applied to eliminate rogue or/and solitary sequences. The remaining sequences constituted the set SEQ_{k+1} , if a further iteration k+1 was executed.

Choosing and assessing the phylogenetic robustness of sequence sets: Taking the output of the last iteration $Iter_{last} = \{tr_{last}, SEQ_{last}\}$ that consisted of $u = |SEQ_{last}|$ sequences as a reference, up to 15 datasets $Iter_k^+$ were selected that contained approximately evenly distributed between u and maximally 500 sequences. Using the trees tr_k^+ , a supertree was computed with the help of `PluMiST` (Kupczok, 2011) and default parameters. The deviation from this supertree was determined for all tr_k^+ by means of the `bitstring` method implemented in the `Bio:Phylo` package (Talevich et al., 2012) and all trees classified as dissimilar were discarded. The m remaining sets $Alt_{s,s=1..m} = \{tr_s, SEQ_s\}$ were chosen for further analysis of tree robustness.

For each dataset $Alt_s = \{tr_s, SEQ_s\}$, 100 trees were created by means of `MrBayes` (parameters as above), which were based on sequence sets SEQ_s^* that consisted of SEQ_s plus 10 randomly selected sequences chosen from the initial data set SEQ_1 . These 100 trees were pruned to the data set SEQ_s and a consensus tree tr_s^* was determined by means of the `prune` and `consensus` method implemented in the `Bio:Phylo` package (Talevich et al., 2012).

3.4.3 Indicators of ASR Suitability

`FitSS4ASR` lists for each of the final datasets $Alt_s = \{tr_s, SEQ_s\}$ the following parameters:

$$tax\_num(Alt_s) = \#\ species\ to\ be\ found\ in\ SEQ_s \tag{3.1}$$

$tax\_num(Alt_s)$ is the number of species contributing to the respective sequence set.

$$branch\_distr(Alt_s) = \#\ branches\ shorter\ than\ 1\ mutation\ per\ site\ /\#\ all\ branches \tag{3.2}$$

The value of $branch\_distr(Alt_s)$ is 1.0, if all branches are shorter than 1 mutation per site and decreases with the number of exceedingly long branches.

$$pp\_distr(Alt_s) = \sum_{v, pp(v) \geq 0.75} pp(v) \cdot dist(root, v) \Big/ \sum_{v} pp(v) \cdot dist(root, v) \tag{3.3}$$

Here, $pp(v)$ is the posterior probability of branch v and $dist(root, v)$ is the distance of v from the root, i. e. the number of branches. The value of $pp\_distr(Alt_s)$ reaches 1.0, if all branches near the root are well supported by posterior probabilities of at least 0.75.

$$tr\_rob(Alt_s) = \ the\ fraction\ of\ shared\ bipartitions \tag{3.4}$$

The value of $tr\_rob(Alt_s)$ is 1.0, if tr_s and the consensus tree tr_s^* resulting from our perturbation approach are identical and decreases with the number of differing bipartitions determined by means of `bitstring`.

$$tr\_mf(Alt_s) = 1/(1 + \ the\ number\ of\ multifurcations) \tag{3.5}$$

The value of $tr\_mf(Alt_s)$ is 1.0, if tr_s does not contain a multifurcation which are introduced by some programs during the computation of consensus trees and impede ASR.

$$ASR\_score(Alt_s) = branch\_distr(Alt_s) \cdot pp\_distr(Alt_s) \cdot tr\_rob(Alt_s) \cdot tr\_mf(Alt_s) \tag{3.6}$$

$ASR\_score(Alt_s)$ is close to 1.0, if a tree fulfils all stability parameters.

3.4.4 Ancestral Sequence Reconstruction

`NJplot` (Perriere and Gouy, 1996) was used for midpoint rooting a phylogenetic tree between Bacteria and Archaea. Ancestral sequences that may contain indels, were computed by means

of the marginal reconstruction approach of `FastML` (Ashkenazy et al., 2012), the substitution model `JTT`, and a gamma distribution. In order to adjust the length of reconstructed and of recent sequences, the parameter `probability cutoff to prefer ancestral indel over character` was set to 0.8 (manual approach) or 0.9 (`FitSS4ASR`) in order to compensate a bias towards longer than true ancestors (Vialle et al., 2018). For each internal node of the tree, the most probable sequence was determined.

3.5 Supplemental Figures and Tables

Table 3.2: MSA consisting of the 87 sequences of *GGGPS2_man* and reconstructed predecessors in FASTA format. Sequences can be found as Digital Supplemental Data on the provided data storage medium.

Table 3.3: MSA consisting of the 61 sequences of *GGGPS2_auto* and reconstructed predecessors in FASTA format. Sequences can be found as Digital Supplemental Data on the provided data storage medium.

Table 3.4: Phylogenetic tree deduced for *GGGPS2_man* in Newick format. Phylogenetic tree can be found as Digital Supplemental Data on the provided data storage medium.

Table 3.5: Phylogenetic tree deduced for *GGGPS2_auto* in Newick format. Phylogenetic tree can be found as Digital Supplemental Data on the provided data storage medium.

Chapter 4

The Ancient Nature of Allostery and Substrate Channeling in the Tryptophan Synthase Complex

Florian Busch, Chitra Rajendran, Kristina Heyn, Sandra Schlee, Rainer Merkl, and Reinhard Sterner

Cell chemical biology, 23(6), 709-715.

Summary

Modern enzyme complexes are characterized by a high catalytic efficiency and allosteric communication between the constituting protein subunits. We were interested whether primordial enzyme complexes from extinct species displayed a similar degree of functional sophistication. To this end, we used ancestral sequence reconstruction to resurrect the α- and β-subunits of the tryptophan synthase (TS) complex from the last bacterial common ancestor (LBCA), which presumably existed more than 3.4 billion years ago. We show that the LBCA TS subunits are thermostable and exhibit high catalytic activity. Moreover, they form a complex with $\alpha\beta\beta\alpha$ stoichiometry whose crystal structure is similar to the structure of modern TS. Kinetic analysis revealed that the reaction intermediate indole is channeled from the α- to the β-subunits and suggests that allosteric communication already occurred in LBCA TS.

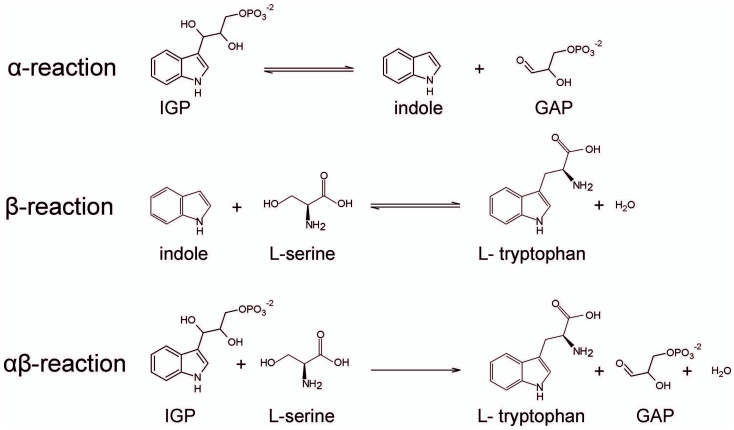

Figure 4.1: Reactions catalyzed by the α-subunit (α-reaction), the β-subunit (β-reaction), and the TS complex ($\alpha\beta$-reaction). IGP: indole-3-glycerol-phosphate; GAP: glyceraldehyde-3-phosphate.

4.1 Introduction

Metabolic pathways comprise a series of reactions that are catalyzed by different enzymes, some of which assemble to complexes. Several of those complexes feature a high level of sophistication, which includes substrate channeling and allosteric communication between the constituting protein subunits (Huang et al., 2001; Raushel et al., 2003). One prominent example is the $\alpha\beta\beta\alpha$ tryptophan synthase complex (TS), which is responsible for the last two steps in tryptophan biosynthesis (**Figure 4.1**): The α-subunit catalyzes the aldolytic cleavage of indole-3-glycerol-phosphate (IGP) to glyceraldehyde-3-phosphate (GAP) and indole. The latter is transferred via an intermolecular 25 Å long hydrophobic tunnel to the active site of the adjacent β-subunit (Hyde et al., 1988), where it condenses with L-serine in a pyridoxal phosphate (PLP) -dependent reaction to L-tryptophan and water. In order to keep the two reactions in phase, the α- and β-subunits reciprocally activate each other (Casino et al., 2007; Dunn, 2012; Miles, 2001). Most investigations have focused on the TS from *Salmonella typhimurium* (stTS) and *Escherichia coli* (ecTS) and thus highlight molecular mechanisms which have been developing over billion years of evolution. In contrast, little is known about allosteric interactions and substrate channeling in earlier phases of life, due to the lack of macromolecular fossils from ancient proteins.

However, recent advances in computational biology allow one to resurrect primordial proteins using a technique named ancestral sequence reconstruction (ASR). Since the pioneering study by Benner and co-workers on ribonucleases (Stackhouse et al., 1990), this method has been successfully applied to gain insights into the structure, thermostability, catalytic activity, conformation flexibility, and assembly properties of enzymes from extinct organisms (Liberles, 2007; Merkl and Sterner, 2016; Risso et al., 2014; Thornton, 2004). For example, we have recently shown that the reconstructed cyclase subunit HisF of the imidazole glycerol phosphate synthase from the last universal common ancestor (LUCA) of cellular organisms is catalytically active

and capable of interacting with the extant glutaminase subunit HisH from *Zymomonas mobilis*. The presence of allosteric communication and substrate channeling within the formed glutamine amidotransferase complex gave first indication that sophisticated complexes might have existed already several billion years ago. However, the sequence heterogeneity within the HisH enzyme family impeded the reconstruction of a catalytically active LUCA-HisH. As a consequence, it has remained impossible to study the properties of the LUCA-HisF/LUCA-HisH complex or of other ancient complexes (Reisinger et al., 2014b). We have now set out to close this knowledge gap by using the TS complex as a test case. For this purpose, the α- and β-subunits of the LBCA, which existed more than 3.4 billion years ago (Battistuzzi et al., 2004), were resurrected by ASR and characterized. The results show that the LBCA TS complex was a highly efficient enzyme, quite unalike the unspecific generalists that presumably dominated the early phases of biological evolution (Jensen, 1976).

4.2 Results and Discussion

4.2.1 Sequence Reconstruction of LBCA TS Subunits

For ASR, we used a representative set of 52 concatenated α- and β-subunit sequences, which was derived from organisms of the bacterial phyla of Chloroflexi, Deinococci, Nitrospirae, Verrucomicrobia, Proteobacteria and Firmicutes, and the archaeal phylum of Euryarchaeota. Other archaeal phyla were not included as they contain a different type of TS β-subunit (Busch et al., 2014; Merkl, 2007). A maximum-likelihood tree was created based on the most probable substitution model (**Figure 4.5**). We rooted the tree within the bacteria, because Euryarchaeota have most likely obtained the TS by a more recent horizontal gene transfer event from a bacterial predecessor (Merkl, 2007). The hierarchy of Firmicutes, Chloroflexi, Deinococcales, and Proteobacteria within the constructed TS tree is consistent with their relationship determined for the iTOL project, which lacks Nitrospirae and Verrucomicrobia, however (Ciccarelli et al., 2006). The deduced α- and β-subunit sequences (**Figure 4.6**) at the root correspond to those of the last common ancestor of bacteria (LBCA) (**Figure 4.5**). The reconstructed LBCA TS α-subunit shares a sequence identity of no more than 57 % with its closest extant relative from *Clostridium arbusti* and the reconstructed LBCA TS β-subunit shares a sequence identity of 78 % with its closest extant relative from *Caldanaerobacter subterraneus*. This is consistent with the generally lower overall sequence conservation within the family of TS α-subunits as compared to the family of TS β-subunits.

The relevance of our study critically depends on the reliability of the reconstructed ancestral sequences that represent the most probable ancestor (MPA) of LBCA TS. Due to limitations of the underlying evolutionary models, ASR is unavoidably uncertain to some extent. Critical parameters of ASR are the topology of the tree and the length of individual edges. However, it

has been show that the algorithms of ASR tolerate a certain degree of phylogenetic uncertainty (Hanson-Smith et al., 2010). Notwithstanding, for the tree used here the posterior probabilities of only four of the 101 edges are below 0.99 and the lowest value is 0.92, which testifies to a highly reliable topology (**Figure 4.5**). Moreover, the length of the longest edge and the mean edge length correspond to only 0.68 and 0.27 mutations per site, respectively, further supporting the fidelity of the reconstruction. For these reasons and because the most probable LBCA TS was stable and active (see below), no alternative and somewhat less likely ancestral TS complexes were produced and characterized. The characterization of such ensembles with similar sequences as the MPA (Bar-Rogovsky et al., 2015) might be advisable, however, in the case of precarious phylogenetic trees.

4.2.2 Stabilities of LBCA TS Subunits and Subunit Interaction

The genes that encode the LBCA α- and β-subunits were synthesized and expressed in *Escherichia coli*. Thermal denaturation of the purified subunits was monitored by circular dichroism spectroscopy and differential scanning calorimetry. The results showed that the LBCA α-subunit unfolds in two steps (T_m = 71 °C and 102 °C) and that the LBCA β-subunit unfolds in one step (T_m = 99 °C). Both subunits do not aggregate up to 115 °C as indicated by the symmetry of unfolding signals (**Figure 4.7**).

We analyzed the association states of the LBCA TS subunits and their interaction by analytical size exclusion chromatography. According to the determined molecular weights (MR), the reconstructed proteins have the same oligomeric state as extant α- and β-subunits: The LBCA TS α-subunit is a monomer (determined MR: 33.5 kDa; expected MR: 29.9 kDa) and the LBCA TS β-subunit is a dimer (determined MR: 80.4 kDa; expected MR: 87.2 kDa) in solution. The determined molecular weight of the LBCA TS is compatible with a $\alpha_2\beta_2$ complex (determined MR: 176.3 kDa; expected MR: 147.0 kDa) (**Figure 4.2 A**). Fluorescence titration experiments unambiguously confirmed the 1:1 subunit stoichiometry in the LBCA TS and testified to a tight interaction with a thermodynamic dissociation constant in the nanomolar range (**Figure 4.2 B**). Moreover, the UV/Vis properties of LBCA TS, which are a sensitive measure for the operating mode of the complex, are identical to those of stTS (Schiaretti et al., 2004): the spectra of LBCA TS, LBCA TS in the presence of 10 mM L-serine and LBCA TS in the presence of 1 mM L-tryptophan show maxima at 412 nm (as described for the internal aldimine in stTS), 350 nm (as described for the α-aminoacrylate in stTS), and 476 nm (as described for the tryptophan quinonoid in stTS), respectively.

4.2.3 Crystal Structure and Substrate Channeling of LBCA TS

We crystallized the LBCA TS complex in presence of the α-subunit ligand glycerol 3-phosphate (GP) and the β-subunit substrate L-serine. The structure was solved at 1.97 Å resolution by

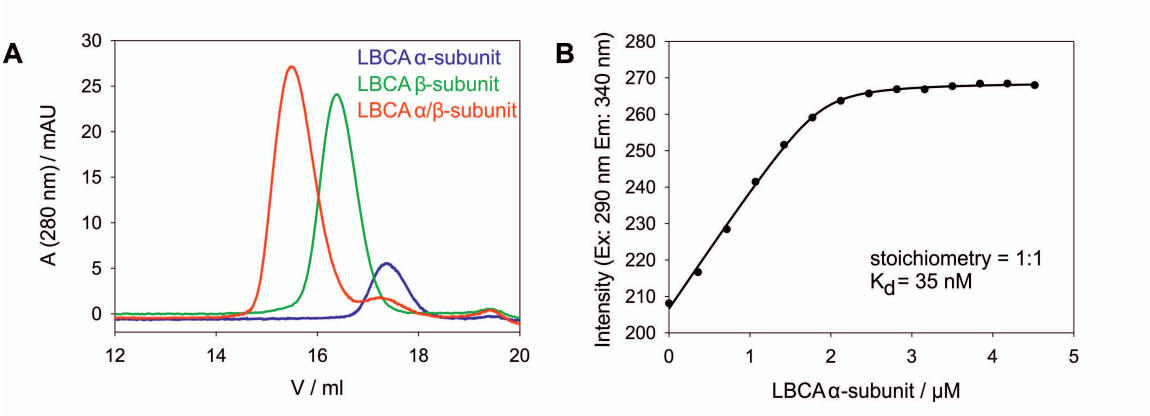

Figure 4.2: Assembly of LBCA α- and β-subunits to the TS complex. (A) Analytical size exclusion chromatograms of LBCA α-subunit, LBCA β-subunit, and a LBCA TS α/β-subunit mixture (10 μM monomer concentration in each case). The subunits were applied on a S200 analytical column equilibrated with 50 mM potassium phosphate pH 7.5 and 300 mM KCl. Elution was performed with a flow rate of 0.5 ml/min at 25 °C and monitored by measuring the absorbance at 280 nm. **(B)** Fluorescence titration of 2 μM LBCA β-subunit (monomer concentration) with LBCA α-subunit in 10 mM potassium phosphate pH 7.5 at 25 °C. Fluorescence emission at 340 nm was determined following excitation at 290 nm. Data points were fitted with a quadratic function.

molecular replacement using stTS as template (**Figure 4.4**). We found that GP is bound at both α-subunits and the cofactor PLP is bound at both β-subunits. The LBCA TS structure superimposes with the structure of the GP-bound stTS (PDB: 1WBJ) with an overall root-mean-square deviation of less than 1 Å. Like in stTS, the active sites of the LBCA TS subunits seem to be connected by a hydrophobic channel that is the pre-requisite for indole channeling (**Figure 4.3**).

We rechecked the existence of this channel by using a biochemical assay that is based on the consolidated knowledge that large externally added nucleophiles approach the β-site via the channel (Dunn et al., 1990). The presence of the α-subunit and its ligand GP confer blockage of the entrance to the channel and thereby decrease the rates at which those bulky nucleophiles react at the β-site. Indeed, we found that the reaction rate with which the large nucleophile benzimidazole displaces the preformed aniline quinonoid at the β-site ($k_1 = 580\,s^{-1}$; $k_2 = 27.2\,s^{-1}$) is decreased by complex formation with the α-subunit ($k = 14.0\ s^{-1}$). This effect is more pronounced in the presence of GP ($k = 0.094\ s^{-1}$) (**Figure 4.8 A**). In contrast, the reaction rate of the small nucleophile N-methylhydroxylamine with the preformed aminoacrylate was only slightly affected by the presence of the α-subunit and by GP ($k = 12.7 - 49.1\ s^{-1}$) (**Figure 4.8 B**). Taken together, these results provide strong evidence for the existence of a hydrophobic channel connecting the active sites of the LBCA α- and β-subunits as found in extant TS (Dunn et al., 1990).

The detailed analysis of the α/β interface showed that fourteen inter-subunit hydrogen bonds are identical between stTS and LBCA TS whereas two H-bonds are specific for LBCA and

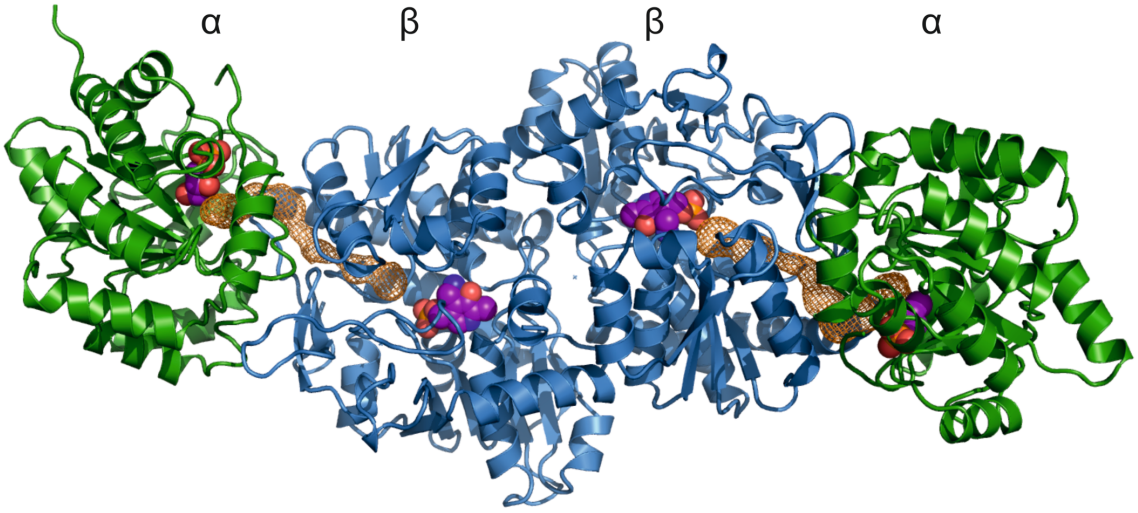

Figure 4.3: Crystal structure of the LBCA TS complex (PDB: 5ey5). The α-subunits are green-colored and the β-subunits are blue-colored. Subunits are shown as cartoon diagrams, and ligands and cofactors are shown as spheres. Glycerol 3-phosphate is bound at α, the cofactor PLP is bound at β. The putative indole channel connecting the active site of the α -subunit with the active site of the β-subunit was visualized with `MOLE` (Sehnal et al., 2013) as an orange mesh.

eight H-bonds are specific for stTS (**Figure 4.9**). For example, the H-bond between βSer178 and αGly181 in stTS, which has been shown to be involved in allosteric communication (Raboni et al., 2005), is not present in the LBCA TS (**Figure 4.4**). Moreover, an H-bond network in the functionally important monovalent cation binding loop (MVC) of stTS (Dierkers et al., 2009) involving βArg141 and βAsp305 is missing in the β-subunit of LBCA.

4.2.4 Impact of the β-subunit for the Catalytic Efficiency of the α-subunit

We next investigated enzymatic activity and allosteric effects by monitoring the aldolytic cleavage of IGP (α-reaction) and the condensation of L-serine with indole (β-reaction), both with isolated and complexed subunits. Moreover, we followed the physiologically relevant coupled $\alpha\beta$-reaction (**Figure 4.1**).

The steady-state kinetic parameters for the α-reaction of the isolated subunit and of the subunit within the LBCA TS were determined by a continuous assay. The steady-state kinetic parameters for the $\alpha\beta$-reaction were monitored by a HPLC-based discontinuous assay to unambiguously differentiate between the formation of indole and L-tryptophan. We found that the catalytic efficiency for the aldolytic cleavage of IGP ($k_{cat}/K_M{}^{IGP}$) is approximately 60-fold enhanced for the complexed α-subunit compared to the isolated α-subunit, mainly due to an increase in k_{cat}. For ecTS, the corresponding increase in $k_{cat}/K_M{}^{IGP}$ is 145-fold (**Table 4.1**). We further observed that the $k_{cat}/K_M{}^{IGP}$ for the physiological $\alpha\beta$-reaction is increased 1.2-fold compared to the α-reaction within the LBCA complex. For ecTS, the corresponding increase

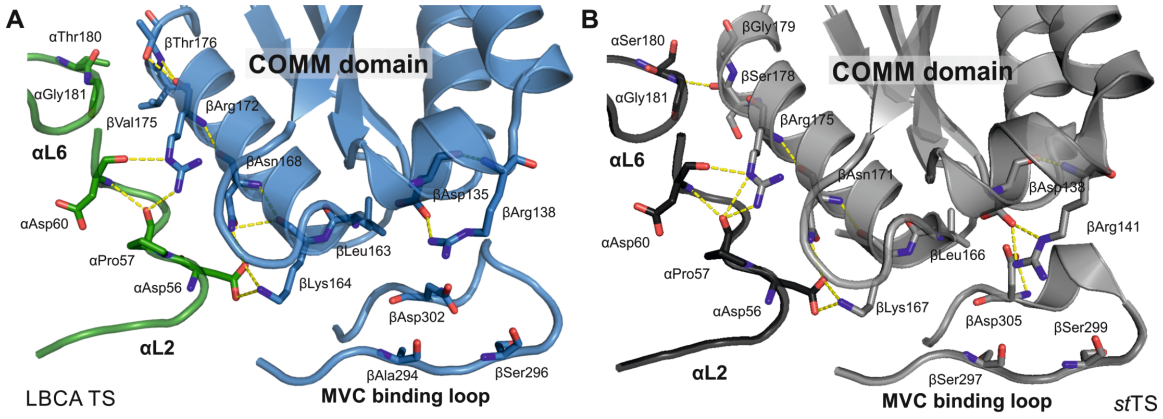

Figure 4.4: Comparison of H-bonds between LBCA TS and stTS. (A) Interface of the LBCA TS complex. The α-subunit is colored in green and the β-subunit is colored in blue. **(B)** Interface of the stTS complex. The α-subunit is colored in dark grey and the β-subunit is colored in light gray. Side-chains of interface residues are shown as sticks and hydrogen-bonds are indicated by yellow dashed lines. Loop 2 and loop 6 of the α-subunit (αL2 and αL6) as well as the communication (COMM) domain (Schneider et al., 1998) and the monovalent cation binding loop (MVC) of the β-subunit are indicated.

	LBCA TS			**ecTS**
	k_{cat} (s^{-1})	K_m^{IGP} (mM)	k_{cat}/K_m^{IGP} (s^{-1} M^{-1})	k_{cat}/K_m^{IGP} (s^{-1} M^{-1})
α	0.022	0.21	$1.0 \cdot 10^2$	$3.3 \cdot 10^0$
***α* in complex**	0.51	0.083	$6.1 \cdot 10^3$	$4.8 \cdot 10^2$
αβ	0.18	0.024	$7.5 \cdot 10^3$	$2.0 \cdot 10^4$

Table 4.1: Steady-state enzymatic parameters for the α-reaction of LBCA TS and ecTS. The listed values of LBCA TS were determined at 60 °C; they are the average of two experiments and the deviations were less than 30 %. Data for ecTS were gained from measurements performed at 25 °C (Hettwer and Sterner, 2002).

of k_{cat}/K_M^{IGP} is 42-fold (**Table 4.1**). These findings suggest that the extent to which the β-subunit and the presence of L-serine enhance IGP cleavage in the α-subunit of stTS (Anderson et al., 1991) and ecTS (Lane and Kirschner, 1991) might have been somewhat less pronounced in the LBCA era.

4.2.5 Impact of the α-subunit for the Catalytic Efficiency of the β-subunit

As for the α-subunit, the steady-state kinetic parameters for the β-reaction of the isolated subunit and of the subunit within the LBCA TS were determined by a continuous assay, and the steady-state kinetic parameters for the $\alpha\beta$-reaction were determined by a discontinuous assay. We found that the catalytic efficiency for the turnover of L-serine under saturating conditions for indole ($k_{cat}/K_M^{L\text{-serine}}$) is, unexpectedly, almost five-fold reduced for the β-subunit within the LBCA TS compared to the isolated β-subunit, due to a decrease of the turnover number.

	LBCA TS			**ecTS**
	k_{cat} (s^{-1})	K_m L-serine (mM)	k_{cat}/K_m L-serine (s^{-1} M^{-1})	k_{cat}/K_m L-serine (s^{-1} M^{-1})
β	4.8	1.3	$3.7 \cdot 10^3$	$1.9 \cdot 10^2$
***β* in complex**	0.95	1.2	$7.9 \cdot 10^2$	$1.2 \cdot 10^4$
αβ	0.15	0.025	$6.0 \cdot 10^3$	$4.1 \cdot 10^3$

Table 4.2: Steady-state enzymatic parameters for the *β*-reaction of LBCA TS and ecTS. The listed values of LBCA TS were determined at 60 °C; they are the average of two experiments and the deviations were less than 30 %. Data for ecTS were gained from measurements performed at 25 °C (Hettwer and Sterner, 2002).

In contrast, in ecTS a 63-fold increase of k_{cat}/K_M L-serine is observed for the β-subunit upon complex formation (**Table 4.2**). We further found that in LBCA TS, k_{cat}/K_M L-serine for the physiological $\alpha\beta$-reaction is almost eight-fold higher than k_{cat}/K_M L-serine for the complexed β-subunit, due to a drastic decrease of the Michaelis constant for L-serine. A different picture emerges for ecTS where k_{cat}/K_M L-serine for the physiological $\alpha\beta$-reaction is almost three-fold lower than k_{cat}/K_M L-serine for the complexed β-subunit (**Table 4.2**). Taken together, in both LBCA TS and ecTS the α-subunit and its substrate IGP influence the catalytic activity of the corresponding β-subunit, albeit in a different manner: Whereas the α-subunit accelerates the reaction of the β-subunit in ecTS and decelerates it in LBCA TS, the inverse effect is observed for IGP.

Allosteric signals are assumed to be transmitted in modern proteins through several, pre-existing pathways and mutations are thought to alter only the contribution of pathway ensembles and not to create new pathways (del Sol et al., 2009). It might be that their fine-tuning was optimized only at a later stage of TS evolution, although we cannot exclude that the slight differences in allostery between modern TS and LCBA TS are caused by inevitable minor ambiguities in inferring ancestral sequences. In any case, our findings strongly suggest that such pathway ensembles existed already in early phases of bacterial speciation.

4.3 Significance

It is well established that modern enzyme complexes are characterized by high catalytic efficiencies as well as structural and functional interactions between the constituting protein subunits. However, very little is known about ancient enzyme complexes from extinct species, due to the lack of macromolecular fossils. We have used ancestral sequence reconstruction to resurrect the α- and β-subunits of a primordial tryptophan synthase (TS) complex from the LBCA. The LBCA TS subunits, which were produced in *Escherichia coli*, formed a hetero-tetrameric $\alpha\beta\beta\alpha$ complex that was characterized by high catalytic activity, the channeling of a reaction inter-

mediate, and allosteric interactions. These sophisticated properties of the reconstructed LBCA TS suggest that enzyme complexes from the LBCA era (more than 3.4 billion years ago) were no longer the primitive and promiscuous generalists that presumably dominated the very early phases of biological evolution (more than 4 billion years ago). Instead, LBCA complexes were presumably already as sophisticated biocatalysts as their modern descendants found in extant organisms.

4.4 Experimental Procedures

4.4.1 Sequence Reconstruction

`BLAST` (Altschul et al., 1997) and the nr database of the NCBI were used to search for α- and β-subunit sequences. For ASR, we used a representative set of 52 concatenated sequences, which were derived from organisms of the bacterial phyla of Chloroflexi, Deinococci, Nitrospirae, Verrucomicrobia, Proteobacteria and Firmicutes, and the archaeal phylum of Euryarchaeota. `MAFFT` (Katoh and Standley, 2013) was used to generate a multiple sequence alignment (MSA; **Table 4.5**). The program `pb` (version 3.3 of `PhyloBayes`, (Lartillot et al., 2009)) with options `-cat -gtr` was used to compute in four independent Monte Carlo Markov Chains 50 000 samples each. The options `-cat -gtr` induce an infinite mixture model, whose components differ by their equilibrium frequencies. For computing this phylogenetic tree that guides the subsequent reconstruction, positions containing more than 50 % gaps were removed by using `GBlocks` (Castresana, 2000). The quality of mixing was assessed by computing for each pair of chains the discrepancy index (*maxdiff*, which was `<` 0.03 in all cases) by means of `bpcomp` and the *minimum effective size* (which was `>` 200 in all cases) with `tracecomp`. These results indicate the convergence of the chains. A consensus tree was determined by means of `readpb`, the burnin was 5000. The resulting tree for the concatenated subunits is depicted in **Figure 4.5**. We used the phylogenetic-aware gap placement algorithm `PRANK` (Löytynoja and Goldman, 2008) with the option `-showanc` and the full MSA to deduce the ancient subunit sequences. The amino acid sequences and nucleotide sequences as optimized for codon-usage in *E. coli* are shown in **Figure 4.6**.

4.4.2 Cloning and Expression

Genes were synthesized by Life Technologies and cloned into pET21a(+) at the *Nde*I/*Xho*I restriction sites to allow for the expression of proteins with a C-terminal His_6-tag. *E. coli* (DE3) cells were transformed with pET21a(+)-*LBCA*-α and pET21a(+)-*LBCA*-β, respectively. The cells were grown in Luria broth (LB) medium with 150 mg/ ml ampicillin; 20 μM PLP were added to the medium for the expression of LBCA β-subunit. At a cell density of $OD_{600} = 0.5$, protein expression was induced by the addition of 0.5 mM isopropyl-β-thiogalactopyranoside

(IPTG). After growth over night at 20 °C, the cells were harvested by centrifugation (Avanti J-26 XP, JLA 8.1000, 15 min, 4000 rpm, 4 °C). Cell pellets were suspended in 50 mM potassium phosphate buffer pH 7.5 with 300 mM KCl (and 40 μM PLP for purification of the LBCA β-subunit). The cells were disrupted by sonication (Branson Sonifier W-250D, amplitude 50 %, 2 x 2 min, 30 sec pulse/30 sec pause). The cell debris was removed by centrifugation (Sorvall RC5B, SS34, 30 min, 14000 rpm, 4 °C) and soluble proteins were purified by metal chelate affinity chromatography (GE Healthcare, HisTrap FF Crude). The proteins were eluted in 50 mM potassium phosphate buffer pH 7.5 with 300 mM KCl using a linear gradient of 10 to 1000 mM imidazole. Fractions containing sufficiently pure protein were pooled and dialyzed against 50 mM potassium phosphate buffer pH 7.5. Protein concentrations were determined by absorbance spectroscopy (Jasco, V650 spectrophotometer) using a commercial Bradford reagent (Biorad, Bradford protein assay). For crystallization, the subunits were mixed and the formed LBCA TS was purified by preparative size exclusion chromatography with a HiLoad Superdex 75 PG column (GE Healthcare, 120 ml) using 100 mM potassium phosphate buffer pH 7.5 with 300 mM KCl as running buffer. Fractions containing sufficiently pure LBCA TS were pooled and dialyzed against 10 mM HEPES/KOH containing 25 mM KCl.

4.4.3 Absorbance and Circular Dichroism (CD) Spectroscopy

The formation of reaction intermediates was followed by absorption spectroscopy with 50 μM LBCA in 50 mM potassium phosphate buffer pH 7.5 +/- 10 mM L-serine or 1 mM L-tryptophan (NanoDrop 2000c spectrophotometer, Thermo).

The thermal denaturation of LBCA TS α- and β-subunits in 50 mM potassium phosphate buffer pH 7.5 was monitored CD spectroscopy at 220 nm in a 1 mm cuvette (Jasco, spectropolarimeter J-815). Thermal unfolding was induced by increasing the temperature at a rate of 1 °C per min. Data points were connected by LOESS smoothed curves.

4.4.4 Differential Scanning Calorimetry (DSC)

The thermal stability of LBCA TS α- and β-subunits in 50 mM potassium phosphate buffer pH 7.5 was also investigated by DSC. The proteins were heated with a scan rate of 1 °C per min in a microcalorimetry system (Malvern, MicroCal VP-DSC). Melting temperatures were determined by the implemented Origin analysis software.

4.4.5 Analytical Size Exclusion Chromatography

Analytical size exclusion chromatography was performed at 25 °C with a chromatographic device (GE Healthcare, Äkta basic 10). Proteins were eluted from an analytical Superdex 200 column

(GE Healthcare, 10/300 Gl) in 50 mM potassium phosphate buffer pH 7.5 and 300 mM KCl at a flow rate of 0.5 ml/min.

4.4.6 Fluorescence Titration

Increasing amounts of LBCA TS α-subunit were added to 2 μM LBCA TS β-subunit in 10 mM potassium phosphate buffer pH 7.5 in a fluorescence cuvette (1 cm) at 25 °C. Following excitation at 290 nm, the fluorescence emission was detected at 340 nm (Jasco, fluorescence spectrometer FP-6500). Apparent K_d values were calculated by a quadratic fit (Reinstein et al., 1990).

4.4.7 Transient Kinetics

Stopped-flow kinetics were recorded at 25 °C and 60 °C in 50 mM EPPS/KOH pH 7.8, 40 mM L-serine using a SX20 instrument (Applied Photophysics). 75 mM NMHA was mixed in a 1:1 ratio with 10 μM LBCA β-subunit or a α/β-subunit mixture (2 μM monomer concentrations), with or without 50 mM GP. Similarly, 10 mM BZI was mixed with 5 μM LBCA β-subunit or a α/β-subunit mixture (5 μM monomer concentrations), with or without 50 mM GP in the presence of 100 mM anilinium chloride. All concentrations are cell concentrations. The absorbance was monitored at the appropriate wavelength, 6-8 traces were averaged, and resulting transients were fitted to exponential functions by non-linear regression using `ProData SX`.

4.4.8 Steady-state Kinetics

The cleavage of IGP to indole and GAP (α-reaction) was measured at 60 °C by absorbance spectroscopy (Jasco, V650 spectrophotometer) using a coupled enzymatic assay (Creighton, 1970). Initial velocities were recorded in 100 mM EPPS/KOH pH 7.5, 180 mM KCl, 40 μM PLP, 6 mM NAD^+ and 20 mM arsenate at different concentrations of IGP. The reduction of GAP by 5.5 μM glyceraldehyde-3-phosphate dehydrogenase from *Thermotoga maritima* was determined spectroscopically using $\Delta\epsilon$(NADH-NAD^+) = 6.22 $mM^{-1}cm^{-1}$. The condensation of L-serine and indole to L-tryptophan (β-reaction) was measured at 60 °C by absorbance spectroscopy using $\Delta\epsilon$(tryptophan-indole) = 1.89 $mM^{-1}cm^{-1}$. Initial velocities were recorded in 100 mM EPPS/KOH pH 7.5, 180 mM KCl, 40 μM PLP with saturating concentrations of L-serine or L-indole. The values for the steady-state parameters were determined by fitting data points with a hyperbolic function. The reaction of IGP and L-serine to L-tryptophan and GAP ($\alpha\beta$-reaction) was measured at 60 °C by a discontinuous assay. Initial velocities were determined in 100 mM EPPS/KOH pH 7.5, 180 mM KCl, 40 μM PLP with saturating concentrations of IGP or L-serine. Reactions were quenched by the addition of 0.5 volume of 1N KOH. The quenched reactions were subsequently mixed with an equal amount of methanol and analyzed by reversed phase high pressure liquid chromatography) using an Agilent instrument (1200 Series). The

separation was performed at 25 °C with a flow rate of 0.5 ml/min using a Kromasil C18 column (Bischoff, 4 mm x 250 mm, 5 μM particle size) with 0.2 % (w/v) sodium bicarbonate in water as solvent A and methanol as solvent B. The program was as follows: hold with 5 % B for 1.5 min, linear gradient 5-100 % B in 15 min, hold with 100 % B for 5 min, recycle 100-5 % B in 1 min, and re-equilibrate for 7.5 min. The elution time of L-tryptophan was 13.2 min.

4.4.9 Crystallization and Structure Determination

Crystallization was performed with the hanging drop vapor diffusion method in 15 well crystallization plates (Quiagen, EasyXtal). Drops contained a mixture of 1 μl of the reservoir solution and 1 μl protein solution (37.6 mg/ml LBCA TS complex plus 100 mM GP and 200 mM L-serine). Crystals were obtained with 0.1 M sodium citrate tribasic dehydrate pH 5.0 (titrated with HCl), 200 mM NaCl and 15 % PEG 6000 (w/v). After flash freezing in liquid nitrogen, data of single crystals were collected at the synchrotron beamline PX3 (SLS) at 100 K. Data were processed using `XDS` (Kabsch, 2010) and the data quality assessment was done using `phenix.xtriage` (Adams et al., 2002). Molecular replacement was performed with `MOLREP` within the `CCP4i` suite (Potterton et al., 2004). A homology model of LBCA TS with the TS complex from *S. typhimurium* (stTS) (PDB: 2J9X) was built with `MODELLER` (Sali and Blundell, 1993) and served as a search model. Initial refinement was performed using `REFMAC` (Murshudov et al., 1997). The model was further improved in several refinement rounds using automated restrained refinement with the program `PHENIX` (Adams et al., 2002) and interactive modeling with `Coot` (Emsley and Cowtan, 2004). The data collection and refinement statistics are summarized in **Table 4.4**. The final model was analyzed using the program `MolProbity` (Davis et al., 2007).

Author Contribution

F.B., C.R., S.S. conducted experiments, K.H. and R.M. performed ancestral sequence reconstruction, R.M. and R.S. supervised research, all authors contributed to the writing of the manuscript.

Acknowledgements

This work was supported by the Deutsche Forschungsgemeinschaft (ME2259/2-1, STE891/9-1). We thank Hermine Reisner and Sonja Fuchs for expert technical assistance.

4.5 Supplemental Figures and Tables

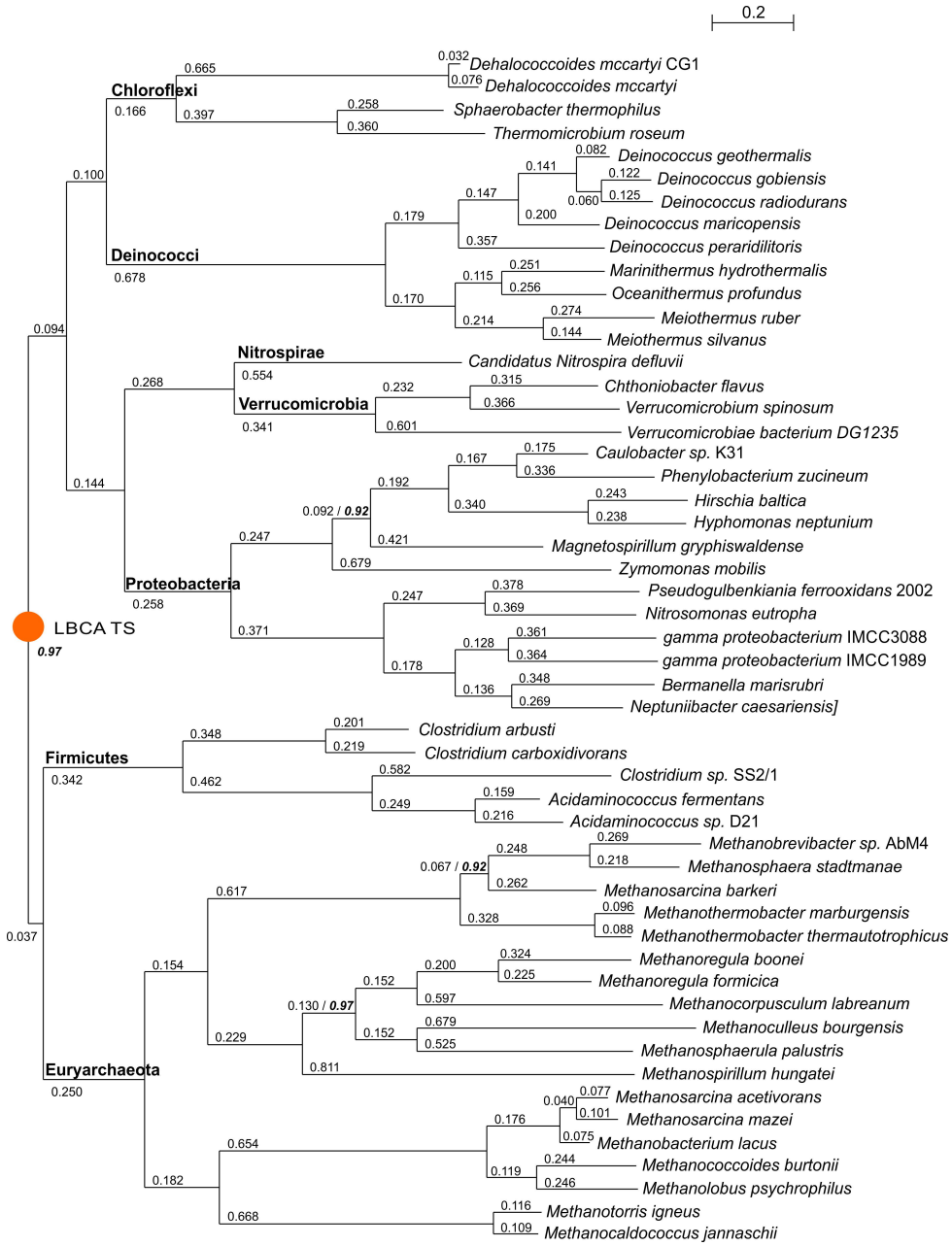

Figure 4.5: Phylogenetic tree for the reconstruction of LBCA TS. *Related to Experimental Procedures.* The constructed tree was midpoint-rooted by taking the phylogenetic relationship between extant TS into consideration. The location of LBCA TS within the phylogenetic tree corresponds to the root of the constructed tree and is indicated by an orange-colored dot. The length of the bar at the top corresponds to 0.2 mutations per site. For each edge, the corresponding rate of mutations per site is given; 97 of the posterior probabilities are ≥ 0.99; the four smaller ones, which are ≥ 0.92, are indicated as numbers formatted in italics and bold.

Figure 4.6: Amino acid sequences of LBCA TS subunits. *Related to Experimental Procedures.* Amino acids corresponding to the expression vector and the His_6-tag are colored red. The single tryptophan in the LBCA β-subunit is colored green. **The sequences can be found as Digital Supplemental Data on the provided data storage medium.**

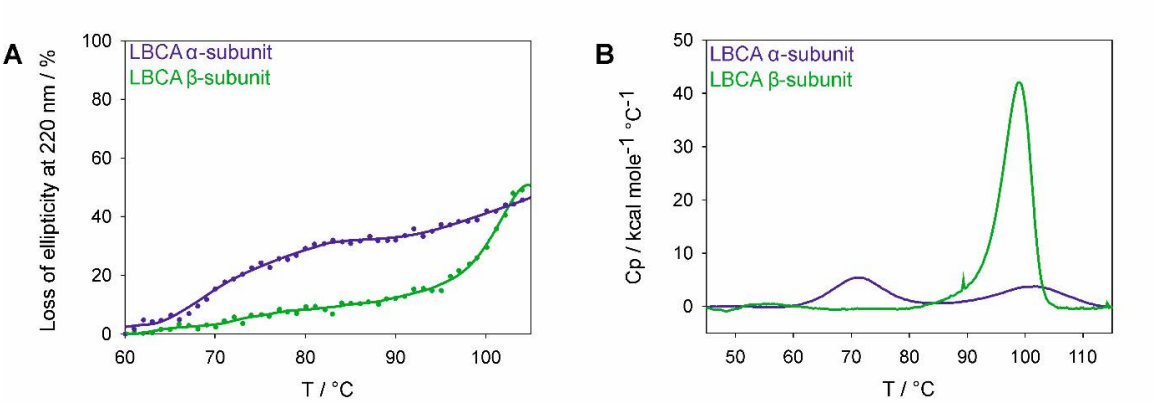

Figure 4.7: Thermal stability of LBCA α- and β-subunits. *Related to Experimental Procedures.* **(A)** Thermal denaturation monitored by CD-spectroscopy. The loss of ellipticity at 220 nm of 10 μM subunit (monomer concentration) in 50 mM potassium phosphate buffer pH 7.5 was monitored from 60 °C to 105 °C with a scan rate of 1 °C/min. The curves connecting the data points were LOESS smoothed. **(B)** Thermal denaturation monitored by DSC. Changes in heat capacity of 15 μM subunit (monomer concentration) in 50 mM potassium phosphate buffer pH 7.5 were detected from 45 °C to 115 °C with a scan rate of 1 °C/min. The curves were baseline corrected.

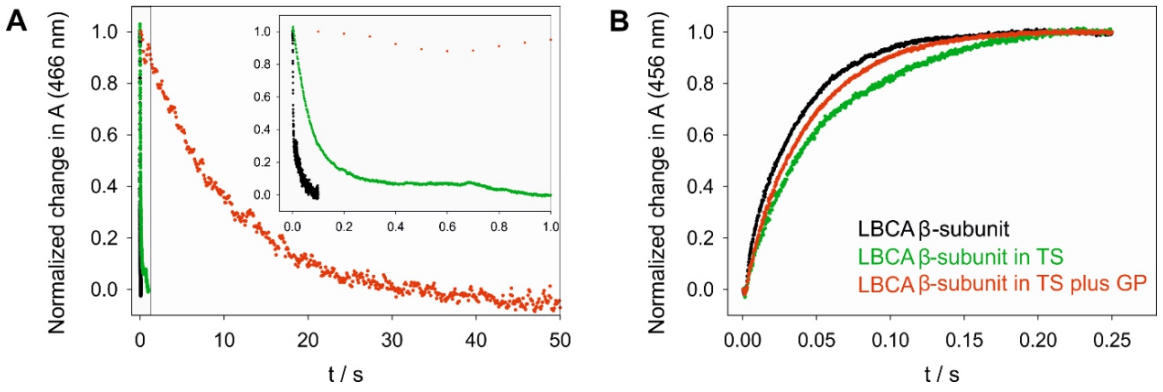

Figure 4.8: Reaction course of two different nucleophiles at the LBCA β-subunit active site. *Related to Experimental Procedures.* **(A)** Effect of TS formation and GP binding on the reaction of the large nucleophile benzimidazole with aniline-quinonoid at 25 °C. Reaction kinetics of 10 mM BZI with aniline-quinonoid as observed for the isolated β-subunit (5 μM monomer concentration; black trace; $k_1 = 580\,s^{-1}$; $k_2 = 27.2\,s^{-1}$), for a α/β-subunit mixture (5 μM monomer concentrations; green trace; $k = 14.0\,s^{-1}$), and for a α/β-subunit mixture (5 μM monomer concentrations) plus 50 mM GP (red trace; $k = 0.094\,s^{-1}$) in 50 mM EPPS/KOH pH 7.8, 40 mM L-serine, 100 mM anilinium chloride. **(B)** Effect of TS formation and GP binding on the reaction of the small nucleophile N-methylhydroxylamine with aminoacrylate at 25 °C. Reaction kinetics of 75 mM NMHA with aminoacrylate as observed for the isolated β-subunit (10 μM monomer concentration; black trace; $k = 28.3\,s^{-1}$), for a α/β-subunit mixture (2 μM monomer concentrations; green trace; $k_1 = 49.1\,s^{-1}$; $k_2 = 12.7\,s^{-1}$), and for a α/β-subunit mixture (2 μM monomer concentrations) plus 50 mM GP (red trace; $k = 23.7\,s^{-1}$) in 50 mM EPPS/KOH pH 7.8, 40 mM L-serine.

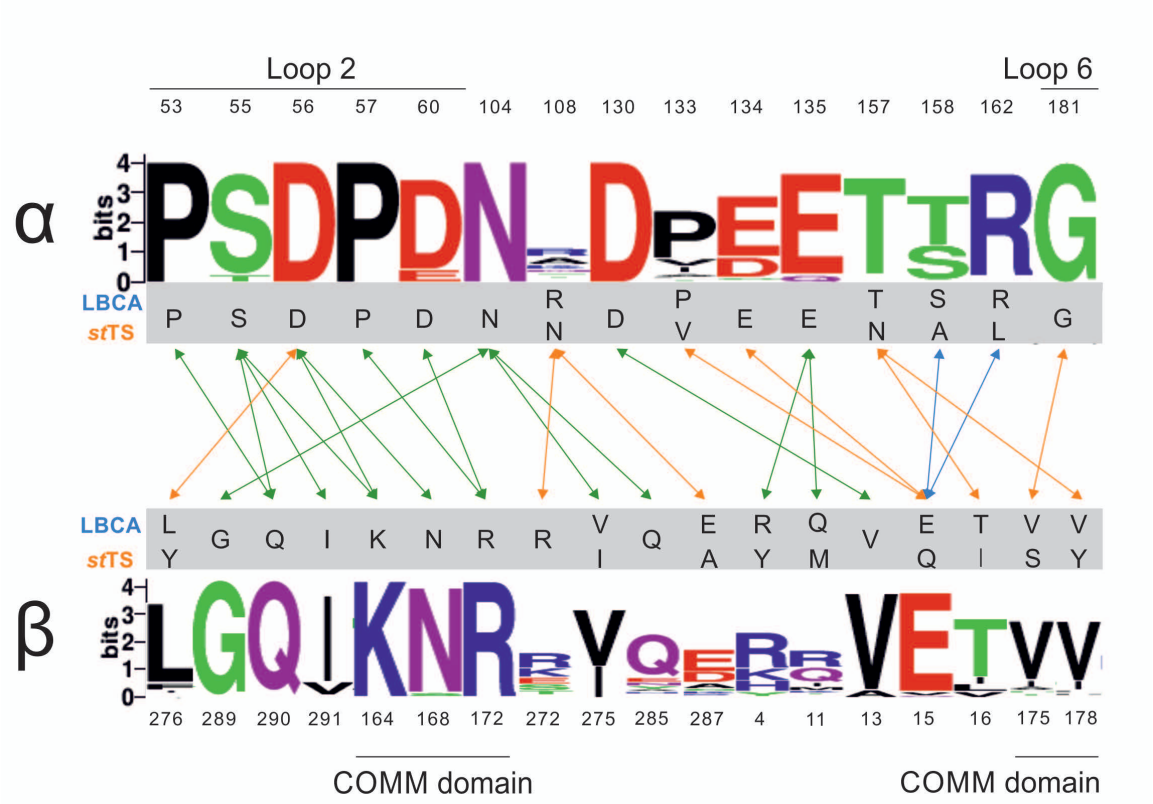

Figure 4.9: Hydrogen bond network at the α/β interfaces of LBCA TS and stTS. *Related to* ***Figure 4.4***. Conservation of interface residues is illustrated with a sequence logo (Jeong and Kim, 2012) deduced from the MSA used for the reconstruction of the LBCA α- and β-subunit. Residues found in LBCA TS and stTS are given in the grey boxes; identical residues are listed only once. Positions are numbered according to the LBCA complex. The H-bonds were determined by analyzing the crystal structures of stTS (PDB: 1WBJ) and LBCA TS (PDB: 5EY5) by means of the PISA server (Krissinel and Henrick, 2007). H-bonds occurring in both interfaces are indicated by green arrows. Blue arrows mark H-bonds exclusively found in LBCA TS and orange ones H-bonds exclusively found in stTS.

Wavelength	1.0 A
Resolution range	47.25 - 1.972 (2.043 - 1.972)
Space group	C 1 2 1
Unit cell	95.96 162.99 78.06 90 93.78 90
Total reflections	302044 (17598)
Unique reflections	81134 (6804)
Multiplicity	3.7 (2.6)
Completeness (%)	97 (81)
Mean I/sigma(I)	8.73 (1.93)
Wilson B-factor	34.96
R-merge	0.08159 (0.4591)
R-meas	0.0933 (0.562)
CC1/2	0.996 (0.801)
CC*	0.999 (0.943)
Reflections used in refinement	81127 (6803)
Reflections used for R-free	4057 (340)
R-work	0.179 (0.292)
R-free	0.231 (0.340)
CC(work)	0.954 (0.790)
CC(free)	0.897 (0.728)
Number of non-hydrogen atoms	9940
macromolecules	9503
ligands	55
Protein residues	1272
RMS(bonds)	0.007
RMS(angles)	0.87
Ramachandran favored (%)	96
Ramachandran allowed (%)	3
Ramachandran outliers (%)	0.79
Average B-factor	45.56
macromolecules	45.51
ligands	54.50
solvent	45.33

Table 4.4: Crystal structure of the LBCA TS: Data collection and refinement statistics. (Statistics for the highest-resolution shell are shown in parentheses.) *Related to* ***Figure 4.3***

Table 4.5: Multiple sequence alignment of concatenated α- and β-subunits of modern TS and sequences of LBCA α- and β-subunits. *Related to Experimental Procedures.* **Sequences can be found as Digital Supplemental Data on the provided data storage medium.**

Chapter 5

Combining Ancestral Sequence Reconstruction with Protein Design to Identify an Interface Hotspot in a Key Metabolic Enzyme Complex

Alexandra Holinski*, Kristina Heyn*, Rainer Merkl, Reinhard Sterner
* These authors contributed equally to this work.

Proteins: Structure, Function, and Bioinformatics, 85(2), 312-321.

Short title: Identification of a protein interface hotspot
Key words: HisF, HisH, imidazole glycerol phosphate synthase, *in silico* mutagenesis, protein-protein interaction.

Abstract

It is important to identify hotspot residues that determine protein-protein interactions in interfaces of macromolecular complexes. We have applied a combination of ancestral sequence reconstruction and protein design to identify hotspots within imidazole glycerol phosphate synthase (ImGPS). ImGPS is a key metabolic enzyme complex, which links histidine and *de novo* purine biosynthesis and consists of the cyclase subunit HisF and the glutaminase subunit HisH. Initial fluorescence titration experiments showed that HisH from *Zymomonas mobilis* (zmHisH) binds with high affinity to the reconstructed HisF from the last universal common ancestor (LUCA-HisF) but not to HisF from *Pyrobaculum arsenaticum* (paHisF), which differ by 103

residues. Subsequent titration experiments with a reconstructed evolutionary intermediate linking LUCA-HisF and paHisF and inspection of the subunit interface of a contemporary ImGPS allowed us to narrow down the differences crucial for zmHisH binding to nine amino acids of HisF. Homology modeling and *in silico* mutagenesis studies suggested that at most two of these nine HisF residues are crucial for zmHisH binding. These computational results were verified by experimental site-directed mutagenesis, which finally enabled us to pinpoint a single amino acid residue in HisF that is decisive for high-affinity binding of zmHisH. Our work shows that the identification of protein interface hotspots can be very efficient when reconstructed proteins with different binding properties are included in the analysis.

5.1 Introduction

Protein-protein interactions are crucial for most cellular processes such as metabolic pathways and signal transduction cascades. Whereas many metabolic reactions are catalyzed by enzyme complexes that consist of permanently assembled protein subunits, signal transduction is often mediated by enzymes that activate or repress each other as a consequence of transient interactions (La et al., 2013). For these interactions well-defined surface patches of the complex-forming partners are responsible, which are called protein-protein interfaces (PPIs). Although PPIs are commonly relatively large (700 - 1500 $Å^2$), binding strength critically depends in many complexes on few residues, which were named "hotspots" (Reichmann et al., 2007; Clackson and Wells, 1995). The identification of such hotspots is a prerequisite for the mechanistic understanding of macromolecular assembly and for the rational design of molecules that modulate it (Watanabe and Osada, 2016).

The crystal structure analysis of numerous protein complexes during the last decades has helped to identify PPIs. However, the quaternary structures deduced from the visual inspection of crystal packing contacts may not reveal the biologically meaningful PPIs, let alone the identification of hotspots. Therefore, in order to localize PPIs and to assess the contribution of individual residues to complex stability, other computational as well as experimental methods are required. Numerous algorithms have been developed that guide the selection of residues potentially belonging to PPIs (Aumentado-Armstrong et al., 2015). The identification and characterization of hotspots requires protein design algorithms that are based on empirical scores or force-fields. Among them are programs like `KFC2` (Zhu and Mitchell, 2011), which is focusing on the prediction of interface hotspots, and `FoldX` (Guerois et al., 2002), which has been designed to predict the effect of mutations on the stability of proteins and protein complexes with known three-dimensional structures. The most widely used experimental approach is the replacement of a particular residue with alanine by site-directed mutagenesis, which in principle allows one to determine the effect of a given amino acid side chain for the dissociation constant of complex formation (Bradshaw et al., 2011; De Genst et al., 2002; Krüger and Gohlke, 2010).

Ancestral sequence reconstruction (ASR) is a computational method that predicts ancient protein sequences from extinct species based on sequences from modern organisms (Harms and Thornton, 2010; Liberles, 2007). The first steps of the ASR pipeline are the compilation of a multiple sequence alignment (MSA) of homologous present-day proteins and the selection of a phylogenetic model which provides mutation frequencies and substitution probabilities for all possible amino acid transitions. Based on the MSA and this model, the most likely phylogenetic tree is calculated and used to infer for the parental nodes the sequences of these predecessors, including the one that corresponds to the last common ancestor of all included present-day proteins. After synthesizing the genes encoding these sequences and their expression in host organisms such as *Escherichia coli*, the ancient proteins can be purified and characterized (Merkl and Sterner, 2016; Wheeler et al., 2016). Numerous attempts of ASR have concentrated on the "resurrection" of up to billions-of-years old proteins and have correlated their stabilities and activities with environmental conditions determining Precambrian life (Busch et al., 2016; Reisinger et al., 2014b; Risso et al., 2014). In addition, the historical approach underlying ASR has been successfully applied to identify amino acid residues key to protein function. This is often not possible by comparing extant proteins due to the large sequence differences observed in homologous proteins. Along these lines, amino acids responsible for the spectroscopic properties of protein pigments could be localized by ASR (Field and Matz, 2010; Ugalde et al., 2004; Yokoyama et al., 2008) as well as residues determining the specificity of steroid receptors for binding their ligands (Bridgham et al., 2006; Eick et al., 2012; Harms et al., 2013; Ortlund et al., 2007). Recently, ASR has been used to identify amino acid residues which are located outside the protein interface but nevertheless control the oligomeric state of an RNA-binding operon attenuator, which is a dimer in a mesophilic but a tetramer in a thermophilic Bacillus species (Perica et al., 2014).

Imidazole glycerol phosphate synthase (ImGPS) is a member of the glutamine amidotransferase family. Enzymes from this family mediate the incorporation of nitrogen into biological molecules by catalyzing the transfer of the amido nitrogen of glutamine to an acceptor substrate (Massiere and Badet-Denisot, 1998; Zalkin and Smith, 1998). In bacteria and archaea, ImGPS consists of the glutaminase subunit HisH and the cyclase subunit HisF, which assemble with high affinity to a bi-enzyme complex (Beismann-Driemeyer and Sterner, 2001). In plants and fungi, the two catalytic domains are fused on a single polypeptide chain (Chittur et al., 2000). HisF uses *N*'-(5'-phosphoribulosyl)-formimino-5-aminoimidazole-4-carboxamide-ribonucleotide (PRFAR) in a cyclization reaction as acceptor for nascent ammonia produced by HisH, leading to the formation of imidazole glycerol phosphate (ImGP), which constitutes an intermediate of histidine biosynthesis, and 5-aminoimidazole-4-carboxamide ribotide (AICAR), which flows into purine biosynthesis. Thereby, ImGPS serves as a branch point enzyme connecting amino acid and nucleotide metabolism (**Figure 5.1**). The sequential HisH and HisF reactions are tightly coupled: hydrolysis of glutamine to glutamate and ammonia by HisH is stimulated by the HisF-substrate PRFAR (Beismann-Driemeyer and Sterner, 2001; Myers et al., 2003), a

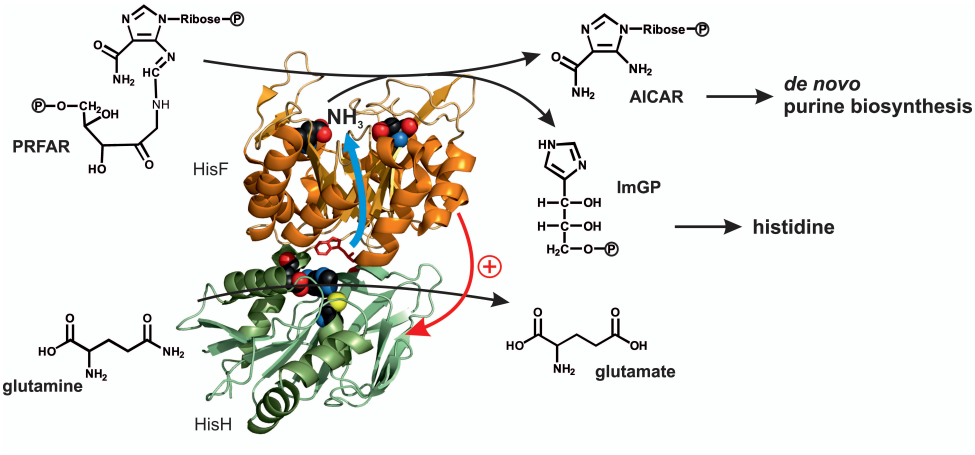

Figure 5.1: Structure and reaction of the ImGP synthase (HisF:HisH complex). Binding of N'-[(5'-phosphoribulosyl)-formimino]-5-aminoimidazole-4-carboxamid-ribonucleotide (PRFAR) to the active site of HisF induces a conformational transition (indicated by the red arrow) which leads to the stimulation of the hydrolytic cleavage of glutamine to glutamate and ammonia at the active site of HisH. Nascent ammonia is then channeled to the active site of HisF (indicated by the blue arrow) where it reacts with PRFAR to imidazole glycerol phosphate (ImGP) and 5-aminoimidazole-4-carboxamide ribotide (AICAR). ImGP is further used in histidine biosynthesis, and AICAR is further used in *de novo* purine biosynthesis. ImGP synthase from *T. maritima* (PDB ID: 1GPW) is shown as a ribbon model. The catalytic aspartate residues of HisF and the catalytic cysteine-histidine-glutamate triad of HisH are shown as spheres; the tryptophan residue of HisH used for fluorescence titration with HisF is indicated in stick representation.

phenomenon that has been ascribed to ligand-induced propagation of conformational motions (Lisi et al., 2016), followed by the channeling of nascent ammonia from the active site of HisH to the active site of HisF (Chaudhuri et al., 2001; Douangamath et al., 2002).

In spite of these detailed biochemical and structural investigations, the individual amino acids responsible for HisF:HisH complex formation, allosteric communication, and ammonia channeling in ImGPS are still unknown (List et al., 2012). In a first step towards this goal, we aimed at the identification of HisF hotspots. For this purpose, we have used a computational approach in which ASR and protein design were combined in order to identify residues in HisF that determine its affinity to a certain HisH protein. The significance of these residues for complex formation was then tested by site-directed mutagenesis and fluorescence titration experiments. This approach allowed us to identify a single hotspot residue in HisF whose mutation leads to a change of the dissociation constant with HisH by 1-2 orders of magnitude. We anticipate that our approach is a general method to speed up the analysis of PPIs and their hotspots.

5.2 Materials and Methods

5.2.1 Cloning and Mutagenesis of *his*F Genes

The gene for *his*F from *Thermotoga maritima* (tm*his*F) was amplified by polymerase chain reaction (PCR) from tm*his*F-pET11c (Beismann-Driemeyer and Sterner, 2001) using the oligonucleotides 5'-ATG CTG **CAT ATG** CTC GCT AAA AGA ATA ATC-3' and 5'-ATG CTG **CTC GAG** *TCA* CAA CCC CTC CAG TCT CAC-3' as 5' and 3' primers, respectively. Pa*his*F was amplified by PCR from genomic DNA from *Pyrobaculum arsenaticum* (DSM13514), which was ordered from the *Leibniz Institute DSMZ*, using the oligonucleotides 5'-GAA GCG **CAT ATG** GCT GTA CGC GTC ATA CC-3' (5'_pa*his*F_*NdeI*) and 5'-GCC CCT **CTC GAG** *TCA* TAG CCT CAC CTC GA-3' (3'_pa*his*F_Stop_*XhoI*) as 5' and 3' primers, respectively. The genes were cloned into the pET-24a(+) (*Stratagene*) vector via the introduced restriction sites *NdeI* and *XhoI*. In order to record the binding of HisF proteins to HisH via fluorescence titration, all tryptophan residues in the HisF proteins were replaced by tyrosines. For the construction of tm*his*F-W156Y, a modified QuikChange mutagenesis protocol (Wang and Malcolm, 1999) was employed using 5'-GCA TAC TTC TGA GAG AC*T ATG* TGG TTG AAG TAG AAA AG-3' and 5'-CTT TTC TAC TTC AAC CAC *ATA* GTC TCT CAG AAG TAT GC-3' as 5' and 3' primers, respectively. For the construction of pa*his*F-W157Y, an overlap extension PCR (Ho et al., 1989) was employed. The megaprimers for this PCR were produced in two separate PCRs using 5'_pa*his*F_*NdeI* and 5'-AGA CGC CGT GGA G*TA TG*C TAA AAA GGT GG-3' as 5' primers and 5'-CCA CCT TTT TAG C*AT A*CT CCA CGG CGT CT-3' and 3'_pa*his*F_Stop_*XhoI* as 3' primers. The resulting amplification products were used in a third PCR, together with the primers 5'_pa*his*F_*NdeI* and 3'_pa*his*F_Stop_*XhoI*. LUCA-*his*F-W138Y+W156Y was available in the pET-24a(+) vector (Reisinger et al., 2014b). Anc1pa-*his*F, Anc1pa-*his*F*, Anc1tm-*his*F and Anc2tm-*his*F were optimized for their expression in *E. coli* and ordered as GeneArt Strings DNA fragments from *Thermo Fisher Scientific*. Their nucleotide and deduced amino acid sequences are given in **Table 5.2**. For the construction of Anc1pa-*his*F-W138Y+W156Y, QuikChange mutagenesis reactions were performed using 5'-CGT GTG GGT GGT GGT *TAT* GAA GTT TTT GTT CG-3' and 5'-CTG GAT GCA GTT GAA *TAT* GCA AAA AAA GTT GAA G-3' as 5' primers, and 5'-CGA ACA AAA ACT TC*A TA*A CCA CCA CCC ACA CG-3' and 5'-CTT CAA CTT TTT TTG C*AT A*TT CAA CTG CAT CCA G-3' as 3' primers. The other reconstructed genes were ordered as variants in which all tryptophan residues were replaced by tyrosines. The GeneArt Strings DNA fragments were cloned into pET-24a(+) vectors using the terminal restriction sites for *NdeI* and *XhoI*. For the construction of Anc1pa-*his*F-A72Q+S74F, overlap extension PCR was performed. The megaprimers for this PCR were produced in two separate PCRs using 5'-GAA GCG **CAT ATG** CTG GCA AAA CGT ATT ATT CC-3' (5'_Anc1pa-*his*F_*NdeI*) and 5'-CGT ACC GCA GAA *CAG* GTT *TTT* ATT CCG CTG ACC-3' as 5' primers, and 5'-GGT CAG CGG AAT *AAA* AAC *CTG* TTC TGC GGT ACG-3' and 5'-GAA GCG **CTC GAG** *TTA* CAG ACG

CAC TTC AAT ACC-3' (3'_Anc1pa-*his*F_Stop_*XhoI*) as 3' primers. The resulting amplification products were used in a third PCR, together with the primers 5'_Anc1pa-*his*F_*NdeI* and 3'_Anc1pa-*his*F_Stop_*XhoI*. Using Anc1pa-*his*F-A72Q+S74F as the template, Anc1pa-*his*F-A72Q and Anc1pa-*his*F-S74F were constructed in two separate QuikChange mutagenesis reactions using 5'-CGT ACC GCA GAA CAG GTT *AGC* ATT CCG CTG ACC GTT GG-3' and 5'-CGT CGT ACC GCA GAA *GCA* GTT TTT ATT CCG CTG-3' as 5' primers, and 5'-CCA ACG GTC AGC GGA AT*G CT*A ACC TGT TCT GCG GTA CG-3' and 5'-CAG CGG AAT AAA AAC *TGC* TTC TGC GGT ACG ACG-3' were used as 3' primers. For the construction of LUCA-*his*F-F74S, Anc2tm-*his*F-F74D and tm*his*F-D74F QuikChange mutagenesis reactions were performed. For this purpose, 5'-CGT ACC GCA GAA CAG GTT *AGC* ATT CCG CTG ACC GTT GG-3', 5'-GTT GCC GAA CAG GTT *GAT* ATT CCG CTG ACC GTT GG-3' and 5'-GTG GCC GAG CAG ATC *TTT* ATT CCG TTC ACT GTT GG-3' were used as 5' primers, and 5'-CCA ACG GTC AGC GGA AT*G CT*A ACC TGT TCT GCG GTA CG-3', 5'-CCA ACG GTC AGC GGA AT*A TC*A ACC TGT TCG GCA AC-3' and 5'-CCA ACA GTG AAC GGA AT*A AA*G ATC TGC TCG GCC AC-3' were used as 3' primers. A stop codon was integrated at the end of each *his*F gene. In all primers the newly introduced restriction sites for *NdeI* and *XhoI* are written in bold, the stop codon is indicated in cursive, and the codon for a newly introduced amino acid is underlined. The oligonucleotides were ordered from *biomers*. Zm*his*H was available in the pET24a(+) vector (Reisinger et al., 2014b).

5.2.2 Heterologous Expression and Purification of HisF Proteins and zmHisH

All genes were expressed in *E. coli* BL21-Gold (DE3) cells (*Agilent Technologies*) transformed with the respective pET-24a(+) plasmid. For this purpose, 1 to 4 L of LB medium supplemented with 75 μg/mL kanamycin were inoculated with a preculture and incubated at 37 °C. After an OD_{600} of 0.6 was reached, the temperature was lowered to 30 °C in the case of the HisF proteins, and 20 °C in the case of zmHisH. Expression was induced by adding 0.5 mM IPTG, and growth was continued overnight. Cells were harvested by centrifugation (*Beckman Coulter* Avanti J-26SXP, JLA-8.1000, 20 min, 4000 rpm, 4 °C), suspended in 50 mM potassium phosphate pH 7.5 (HisF proteins) or 50 mM Tris/HCl pH 9.0 (zmHisH), and lysed by sonification (Branson Sonifier W-250D, 3 min in 5 s intervals, 45 % pulse, 0 °C). To separate the soluble from the insoluble fraction of the cell extract, the cells were centrifuged again (*Beckman Coulter* Avanti J-26SXP, JA-25.50, 30 min, 13.000 rpm, 4 °C). The soluble supernatant was subjected to ion exchange chromatography using a MonoQ column (HR 16/10, 20 mL, GE Healthcare), which had been equilibrated with 50 mM potassium phosphate pH 7.5 and 50 mM Tris/HCl pH 9.0 for the purification of HisF proteins and zmHisH, respectively. The column was washed with equilibration buffer and the proteins were eluted by applying a linear gradient of 0-1.5 M NaCl. Fractions which contained HisF or HisH, respectively, were pooled and dialyzed against 50 mM potassium phosphate pH 7.5 (HisF proteins) or 50 mM Tris/HCl pH 7.5 (zmHisH). In the following step,

the proteins were precipitated with 80 % ammonia sulfate, centrifuged (*Beckman Coulter* Avanti J-26SXP, JA-25.50, 30 min, 13.000 rpm, 4 °C), dissolved in either 50 mM potassium phosphate pH 7.5 (HisF proteins) or 50 mM Tris/HCl pH 7.5 and 300 mM KCl (zmHisH) and further purified via size exclusion chromatography. To this end a Superdex75 (HiLoad 26/60, 320 mL, GE Healthcare) was operated with 50 mM potassium phosphate pH 7.5 in the case of HisF proteins, and 50 mM Tris/HCl pH 7.5 and 300 mM KCl in the case of zmHisH. Fractions with pure protein were pooled and in the case of zmHisH dialyzed against 50 mM Tris/HCl pH 7.5. According to SDS-PAGE (12.5 % acrylamide), all proteins were more than 95 % pure. In the case of HisF proteins the yield was about 15 mg to 220 mg per liter of culture, and about 10 mg of zmHisH protein were obtained per liter of culture.

5.2.3 Fluorescence Titration

Upon complex formation with HisF, a conserved tryptophan residue of HisH (Trp134 in zmHisH) is shielded from the solvent (Beismann-Driemeyer and Sterner, 2001). The resulting increase of the fluorescence emission was used in titration experiments to determine the dissociation constants for the formation of the various HisF:zmHisH complexes. For this purpose, 5 μM zmHisH were titrated at 25 °C with the different tryptophan-free HisF variants in 50 mM potassium phosphate pH 7.5. The tryptophan fluorescence was excited at 295 nm and the change of fluorescence emission intensity at 318 nm in dependence of the applied HisF concentration was followed. As we observed that the addition of HisF led to a fluorescence emission background signal at 318 nm, we corrected all titrations for this signal. The corrected fluorescence at 318 nm was plotted against the HisF concentration and the resulting curves were analyzed with a quadratic fit (Reinstein et al., 1990). For the illustration of the curves, the normalized fluorescence signal was plotted against the molar ratio of HisF/HisH, and only the data points until a molar ratio HisF/HisH of 2 are shown in **Figure 5.6**. All titrations were done in duplicate at a Cary-Eclipse fluorimeter (*Varian*).

5.2.4 Far-UV CD-Spectroscopy

Far-UV circular dichroism spectra of all HisF variants were recorded in a JASCO J-815 spectrometer. The spectra were measured with 20 μM or 30 μM HisF protein in 10 mM potassium phosphate, pH 7.5, in a 0.02 cm cuvette at 25 °C. The curves shown in **Figure 5.7** are the result of five accumulations and were smoothed (Savitzky and Golay, 1964). The observed ellipticity was standardized according to the molar ellipticity per amino acid.

5.2.5 ASR of Intermediate Sequences

The MSA of 87 modern-day HisF sequences, the reconstructed LUCA-HisF sequence, and the phylogenetic tree, whose topology guided ASR, were taken from our previous publication

(Richter et al., 2010). This MSA, which contains paHisF, tmHisF, and zmHisF, and the tree were utilized to determine the intermediates Anc1tm-HisF, Anc2tm-HisF, and Anc1pa-HisF by means of the `FastML` server (Ashkenazy et al., 2012), which is an implementation of a fast and accurate ASR algorithm (Randall et al., 2016). For this calculation, the substitution model `JTT` and a `gamma distribution` were chosen to compute a ML tree with optimized branch lengths and the ancestral sequences by means of the joint reconstruction approach.

5.2.6 Interface Prediction

The positions of tmHisF interface residues were determined by analyzing the tmHisF:tmHisH complex (PDB ID: 1GPW, CD) with the help of the webserver `PISA` (*http://www.ebi.ac.uk/pdbe-/pisa/*). Using an MSA containing the 87 modern-day HisF sequences utilized for reconstruction and the sequences of four predecessors, these positions were transferred to the sequences of LUCA-HisF, Anc1tm-HisF, Anc2tm-HisF, Anc1pa-HisF, and paHisF.

5.2.7 Homology Modelling

Homology models for the complex structures of LUCA-HisF:zmHisH and Anc1pa-HisF:zmHisH were determined by means of `YASARA` (Krieger et al., 2009). The crystal structure of tmHisF:tmHisH (PDB ID: 1GPW, CD) served as a template; `YASARA` was run with default parameters. All complex structures were visualized by means of `PyMol` (Schrödinger, LLC, 2015).

5.2.8 Calculating the Interaction Energy of Protein Complexes

For each complex, the three `FoldX` (Guerois et al., 2002) functions `RepairPDB`, `BuildModel`, and `AnalyseComplex` were used to *i*) relax the 3D structure of the complex, *ii*) introduce mutations, and *iii*) predict their effects on the interaction free energy ($\Delta\Delta$G) of the full complex. Using the `FoldX` function `Pssm`, an *in silico* saturation mutagenesis was carried out at HisF position 74 to assess the contribution of all 20 residues to the complex stability of LUCA-HisF:zmHisH and Anc1pa-HisF:zmHisH. All functions of `FoldX` version 4 were utilized with default parameters.

5.2.9 Predicting Hotspots

Hotspots were predicted with `FoldX`, as described above and additionally by using the `KFC2` (Zhu and Mitchell, 2011) server. The only input required by `KFC2` is the three-dimensional structure of a protein complex and the specification of the protein chains forming the complex; no other parameters can be chosen. More specifically, the input was the crystal structure of the tmHisF:tmHisH complex (PDB ID: 1GPW, CD) or alternatively the homology model determined by means of `YASARA` for the complex LUCA-HisF:zmHisH. The predicted hotspots and their confidence scores were extracted from the result page.

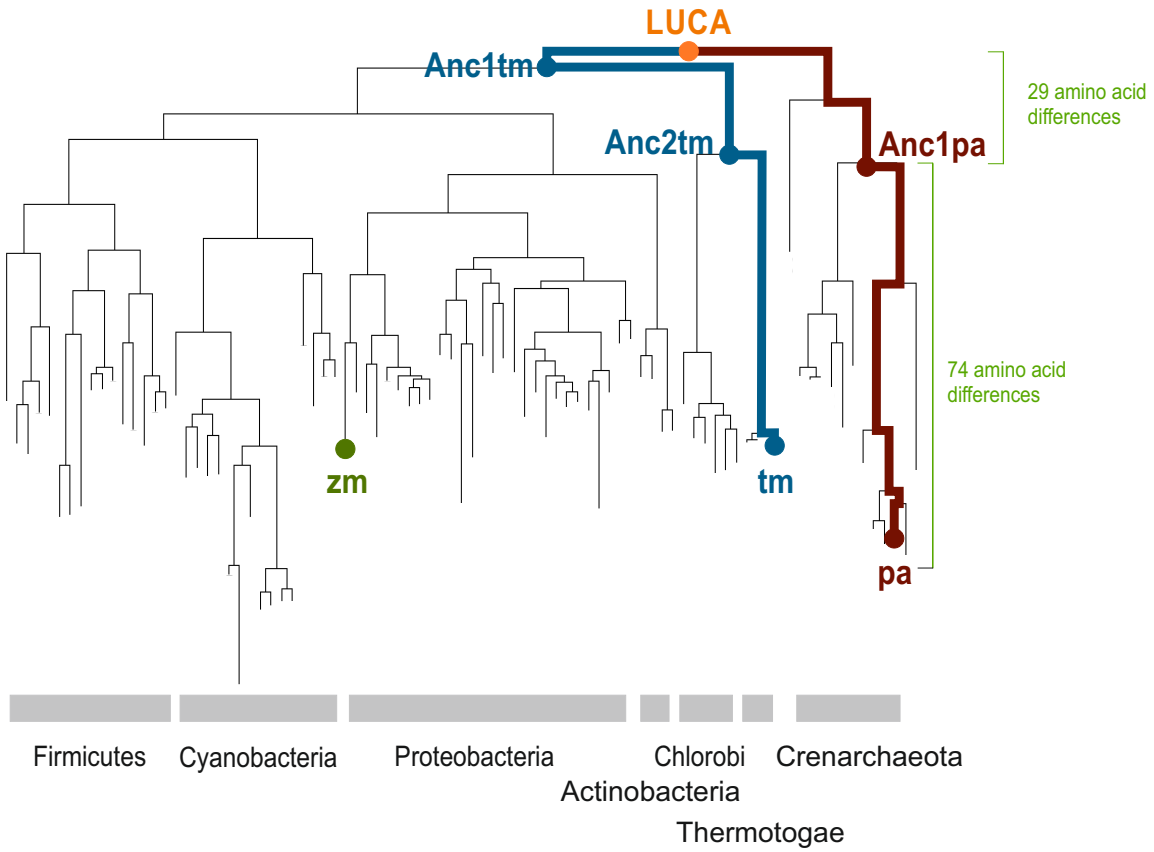

Figure 5.2: Phylogenetic tree based on 87 concatenated HisF and HisH sequences from seven phylogenetic clades. This tree and the modern-day sequences were used for ASR. The evolutionary paths from the last universal common ancestor (LUCA) to *P. arsenaticum* (pa) and *T. maritima* (tm) are marked in bold brown and bold blue, respectively. LUCA-HisF and the evolutionary intermediates Anc1pa-HisF, Anc1tm-HisF, and Anc2tm-HisF were reconstructed and characterized with respect to their binding to HisH from *Z. mobilis* (zm). The number of amino acid exchanges between LUCA-HisF and Anc1pa-HisF and between Anc1pa-HisF and paHisF are indicated in green.

5.3 Results

We have previously used ASR to determine the sequences of HisF and HisH from the last universal common ancestor of all modern organisms (LUCA-HisF, LUCA-HisH) (Richter et al., 2010); the phylogenetic tree used for reconstruction is depicted in **Figure 5.2**. The genes for LUCA-HisF and LUCA-HisH were then synthesized and expressed in *E. coli*, and the proteins were purified and characterized (Reisinger et al., 2014b). The formation of the complex between the two reconstructed proteins was recorded by fluorescence titration experiments, which are based on the shielding of a conserved tryptophan residue of LUCA-HisH upon interaction with LUCA-HisF (**Figure 5.1**). LUCA-HisF and LUCA-HisH form a high-affinity complex (Reisinger et al., 2014b). Remarkably, LUCA-HisF also bound to HisH from the present-day bacterium *Zymomonas mobilis* (zmHisH), and the low K_d-value of less than 50 nM for the LUCA-HisF:zmHisH complex suggested that the ImGPS interface has been conserved through-

HisH	HisF	K_d[nM]
zmHisH	LUCA-HisF	< 50
	paHisF	> 5000
	Anc1pa-HisF	2730 (2400; 3050)
	Anc1pa-HisF*	< 50
	Anc1pa-HisF-A72Q+S74F	< 50
	Anc1pa-HisF-A72Q	3080 (2630; 3520)
	Anc1pa-HisF-S74F	< 50
	LUCA-HisF-F74S	1440 (1040; 1830)
	tmHisF	1560 (943; 2170)
	tmHisF-D74F	< 50
	Anc1tm-HisF	< 50
	Anc2tm-HisF	< 50
	Anc2tm-HisF-F74D	380 (312; 447)

Table 5.1: Dissociation constants for the interaction of zmHisH with various HisF proteins. The shown K_d-values are the mean as determined from two independent titration experiments. (The individually determined values are given in parenthesis.) For the high-affinity complexes, only an upper limit of $K_d < 50$ nM can be given. No complex formation could be detected between paHisF and zmHisH, indicating that $K_d > 5000$ nM.

out evolution. This and all other titration curves recorded in this work are shown in **Figure 5.6**, and the resulting dissociation constant (K_d) values are summarized in **Table 5.1**.

To further test the conservation of the HisF-HisH interface, we performed titration experiments between HisF from the archaeon *Pyrobaculum arsenaticum* (paHisF) and zmHisH. However, no binding of the two proteins could be detected under the given experimental conditions indicating that the K_d-value for the formation of the paHisF:zmHisH complex must be higher than 5000 nM. Thus, although residues of PPIs are generally more conserved than other surface residues (Janin et al., 2008), our results demonstrate that the interfaces of at least some modern HisH and HisF enzymes are incompatible due to their species-specific evolution in the post-LUCA era. In agreement with this finding, the number of identical residues in the interfaces of zmHisF, paHisF, and tmHisF, which comprise between 34 and 36 residues, is only in the range of 47 % - 59 %.

Based on our interest to identify HisF hotspots we wanted to pinpoint residues being responsible for the different affinities of LUCA-HisF and paHisF for zmHisH. LUCA-HisF and paHisF contain about 250 amino acids of which 103 residues are different between the two proteins. In order to narrow down the number of candidate residues, we utilized `FastML` (Ashkenazy et al., 2012) to reconstruct evolutionary intermediates linking LUCA-HisF with paHisF and tmHisF, respectively. The sequences of the characterized ancestors (**Table 5.2**) and of all 87 modern HisF enzymes used for the phylogenetic analysis (**Table 5.3**) as well as the phylogenetic tree (**Figure 5.8**) and the log likelihood values together with the posterior probabilities of each position (**Table 5.4**) are shown in the Supporting Information.

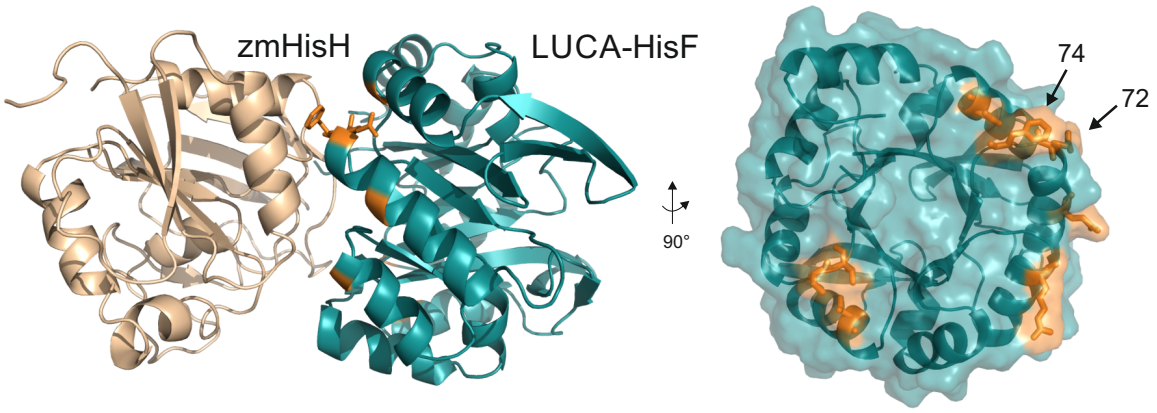

Figure 5.3: Model of the LUCA-HisF:zmHisH complex. The nine interface residues in which LUCA-HisF and Anc1pa-HisF differ are marked in orange. Positions 72 and 74, which were predicted by `FoldX` (Guerois et al., 2002) as crucial for complex stability, are marked with arrows. The model was created by means of `YASARA` (Krieger et al., 2009).

To begin with, we produced Anc1pa-HisF, which differs from LUCA-HisF and paHisF by 29 and 74 residues, respectively (**Figure 5.2**). Titration experiments yielded a K_d-value of 2730 nM for the formation of the Anc1pa-HisF:zmHisH complex, corresponding to a binding affinity which is at least 55-fold lower than that of the LUCA-HisF:zmHisH complex. We reasoned that residues responsible for these different binding affinities might be mainly localized in the interface region, which we inferred from the known crystal structure of ImGPS from *T. maritima* (Douangamath et al., 2002). When analyzing the 29 residue differences between LUCA-HisF and Anc1pa-HisF, we found that nine of them are located in the putative interface with zmHisH (**Figure 5.3; Figure 5.4 A**). The nine residues of Anc1pa-HisF were replaced by the corresponding residues from LUCA-HisF, yielding Anc1pa-HisF* (**Figure 5.4 B**). A K_d-value of less than 50 nM was determined for the formation of the Anc1pa-HisF*:zmHisH complex, which proves that nine residues at most are decisive for the high affinity between these two proteins. In order to estimate whether some of these residues are more important than others, we analyzed several ImGPS interfaces by a combination of *in silico* approaches.

As a first step, homology models for LUCA-HisF:zmHisH and Anc1pa-HisF:zmHisH were built by means of `YASARA` (Krieger et al., 2009). To identify hotspots, the native tmHisF:tmHisH complex and LUCA-HisF:zmHisH were analyzed by using the `KFC2` server. The results, which are summarized in **Table 5.5**, suggested that among the nine residue differences between LUCA-HisF and Anc1pa-HisF, only position 74 (Phe in LUCA-HisF, Ser in Anc1paHisF) is a hotspot. As an alternative, we utilized `FoldX` that additionally assesses the effect of mutations on protein and complex stability. Thus, we reciprocally exchanged the nine residues in all combinations between LUCA-HisF and Anc1pa-HisF (**Figure 5.5 A**). The 12 combinations of mutual residue exchanges leading to the largest predicted effects on complex stability ($|\Delta\Delta G| > 2$ kcal/mol) are depicted in **Figure 5.5 B**. Most strikingly, `FoldX` predicted for the single substitution of Ser74

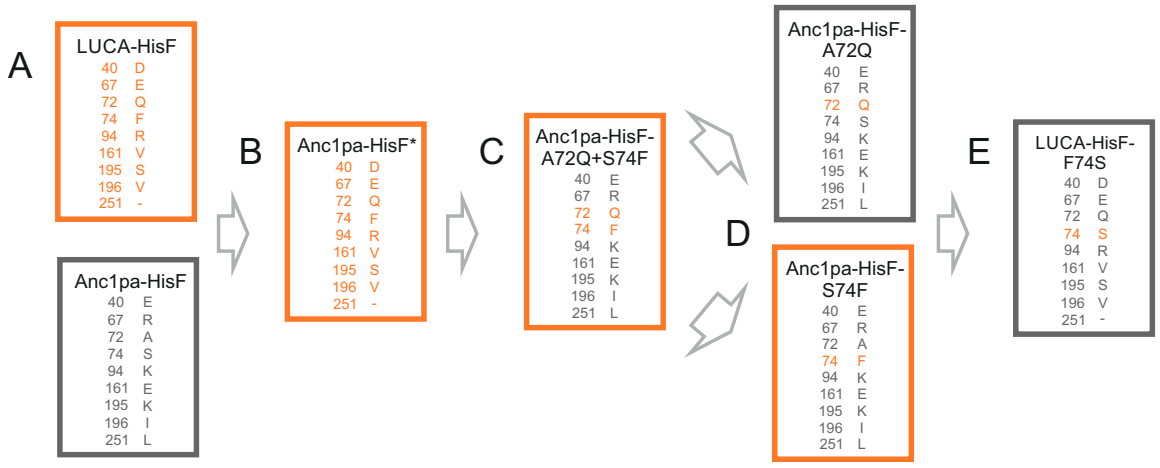

Figure 5.4: Stepwise identification of a HisF hotspot for binding to zmHisH. The rationale for the generation of the HisF variants shown in panels **A - E** is detailed in the text. HisF variants with high affinity ($K_d < 50$ nM) and low affinity ($K_d >> 50$ nM) for zmHisH are framed in orange and grey, respectively. The corresponding fluorescence titration curves are shown in **Figure 5.6** and the determined K_d-values for complex formation are listed in **Table 5.1**.

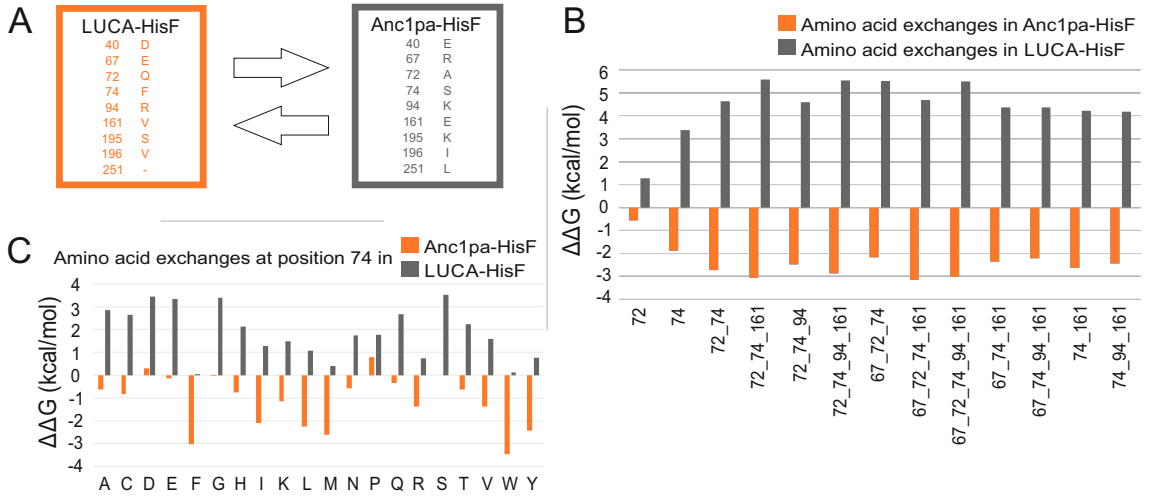

Figure 5.5: Identification of interface residues determining the affinity of LUCA-HisF and Anc1pa-HisF for zmHisH by means of *in silico* design. (A) Nine interface residues distinguish LUCA-HisF and Anc1pa-HisF; these residues were reciprocally exchanged individually and in all possible combinations. For each of these exchanges their effect on complex stability was assessed by means of `FoldX` (Guerois et al., 2002). **(B)** Combinations of reciprocal residue exchanges that have - according to `FoldX` - the largest effects on complex stability. A $\Delta\Delta G$-value < 0 indicates a stabilization of the complex, $\Delta\Delta G > 0$ indicates a destabilization. For comparison, the predicted effects of the individual reciprocal exchanges at positions 72 and 74 are also shown. **(C)** $\Delta\Delta G$-values predicted by `FoldX` for saturation mutagenesis of residue position 74 in LUCA-HisF and Anc1pa-HisF.

from Anc1pa-HisF with Phe74 from LUCA-HisF a stabilization of the complex with zmHisH ($\Delta\Delta G = -1.9$ kcal/mol); for the inverse substitution of Phe74 from LUCA-HisF with Ser74 from Anc1pa-HisF, a destabilization of the complex was predicted ($\Delta\Delta G = +3.4$ kcal/mol). A comparison of the $\Delta\Delta G$ values indicated that the largest destabilizing effects have to be expected for a combination of three or four substitutions, however most of those combinations contained residues 74 and 72. According to `FoldX`, the exclusive mutual exchange of residue 72 has only a weak effect and that of residue 74 is the dominating single substitution. Using the `FoldX` function `Pssm`, we estimated the effect on complex stability for all possible amino acid replacements at position 74 (**Figure 5.5 C**). For the LUCA-HisF:zmHisH complex, all substitution were destabilizing ($\Delta\Delta G > 0$). In Anc1pa-HisF:zmHisH, the replacement of the polar amino acid serine with one of six hydrophobic residues had a strong stabilizing effect ($\Delta\Delta G < -2$ kcal/mol), the most beneficial ones being phenylalanine and tryptophan. Inspecting the homology model of Anc1pa-HisF:zmHisH by means of `PyMol` (Schrödinger, LLC, 2015) indicated that the substitutions Ser74Phe and Ser74Trp does not increase the number of hydrogen bonds but that the bulky aromatic amino acids residues improve shape complementarity by filling a small cavity at the interface of the complex.

In order to test the validity of our *in silico* analyses, Ala72 and Ser74 of Anc1pa-HisF were simultaneously replaced by Gln72 and Phe74 from LUCA-HisF, yielding Anc1pa-HisF-A72Q+S74F (**Figure 5.4 C**). The K_d-value of less than 50 nM as measured for the formation of the Anc1pa-HisF-A72Q+S74F:zmHisH complex resembles the value for LUCA-HisF:zmHisH, narrowing down the crucial differences between LUCA-HisF and Anc1pa-HisF to two residues. In the next step, Ala72 and Ser74 of Anc1pa-HisF were replaced individually by Gln72 and Phe74 from LUCA-HisF, yielding Anc1pa-HisF-A72Q and Anc1pa-HisF-S74F (**Figure 5.4 D**). The K_d-values for the formation of the Anc1pa-HisF-A72Q:zmHisH and Anc1pa-HisF-S74F:zmHisH complexes were 3080 nM and less than 50 nM, respectively, which indicated that the LUCA-HisF residue Phe74 is much more important than residue Gln72 for binding to zmHisH. This conclusion was finally confirmed by the replacement of Phe74 from LUCA-HisF with Ser74 from Anc1pa-HisF (**Figure 5.4 E**): the K_d-value of 1440 nM as determined for the formation of the LUCA-HisF-F74S:zmHisH complex was at least 29-fold higher than for LUCA-HisF:zmHisH but only two-fold lower than for Anc1pa-HisF:zmHisH.

Our analysis showed that Phe74 of LUCA-HisF is decisive for its affinity for zmHisH. Interestingly, this residue is not conserved: Within the 74 bacterial HisF sequences used to compute LUCA-HisF, the frequency of Phe was 72 % and the corresponding 13 archaeal HisF sequences contained Ser (84 %) or Thr (16 %) residues only. We speculated that the occupancy of position 74 might have in general a strong influence on the affinity of a given HisF enzyme for zmHisH. In order to test this hypothesis, we performed a titration experiment between tmHisF, which has an aspartate residue at position 74, and zmHisH. The resulting K_d-value of 1560 nM for the tmHisF:zmHisH complex signaled a low affinity between these two proteins. The replacement of Asp74 with Phe in tmHisF resulted in a K_d-value of less than 50 nM for the

tmHisF-D74F:zmHisH complex, corresponding to an at least 31-fold affinity increase. The evolutionary intermediates Anc1tm-HisF and Anc2tm-HisF, which link LUCA-HisF with tmHisF (**Figure 5.2**) both carry a Phe at position 74. The corresponding genes were synthesized and expressed in *E. coli*, and the two proteins were purified and characterized. The K_d-values for the formation of Anc1tm-HisF:zmHisH and Anc2tm-HisF:zmHisH are less than 50 nM, testifying to high affinities. In contrast, the K_d-value for the formation of Anc2tm-HisF-F74D:zmHisH is 380 nM, corresponding to an at least 8-fold affinity decrease. These data indicate that, indeed, Phe74 of HisF is a hotspot with respect to binding zmHisH. In order to exclude that variants with low affinity for zmHisH are misfolded, the structural integrity of all investigated HisF proteins was demonstrated by far-UV CD spectroscopy (**Figure 5.7**). Moreover, we found that all HisF variants with poor binding to zmHisH nevertheless formed a tight complex with tmHisH or paHisH (data not shown), confirming that they adopt a well-defined tertiary structure.

5.4 Discussion

The residues of sophisticated PPIs have to fulfill demanding tasks like the propagation of allosteric signals and must concurrently ensure complex stability. With respect to stability, key residues have been designated as hotspots. This term has been introduced early on (Clackson and Wells, 1995) and confirmed later on by an analysis of numerous alanine scanning experiments (Bogan and Thorn, 1998) and other datasets. As recently summarized, hotspots directly influence interactions, are often found in central regions of the interface, are enriched in Trp, Arg, and Ser residues, and are more conserved than the remaining interface residues (Zhang et al., 2013). However, neither chemical nor evolutionary conservation has been a reliable measure to predict the effect of specific mutations on complex stability in a comprehensive study of the interface between the human growth hormone and its receptor (Pal et al., 2006). This example illustrates that it is difficult to understand the contribution of individual residues based on simple criteria such as conservation and asks for sophisticated *in silico* analyses.

We were specifically interested to characterize the PPI between the cyclase subunit HisF and the glutaminase subunit HisH of ImGPS, which is a well-studied key metabolic enzyme that links *de novo* purine with histidine biosynthesis (Beismann-Driemeyer and Sterner, 2001). We have used a novel approach to address this problem, namely the combination of ASR with protein design algorithms. The results indicated that position 74 is a hotspot, which could be confirmed by site-directed mutagenesis. However, this hotspot is located at the rim instead of the core of the interface (**Figure 5.3**) and is only moderately conserved. Thus, with respect to these properties, it differs from the ideal hotspot residue.

Why is the HisF hotspot located at the rim of the interface? Most likely, the composition of the ImGPS interface core is dictated by strong functional constraints that are linked to allosteric signaling and ammonia channeling and require the presence of a well-defined set of amino acids.

Indeed, a comparison of LUCA-HisF and modern-day enzymes confirms a strict conservation of most core residues (Reisinger et al., 2014b). On the other hand, the low binding affinities determined for the chimeric complexes paHisF:zmHisH and tmHisF:zmHisH make clear that in the post-LUCA era the evolution of the ImGPS interface continued and affected predominantly rim residues like Phe74.

It is fair to note that ASR is a stochastic method, which means that the reconstructed sequences can differ from the "correct" ancestral sequence to a certain extent (Bar-Rogovsky et al., 2015; Merkl and Sterner, 2016). However, this does not diminish the use of ASR for narrowing down the number of crucial residues as long as the reconstructed proteins can be produced in stable form and differ in their binding properties from each other and/or their modern-day relatives. However, in cases where the ancestral proteins do not differ from their modern descendants, ASR cannot contribute to the identification of hotspots. In any case, ASR alone will suffice to pinpoint crucial residues only in such rare cases where the analyzed proteins deviate from each other by only few residues. If the number of differing residues is larger, the number of candidate residues has to be further limited by applying other computational methods. In our hands, `FoldX` was well suited to drastically narrow down the number of mutagenesis experiments. We could, in principle, have attempted to identify HisF:zmHisH hotspots without using ASR and by exclusively relying on `FoldX` and `KFC2`. In such case, we would have had to analyze *in silico* the differences between modern HisF proteins with high affinity and those with low affinity to zmHisH. However, all modern HisF proteins tested by us were either insoluble when expressed in *E. coli* or showed low affinity to zmHisH. Moreover, even if we had found a soluble HisF protein with high affinity to zmHisH, a comparative *in silico* analysis would most probably have been extremely extensive given that HisF proteins differ by about 125 residues. (HisF proteins contain about 250 residues and the average sequence identity is about 50 %.) Almost 20 of these different residues are located in the interface with HisH and we would have had to analyze all of them in a combinatorial manner with `FoldX` and `KFC2`. In comparison, ASR allowed us to reduce the number of HisF residues with potential significance for binding to zmHisH to 29, and only nine of those are located in the interface with zmHisH. Our example illustrates that ASR is most useful to identify hotspots when modern representatives of a given protein family are not available in stable and soluble form and/or when large amino acid sequences differences between them impede a straightforward *in silico* comparison.

We conclude that in addition to the initial goal of ASR, namely to assess characteristics of primordial proteins, its potential to study mutational pathways (Zhang et al., 2013) offers new possibilities to promote our understanding of the structural basis and the evolution of protein-protein interactions.

Acknowledgement

We thank Veronika Schmid for cloning of several of the constructs, and Jeannette Ueckert and Sonja Fuchs for expert technical assistance. This work was supported by the Deutsche Forschungsgemeinschaft (ME2259/2-1, STE891/9-1).

5.5 Supplemental Figures and Tables

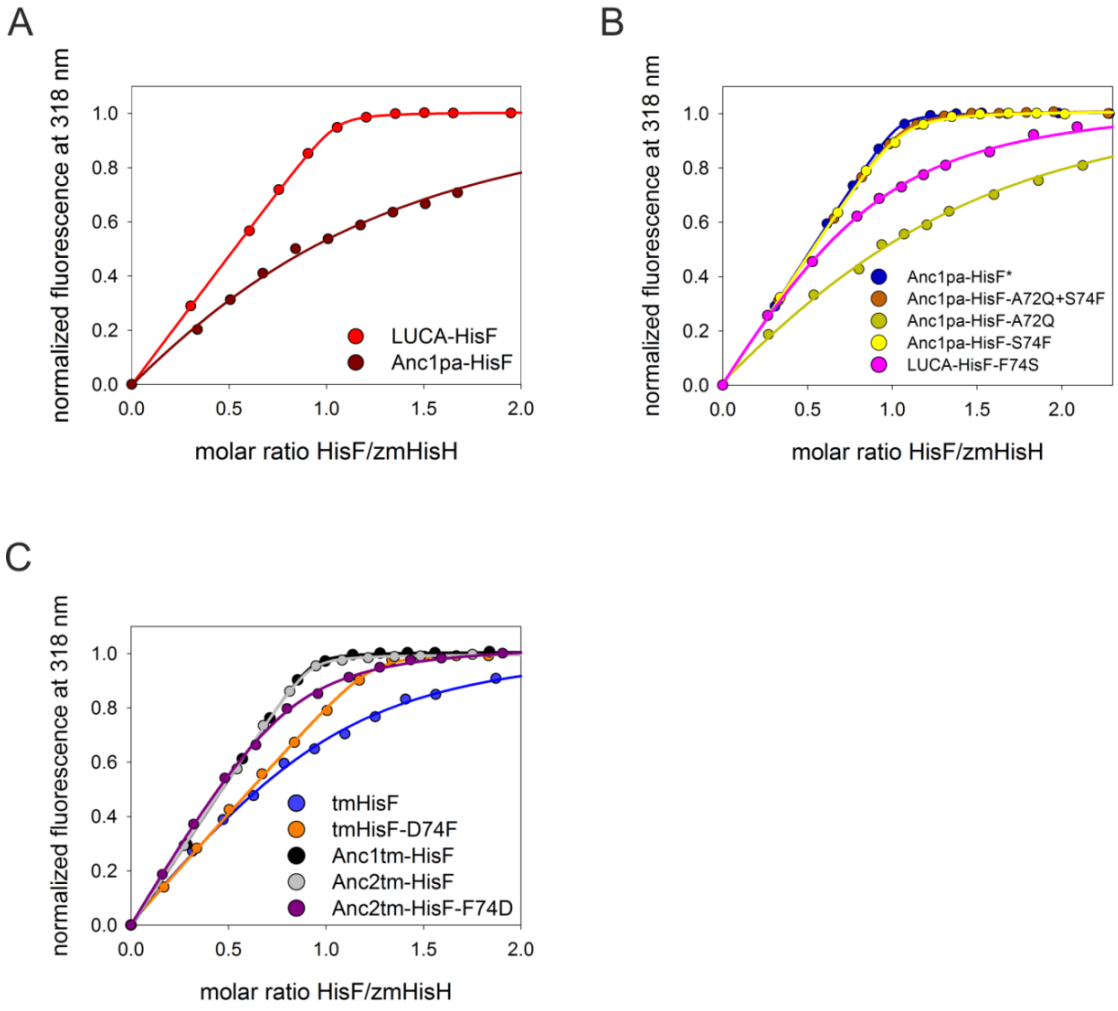

Figure 5.6: Fluorescence titration experiments to determine dissociation constants for the interaction of zmHisH with various HisF subunits. zmHisH at a concentration of 5 μM was titrated with tryptophan-free HisF variants in 50 mM potassium phosphate, pH 7.5, at 25 °C. Fluorescence of the single tryptophan residue Trp134 of zmHisH was excited at 295 nm, and the emission intensity was determined at 318 nm. Data points were fitted with a quadratic equation. The resulting K_d-values are listed in **Table 5.1**.

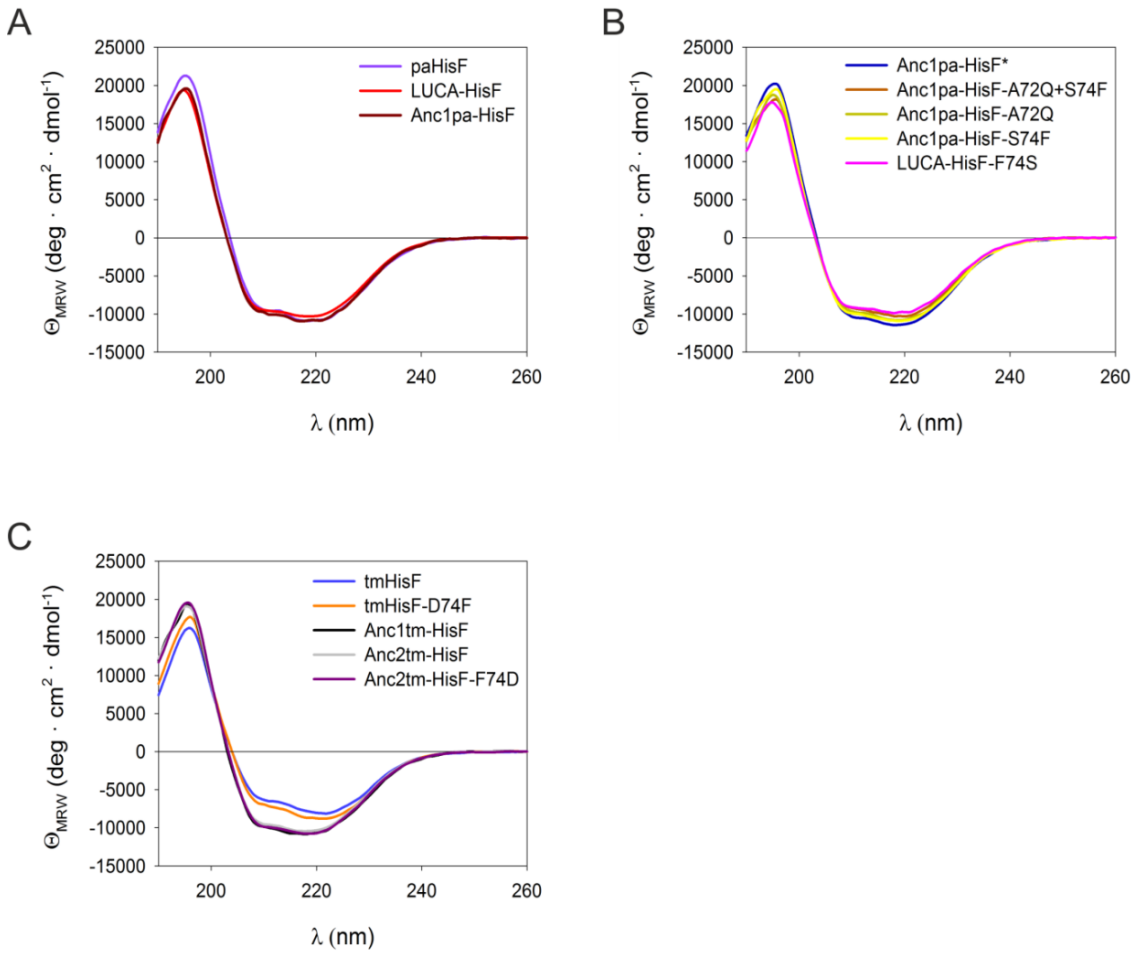

Figure 5.7: Far-UV circular dichroism spectra of HisF proteins used for fluorescence titration with zmHisH. Spectra with proteins at a concentration of 20 μM or 30 μM were recorded in 10 mM potassium phosphate, pH 7.5, at 25 °C. The shown curves are the result of five accumulations and subsequent smoothing. All spectra are indicative of well-defined secondary structures, demonstrating that the proteins are natively folded.

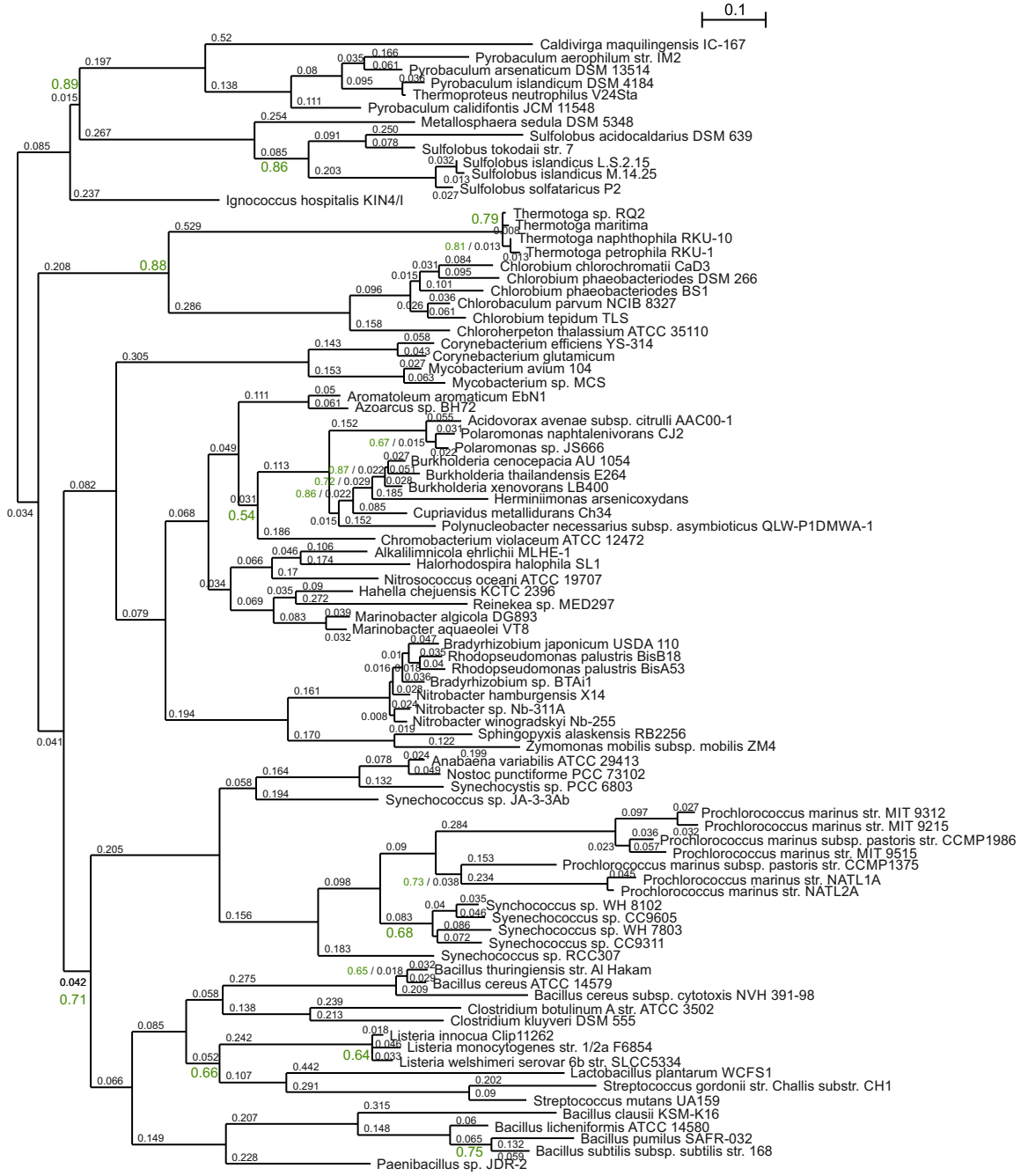

Figure 5.8: Phylogenetic tree used for reconstruction of ancestral HisF sequences after optimization with `FastML`. The indicated lengths of the individual branches correspond to the rate of mutations per site. Posterior probabilities at the splits below 0.95 are shown in green. The horizontal bar indicates the branch length that corresponds to 0.1 mutations per site.

Table 5.2: Nucleotide and amino acid sequences for Anc1pa-HisF, Anc1pa-HisF*, Anc1tm-HisF, and Anc2tm-HisF. The nucleotide sequences were optimized for expression in *E. coli*. The introduced restriction sites for *NdeI* (5') and *XhoI* (3') are in bold, and the introduced stop codon is in italics. **Sequences can be found as Digital Supplemental Data on the provided data storage medium.**

Table 5.3: Aligned sequences of modern HisF proteins used for phylogenetic analysis and of LUCA-HisF. Sequences can be found as Digital Supplemental Data on the provided data storage medium.

Table 5.4: Log likelihood values and posterior probabilities of the reconstructed ancestral sequences at each position. Output is from the joint reconstruction by means of `FastML`. For each position of the alignment, the joint log likelihood from the character reconstruction (first step) and the joint posterior probabilities from the indel reconstruction (second step) are listed. The total joint log likelihood is -15466.2 and is the sum of the position specific log likelihood values. **Sequences can be found as Digital Supplemental Data on the provided data storage medium.**

Residue position	tmHisF:tmHisH	LUCA-HisF:zmHisH	Residues in Anc1pa-HisF
2	L 0.15	L 0.8	L
5		R 0.24	R
45	D 1.15	D 0.96	D
46		E 0.63	E
73	I 0.19		V
74	**D 0.41**	**F 0.96**	S
75	I 0.39	I 0.73	I
76	P 1.48	P 1.35	P
77	F 0.38	L 0.57	L
99		K 0.63	K
123		Q 0.34	Q
167		E 0.24	E
250	R 1.51	R 0.22	R

Table 5.5: Hotspot prediction for HisF residues in ImGPS interfaces. The server `KFC2` was used to analyze the interfaces of the complexes tmHisF:tmHisH (PDB ID: 1GPW, CD) and a homology model of LUCA-HisF:zmHisH. For the residue positions given in column 1, columns 2, and 3 are listing the corresponding residues and the `KFC2A` confidence scores. The higher these scores are, the more likely is the residue a hotspot; if values are missing the residues are unsuspicious in the corresponding complexes. For comparison, the Anc1pa-HisF residues are listed in the last column. The row that corresponds to the Phe74Ser mutation distinguishing the interfaces of LUCA-HisF and Anc1pa-HisF is printed in bold.

Chapter 6

Comprehensive Summary, Discussion and Outlook

Within this work, several diverse applications of ASR are introduced and explained. In addition to the detailed protocol to compile input data and a reliable phylogenetic tree for ASR, a method to elucidate hotspots for protein-protein interactions is given (**chapter 2**). Thus, not only the general ASR procedure is addressed, but also pitfalls, and these comments ease the processing of the data. The identification of an interaction hotspot for ImGPS is an example for a vertical approach and the corresponding protocol is explained in detail (**chapter 2**).

In order to standardize the process of ASR even further, a method to choose a sequence set leading to a suitable phylogenetic tree was developed. The enormous amount of available sequences makes it necessary to filter drastically and to pick sequences that possess a strong phylogenetic signal. To support the user, we have developed `FitSS4ASR` (**chapter 3**) and demonstrated that our program compiles data sets, which allow for a reliable ASR. Although not perfert, `FitSS4ASR` replaces manual sequence selection by a semi-automatic approach.

Using the above-introduced protocols ASR was applied to gain insights into the function and sophistication of ancestral proteins. The classical application of ASR is the characterization of primordial enzymes, which was used here on the example of TS (**chapter 4**). We managed to reconstruct the first ancestral complex, consisting of TrpA and TrpB that form a heterotetramer. Analyzing the crystal structure, we could show that LBCA TS was already a sophisticated protein complex. Moreover, not only protein binding and activity was established, but also an allosteric mechanism to control reciprocally the functions in each subunit.

In addition to the characterization of primordial enzymes, ASR can be utilized to elucidate residues crucial for specific features. Based on the detailed protocol given in **chapter 2**, **chapter 5** gives a closer look on ImGPS, consisting of HisF and HisH that form a heterodimer. With the help of the phylogenetic tree and ancestral HisF sequences, we were able to identify a hotspot of ancestral and extant HisF subunits that is crucial for binding zmHisH. Here, ASR helped to reduce 103 candidate positions down to 9 positions located in the interface. Utilizing

further bioinformatic predictions, we were able to narrow down them to just two positions. Biochemical experiments revealed one of the two positions as an interaction hotspot and indicates the suitability of ASR for hotspot detection.

Similar to the workflow given in this work, several protocols have been published in reviews (Merkl and Sterner, 2016; Joy et al., 2016; Gumulya and Gillam, 2017; Wheeler et al., 2016; Hochberg and Thornton, 2017). However, standard conditions illustrated in **chapter 2** are only specified in few of the published protocols. As the reliability of ASR depends on different sources of errors, it is important to use a suitable data set, a reliable phylogenetic tree, and to apply well-defined quality standards. A main topic discussed frequently (Williams et al., 2006; Hanson-Smith et al., 2010; Bar-Rogovsky et al., 2015) is the accuracy of ASR, which can be degraded for several reasons. A first source of error is the MSA, as the alignment quality affects the accuracy of ASR (Vialle et al., 2018). Especially, the positioning of gaps can lead to quite different ancestral sequences, which demands their handling with special care. There are only a few methods available, like `prank` (Löytynoja and Goldman, 2008), which addresses the correct placement of gaps or `FastML` (Ashkenazy et al., 2012), which supports the careful reconstruction of indels. A further source of error is a too-low reliability of the phylogenetic tree caused by sampling or systematic errors (Huson and Bryant, 2006), which can affect mutation rates in the tree and the amino acid distribution during the reconstruction (Pollock et al., 2012; Goldstein et al., 2015). Checking for reliability can be done with nonparametric bootstrap (Felsenstein, 1985) or multiple samples from the posterior distribution for Bayesian analysis (Rannala and Yang, 1996).

Although, it is not possible to remove all sources of errors, ASR had already its breakthrough in computational biology. This success was due to work demonstrating that properties of ancestors are robust against phylogenetic uncertainty (Hanson-Smith et al., 2010; Risso et al., 2013) and errors in sequence prediction (Akanuma et al., 2013; Gaucher et al., 2008). Additionally, the thermostability of ancestral proteins has been shown to be robust against uncertainties of ASR (Hart et al., 2014) and against the absence of thermophilic sequences in the input sequence set (Akanuma et al., 2015).

Despite the robustness of ASR against several sources of errors, the prediction fidelity of ASR need to be assessed. One way of evaluating the prediction fidelity of ASR is the residue-wise assessment of the posterior probabilities (Bar-Rogovsky et al., 2015), as it has been done in **chapter 5**. Alternatively, sequence ensembles have been generated that varied at critical residue positions according to these amino acid probabilities and constituted an ensemble of near-ancestor sequences corresponding to one ancestral node. The experimental characterization of these proteins helps to assess the prediction reliability and to identify ambiguous positions that could affect the phenotype of the ancestor (Bar-Rogovsky et al., 2015). Ambiguous positions can be identified by characterizing a phenotypic trait, e. g. fluorescence property during the evolution (Chang et al., 2007; Ugalde et al., 2004).

Taken together, accuracy and limitations of ASR have been extensively surveyed in the literature by using different approaches. Due to the reliability of the method, the initial application of resurrecting primordial proteins (**chapter 4**), is no longer the exclusive goal of ASR. A more recent application is the vertical approach, which is used to trace back the evolution from the most ancient sequences to extant ones. After detecting a change in a specific property, crucial positions responsible for this change can be identified and then assessed biochemically, as it is described in the REAP approach (Cole et al., 2013) or in **chapter 5**. A similar approach is to detect crucial positions with a library approach, as it has been shown for the red fluorescence in GFP-like proteins (Field and Matz, 2010) and sulfotransferases (Alcolombri et al., 2011). However, ASR can also be utilized for protein engineering (Zhang et al., 2013). Due to the generally higher thermostability of reconstructed proteins and their tendency to provide promiscuous functions (Wheeler et al., 2016), these proteins represent promising scaffolds for protein design. Thus, ASR was used to design more stable proteins (Wijma et al., 2013), new specificity (Alcolombri et al., 2011), enhanced pharmaceutical properties of protein drugs (Zakas et al., 2017), and novel or enhanced biomolecular function (Cole and Gaucher, 2011).

Can one further improve the methods of ASR? The reconstructions performed for TS (**chapter 4**) and ImGPS (**chapter 5**) showed already high accuracy even for the reconstruction of two subunits. The applied protocol was based on a so-called gene tree, which does only contain the sequences of one gene (product) and is the common approach of ASR. It remains to be shown whether a species tree-aware gene tree, a combination of a species tree and gene tree, improves ASR accuracy (Groussin et al., 2014). In addition to the method of tree construction, the algorithm used for the computation of the ancestral sequences is of great importance. There are two main reconstruction methods available, namely the joint and the marginal reconstruction (Pupko et al., 2000). The marginal reconstruction is able to reconstruct ancestral indels and computes the ancestral states for each internal node of the tree separately, which can introduce a bias. Thus, as the alternative `SubRecon` (Goldstein and Kelso, 2018) was developed to infer ancestral states by means of joint reconstruction, which computes the ancestral states of all nodes at once, while preserving the reconstruction of indels. It remains to be shown that `SubRecon` lead to more accurately reconstructed indels.

Further improvements are expected in the context of selecting extant sequences, which is a critical step of the ASR protocol (**chapter 3**). The current version of `FitSS4ASR` allows for a semi-automatic compilation of sequence sets that promise a robust tree and thus, a reliable reconstruction. Hence, this approach is less error-prone than a manual approach. A further improvement of this program could be the integration of additional filtering techniques, e. g. based on the evolutionary distribution of the sequence within the tree of life or the implementation of faster Bayesian analysis programs to improve the runtime. More importantly, sophisticated methods for the assessment of tree topology (Lemoine et al., 2018) or qualitatively improved phylogenetic programs should be integrated: Darriba et al. (2018) showed the necessity of reliable, high quality scientific software especially when using huge phylogenetic data sets. In

addition, one could consider the integration of multi-gene or species trees: As the input data of `FitSS4ASR` consists of a sequence set, not only gene trees but also multiple gene trees or species trees can be calculated by concatenating the corresponding sequences beforehand. Independent of these hypothetical improvements, the current version of the program assesses the suitability of a sequence set by means of scores that are easy to interpret. Therefore, not only experts but also novices are able to utilize ASR. Thus, `FitSS4ASR` has the potential to further disseminate ASR among life-science to solve important and quite different problems.

Digital Supplemental Data

The following Digital Supplemental Data files can be found on the data storage medium attached to the back cover of the thesis.

Chapter 3:	Tables 3.2 - 3.5
Chapter 4:	Figure 4.6
	Table 4.5
Chapter 5:	Tables 5.2 - 5.4

Abbreviations

$\Delta\Delta G$	interaction free energy
AncGGGPS1	ancestors of group I GGGPS sequences
AncGGGPS2	ancestors of group II GGGPS sequences
AICAR	5-aminoimidazole-4-carboxamide ribotide
ASR	ancestral sequence reconstruction
CA	common ancestor
CD	circular dichroism
COMM domain	communication domain
DNA	deoxyribonucleic acid
DSC	differential scanning calorimetry
E. coli	*Escherichia coli*
ecTS	tryptophan synthase from *Escherichia coli*
EPPS	3-[4-(2-Hydroxyethyl)piperazin-1-yl]propane-1-sulfonic acid
G1P	sn-glycerol-1-phosphate
G3P	sn-glycerol-3-phosphate
GAP	glyceraldehyde-3-phosphate
GGGPS	geranylgeranylglyceryl phosphate synthase
GFP	green flourescent protein
GP	glycerol 3-phosphate
HEPES	2-[4-(2-hydroxyethyl)piperazin-1-yl]ethanesulfonic acid
HGT	horizontal gene transfer
HisF	imidazole glycerol phosphate synthase subunit HisF
HisH	imidazole glycerol phosphate glutaminase subunit HisH
IGP	indole-3-glycerol-phosphate

ImGPS	imidazole glycerol phosphate synthase
IPTG	isopropyl-β-thiogalactopyranoside
K_d	dissociation constant
LB medium	Luria broth medium
LBA	long branch attraction
LBCA	last bacterial common ancestor
LCA	last common ancestor
LUCA	last universal common ancestor
maxdiff	maximum difference of posterior probabilities of tree bipartitions
MCMC	Markov Chain Monte Carlo
ML	maximum likelihood
MPA	most probable ancestor
MR	molecular weights
MSA	multiple sequence alignment
MVC	monovalent cation binding loop
NAD^+	oxidized nicotinamide adenine dinucleotide
NADH	reduced nicotinamide adenine dinucleotide
NMHA	N-methylacetohydroxamic acid
pa	*Pyrobaculum arsenaticum*
paHisF	imidazole glycerol phosphate synthase subunit HisF from *Pyrobaculum arsenaticum*
paHisH	imidazole glycerol phosphate glutaminase subunit HisH from *Pyrobaculum arsenaticum*
PEG	Polyethylenglycol
PLP	pyridoxal phosphate
PPI	protein-protein interaction
PPIs	protein-protein interfaces

PRFAR	N'-[(5'-phosphoribulosyl)-formimino]-5-aminoimidazole-4-carboxamid-ribonucleotide
RBIC	relative bipartition information criterion
S. typhimurium	*Salmonella typhimurium*
SEC	size exclusion chromatography
SeqId	sequence identity
stTS	tryptophan synthase from *Salmonella typhimurium*
TS	tryptophan synthase
tm	*Thermotoga maritima*
tmHisF	imidazole glycerol phosphate synthase subunit HisF from *Thermotoga maritima*
tmHisH	imidazole glycerol phosphate glutaminase subunit HisH from *Thermotoga maritima*
Z. mobilis	*Zymomonas mobilis*
zm	*Zymomonas mobilis*
zmHisH	imidazole glycerol phosphate glutaminase subunit HisH from *Zymomonas mobilis*

References

Abascal, F., R. Zardoya, and D. Posada (2005). ProtTest: selection of best-fit models of protein evolution. *Bioinformatics*, 21(9):2104–5.

Aberer, A. J., D. Krompass, and A. Stamatakis (2012). Pruning rogue taxa improves phylogenetic accuracy: an efficient algorithm and webservice. *Systematic biology*, 62(1):162–166.

Adams, P. D., R. W. Grosse-Kunstleve, L.-W. Hung, T. R. Ioerger, A. J. McCoy, N. W. Moriarty, R. J. Read, J. C. Sacchettini, N. K. Sauter, and T. C. Terwilliger (2002). PHENIX: building new software for automated crystallographic structure determination. *Acta Crystallogr. Section D, Biol. Crystallogr.*, 58(11):1948–54.

Akanuma, S., Y. Nakajima, S. Yokobori, M. Kimura, N. Nemoto, T. Mase, K. Miyazono, M. Tanokura, and A. Yamagishi (2013). Experimental evidence for the thermophilicity of ancestral life. *Proceedings of the National Academy of Sciences, USA*, 110(27):11067–72.

Akanuma, S., S.-i. Yokobori, Y. Nakajima, M. Bessho, and A. Yamagishi (2015). Robustness of predictions of extremely thermally stable proteins in ancient organisms. *Evolution*, 69(11):2954–2962.

Alcolombri, U., M. Elias, and D. S. Tawfik (2011). Directed evolution of sulfotransferases and paraoxonases by ancestral libraries. *Journal of molecular biology*, 411(4):837–853.

Ali, R. H., M. Bark, J. Miró, S. A. Muhammad, J. Sjöstrand, S. M. Zubair, R. M. Abbas, and L. Arvestad (2017). VMCMC: a graphical and statistical analysis tool for markov chain monte carlo traces. *BMC Bioinformatics*, 18(1):97–105.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410.

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402.

Anderson, K. S., E. W. Miles, and K. A. Johnson (1991). Serine modulates substrate channeling in tryptophan synthase. A novel intersubunit triggering mechanism. *The Journal of biological chemistry*, 266(13):8020–33.

Apweiler, R., A. Bairoch, C. H. Wu, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. J. Martin, D. A. Natale, C. O'Donovan, N. Redaschi, and L. S. Yeh (2004). UniProt: the universal protein knowledgebase. *Nucleic Acids Research*, 32(Database issue):D115–9.

Ashkenazy, H., O. Penn, A. Doron-Faigenboim, O. Cohen, G. Cannarozzi, O. Zomer, and T. Pupko (2012). FastML: a web server for probabilistic reconstruction of ancestral sequences. *Nucleic Acids Res*, 40(Web Server issue):W580–4.

Aumentado-Armstrong, T. T., B. Istrate, and R. A. Murgita (2015). Algorithmic approaches to protein-protein interaction site prediction. *Algorithm Mol Biol*, 10:7.

Bar-Rogovsky, H., A. Stern, O. Penn, I. Kobl, T. Pupko, and D. S. Tawfik (2015). Assessing the prediction fidelity of ancestral reconstruction by a library approach. *Protein Eng Des Sel*, 28(11):507–18.

Battistuzzi, F. U., A. Feijao, and S. B. Hedges (2004). A genomic timescale of prokaryote evolution: insights into the origin of methanogenesis, phototrophy, and the colonization of land. *BMC Evolutionary Biology*, 4:44.

Beismann-Driemeyer, S. and R. Sterner (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(23):20387–96.

Bell, E. A., P. Boehnke, T. M. Harrison, and W. L. Mao (2015). Potentially biogenic carbon preserved in a 4.1 billion-year-old zircon. *Proceedings of the National Academy of Sciences*, 112(47):14518–14521.

Bergsten, J. (2005). A review of long-branch attraction. *Cladistics*, 21(2):163–193.

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

Bornscheuer, U. T., G. W. Huisman, R. J. Kazlauskas, S. Lutz, J. C. Moore, and K. Robins (2012). Engineering the third wave of biocatalysis. *Nature*, 485(7397):185–94.

Bouckaert, R., J. Heled, D. Kuhnert, T. Vaughan, C. H. Wu, D. Xie, M. A. Suchard, A. Rambaut, and A. J. Drummond (2014). BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Comput Biol*, 10(4):e1003537.

Boussau, B., S. Blanquart, A. Necsulea, N. Lartillot, and M. Gouy (2008). Parallel adaptations to high temperatures in the archaean eon. *Nature*, 456(7224):942–5.

Bradshaw, R. T., B. H. Patel, E. W. Tate, R. J. Leatherbarrow, and I. R. Gould (2011). Comparing experimental and computational alanine scanning techniques for probing a prototypical protein-protein interaction. *Prot Eng Des Sel*, 24(1-2):197–207.

Bridgham, J. T., S. M. Carroll, and J. W. Thornton (2006). Evolution of hormone-receptor complexity by molecular exploitation. *Science*, 312(5770):97–101.

Brinkmann, H., M. Van der Giezen, Y. Zhou, G. P. De Raucourt, and H. Philippe (2005). An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. *Systematic biology*, 54(5):743–757.

Brooks, D. J. and E. A. Gaucher (2007). *A thermophilic last universal ancestor inferred from its estimated amino acid composition*, Pp. 200–207. Oxford: Oxford University Press.

Busch, F., C. Rajendran, K. Heyn, S. Schlee, R. Merkl, and R. Sterner (2016). Ancestral tryptophan synthase reveals functional sophistication of primordial enzyme complexes. *Cell Chem Biol*, 23(6):709–15.

Busch, F., C. Rajendran, O. Mayans, P. Löffler, R. Merkl, and R. Sterner (2014). TrpB2 enzymes are *O*-Phospho-L-serine dependent tryptophan synthases. *Biochemistry*, 53(38):6078–6083.

Capella-Gutiérrez, S., J. M. Silla-Martínez, and T. Gabaldón (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, 25(15):1972–1973.

Casino, P., D. Niks, H. Ngo, P. Pan, P. Brzovic, L. Blumenstein, T. R. Barends, I. Schlichting, and M. F. Dunn (2007). Allosteric regulation of tryptophan synthase channeling: the internal aldimine probed by trans-3-indole-3'-acrylate binding. *Biochemistry*, 46(26):7728–39.

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

Chang, B., M. Matz, S. Field, J. Müller, and I. van Hazel (2007). Dealing with model uncertainty in reconstructing ancestral proteins in the laboratory: examples from ancestral visual pigments and GFP-like proteins. *Ancestral Sequence Reconstruction. Oxford: Oxford University Press. p*, Pp. 164–180.

Chaudhuri, B. N., S. C. Lange, R. S. Myers, S. V. Chittur, V. J. Davisson, and J. L. Smith (2001). Crystal structure of imidazole glycerol phosphate synthase: a tunnel through a $(\beta/\alpha)_8$ barrel joins two active sites. *Structure*, 9(10):987–97.

Chittur, S. V., Y. Chen, and V. J. Davisson (2000). Expression and purification of imidazole glycerol phosphate synthase from *Saccharomyces cerevisiae*. *Protein Expression and Purification*, 18(3):366–77.

Ciccarelli, F. D., T. Doerks, C. von Mering, C. J. Creevey, B. Snel, and P. Bork (2006). Toward automatic reconstruction of a highly resolved tree of life. *Science*, 311(5765):1283–7.

Clackson, T. and J. A. Wells (1995). A hot spot of binding energy in a hormone-receptor interface. *Science*, 267(5196):383–6.

Cole, M. F., V. E. Cox, K. L. Gratton, and E. A. Gaucher (2013). Reconstructing evolutionary adaptive paths for protein engineering. In *Enzyme Engineering*, Pp. 115–125. Springer.

Cole, M. F. and E. A. Gaucher (2011). Exploiting models of molecular evolution to efficiently direct protein engineering. *Journal of molecular evolution*, 72(2):193–203.

Creighton, T. E. (1970). A steady-state kinetic investigation of the reaction mechanism of the tryptophan synthetase of *Escherichia coli*. *European Journal of Biochemistry*, 13(1):1–10.

Darriba, D., T. Flouri, and A. Stamatakis (2018). The state of software for evolutionary biology. *Molecular Biology and Evolution*, 35(5):1037–1046.

Darwin, C. (1837). Notebook b: Transmutation of species (1837-1838). *Darwin Online, URL: http://darwin-online. org. uk*.

Darwin, C. (1859). *On the Origin of Species*. London: John Murray.

Davis, I. W., A. Leaver-Fay, V. B. Chen, J. N. Block, G. J. Kapral, X. Wang, L. W. Murray, W. B. Arendall `III`, J. Snoeyink, J. S. Richardson, et al. (2007). MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Research*, 35(Web Server issue):W375–83.

Dayhoff, M. O., R. M. Schwartz, and B. C. Orcutt (1978). *A model of evolutionary change in proteins*, volume 5, Pp. 345–352. Washington, DC: National Biomedical Research Foundation.

De Genst, E., D. Areskoug, K. Decanniere, S. Muyldermans, and K. Andersson (2002). Kinetic and affinity predictions of a protein-protein interaction using multivariate experimental design. *The Journal of biological chemistry*, 277(33):29897–907.

de Vienne, D. M., S. Ollier, and G. Aguileta (2012). Phylo-MCOA: a fast and efficient method to detect outlier genes and species in phylogenomics using multiple co-inertia analysis. *Molecular biology and evolution*, 29(6):1587–1598.

del Sol, A., C.-J. Tsai, B. Ma, and R. Nussinov (2009). The origin of allosteric functional modulation: multiple pre-existing pathways. *Structure*, 17(8):1042–1050.

Dereeper, A., V. Guignon, G. Blanc, S. Audic, S. Buffet, F. Chevenet, J. F. Dufayard, S. Guindon, V. Lefort, M. Lescot, J. M. Claverie, and O. Gascuel (2008). Phylogeny.fr: robust phylogenetic analysis for the non-specialist. *Nucleic Acids Res*, 36(Web Server issue):W465–9.

Dierkers, A. T., D. Niks, I. Schlichting, and M. F. Dunn (2009). Tryptophan synthase: structure and function of the monovalent cation site. *Biochemistry*, 48(46):10997–1010.

Douangamath, A., M. Walker, S. Beismann-Driemeyer, M. C. Vega-Fernandez, R. Sterner, and M. Wilmanns (2002). Structural evidence for ammonia tunneling across the $(\beta\alpha)_8$ barrel of the imidazole glycerol phosphate synthase bienzyme complex. *Structure*, 10(2):185–93.

Dunn, M. F. (2012). Allosteric regulation of substrate channeling and catalysis in the tryptophan synthase bienzyme complex. *Archives of Biochemistry and Biophysics*, 519(2):154–66.

Dunn, M. F., V. Aguilar, P. Brzovic, W. F. Drewe Jr, K. F. Houben, C. A. Leja, and M. Roy (1990). The tryptophan synthase bienzyme complex transfers indole between the α- and β-sites via a 25-30 Å long tunnel. *Biochemistry*, 29(37):8598–607.

Edgar, R. C. and S. Batzoglou (2006). Multiple sequence alignment. *Current Opinion in Structural Biology*, 16(3):368–73.

Eick, G. N., J. K. Colucci, M. J. Harms, E. A. Ortlund, and J. W. Thornton (2012). Evolution of minimal specificity and promiscuity in steroid hormone receptors. *PLoS Genetics*, 8(11):e1003072.

Eme, L., A. Spang, J. Lombard, C. W. Stairs, and T. J. Ettema (2017). Archaea and the origin of eukaryotes. *Nature Reviews Microbiology*, 15(12):711.

Emsley, P. and K. Cowtan (2004). Coot: model-building tools for molecular graphics. *Acta Crystallographica. Section D: Biological Crystallography*, 60(Pt 12 Pt 1):2126–32.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of molecular evolution*, 17(6):368–376.

Felsenstein, J. (1985). Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, 39(4):783–791.

Field, S. F. and M. V. Matz (2010). Retracing evolution of red fluorescence in GFP-like proteins from *Faviina* corals. *Molecular Biology and Evolution*, 27(2):225–33.

Frickey, T. and A. N. Lupas (2004). Phylogenie: automated phylome generation and analysis. *Nucleic Acids Res*, 32(17):5231–8.

Frumhoff, P. C. and H. K. Reeve (1994). Using phylogenies to test hypotheses of adaptation: a critique of some current proposals. *Evolution*, 48(1):172–180.

Fuellen, G., M. Spitzer, P. Cullen, and S. Lorkowski (2005). Correspondence of function and phylogeny of ABC proteins based on an automated analysis of 20 model protein data sets. *Proteins*, 61(4):888–99.

Gaucher, E. A., S. Govindarajan, and O. K. Ganesh (2008). Palaeotemperature trend for precambrian life inferred from resurrected proteins. *Nature*, 451(7179):704–7.

Gehling, J. G., G. M. Narbonne, and M. M. Anderson (2000). The first named ediacaran body fossil, *Aspidella terranovica*. *Palaeontology*, 43(3):427–456.

Gerlt, J. A. (2017). Genomic enzymology: web tools for leveraging protein family sequence-function space and genome context to discover novel functions. *Biochemistry*, 56(33):4293–4308.

Goldstein, R. A. and J. Kelso (2018). SubRecon: ancestral reconstruction of amino acid substitutions along a branch in a phylogeny. *Bioinformatics*, 1:3.

Goldstein, R. A., S. T. Pollard, S. D. Shah, and D. D. Pollock (2015). Nonadaptive amino acid convergence rates decrease over time. *Molecular biology and evolution*, 32(6):1373–1381.

Goremykin, V. V., S. V. Nikiforova, and O. R. Bininda-Emonds (2010). Automated removal of noisy data in phylogenomic analyses. *J Mol Evol*, 71(5-6):319–31.

Groussin, M., J. K. Hobbs, G. J. Szöllősi, S. Gribaldo, V. L. Arcus, and M. Gouy (2014). Toward more accurate ancestral protein genotype–phenotype reconstructions with the use of species tree-aware gene trees. *Molecular biology and evolution*, 32(1):13–22.

Guerois, R., J. E. Nielsen, and L. Serrano (2002). Predicting changes in the stability of proteins and protein complexes: A study of more than 1000 mutations. *Journal of Molecular Biology*, 320(2):369–87.

Guindon, S., J. F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk, and O. Gascuel (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol*, 59(3):307–21.

Gumulya, Y. and E. M. Gillam (2017). Exploring the past and the future of protein evolution with ancestral sequence reconstruction: the 'retro' approach to protein engineering. *Biochem J*, 474(1):1–19.

Gupta, R. S. (1998). Protein phylogenies and signature sequences: A reappraisal of evolutionary relationships among archaebacteria, eubacteria, and eukaryotes. *Microbiology and Molecular Biology Reviews*, 62(4):1435–91.

Hanson-Smith, V. and A. Johnson (2016). Phylobot: A web portal for automated phylogenetics, ancestral sequence reconstruction, and exploration of mutational trajectories. *PLoS Comput Biol*, 12(7):e1004976.

Hanson-Smith, V., B. Kolaczkowski, and J. W. Thornton (2010). Robustness of ancestral sequence reconstruction to phylogenetic uncertainty. *Molecular Biology and Evolution*, 27(9):1988–99.

Harms, M. J., G. N. Eick, D. Goswami, J. K. Colucci, P. R. Griffin, E. A. Ortlund, and J. W. Thornton (2013). Biophysical mechanisms for large-effect mutations in the evolution of steroid hormone receptors. *Proceedings of the National Academy of Sciences, USA*, 110(28):11475–80.

Harms, M. J. and J. W. Thornton (2010). Analyzing protein structure and function using ancestral gene reconstruction. *Current Opinion in Structural Biology*, 20(3):360–6.

Hart, K. M., M. J. Harms, B. H. Schmidt, C. Elya, J. W. Thornton, and S. Marqusee (2014). Thermodynamic system drift in protein evolution. *PLoS Biology*, 12(11):e1001994.

Hennig, W. (1965). Phylogenetic systematics. *Annual review of entomology*, 10(1):97–116.

Hettwer, S. and R. Sterner (2002). A novel tryptophan synthase β-subunit from the hyperthermophile *Thermotoga maritima*. Quaternary structure, steady-state kinetics, and putative physiological role. *Journal of Biological Chemistry*, 277(10):8194–8201.

Ho, S. N., H. D. Hunt, R. M. Horton, J. K. Pullen, and L. R. Pease (1989). Site-directed mutagenesis by overlap extension using the polymerase chain reaction. *Gene*, 77(1):51–9.

Ho, S. Y. and L. S. Jermiin (2004). Tracing the decay of the historical signal in biological sequence data. *Systematic Biology*, 53(4):623–637.

Hobbs, J. K., C. Shepherd, D. J. Saul, N. J. Demetras, S. Haaning, C. R. Monk, R. M. Daniel, and V. L. Arcus (2012). On the origin and evolution of thermophily: reconstruction of functional precambrian enzymes from ancestors of *Bacillus*. *Molecular Biology and Evolution*, 29(2):825–35.

Hochberg, G. K. A. and J. W. Thornton (2017). Reconstructing ancient proteins to understand the causes of structure and function. *Annu Rev Biophys*, 46:247–269.

Hoffmann, A. A. and C. M. Sgrò (2011). Climate change and evolutionary adaptation. *Nature*, 470(7335):479.

Holder, M. and P. O. Lewis (2003). Phylogeny estimation: traditional and Bayesian approaches. *Nature reviews genetics*, 4(4):275.

Holinski, A., K. Heyn, R. Merkl, and R. Sterner (2017). Combining ancestral sequence reconstruction with protein design to identify an interface hotspot in a key metabolic enzyme complex. *Proteins*, 85(2):312–321.

Holmes, I. H. (2017). Historian: accurate reconstruction of ancestral sequences and evolutionary rates. *Bioinformatics*, 33(8):1227–1229.

Huang, X., H. M. Holden, and F. M. Raushel (2001). Channeling of substrates and intermediates in enzyme-catalyzed reactions. *Annual Review of Biochemistry*, 70:149–80.

Hug, L. A., B. J. Baker, K. Anantharaman, C. T. Brown, A. J. Probst, C. J. Castelle, C. N. Butterfield, A. W. Hernsdorf, Y. Amano, K. Ise, et al. (2016). A new view of the tree of life. *Nature microbiology*, 1:16048.

Hunter, S., 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, and S. Y. Yong (2012). InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Research*, 40(Database issue):D306–12.

Huson, D. H. and D. Bryant (2006). Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution*, 23(2):254–67.

Hyde, C. C., S. A. Ahmed, E. A. Padlan, E. W. Miles, and D. R. Davies (1988). Three-dimensional structure of the tryptophan synthase $\alpha_2\beta_2$ multienzyme complex from *Salmonella typhimurium*. *Journal of Biological Chemistry*, 263(33):17857–71.

Janda, J. O., A. Meier, and R. Merkl (2013). CLIPS-4D: a classifier that distinguishes structurally and functionally important residue-positions based on sequence and 3D data. *Bioinformatics*, 29(23):3029–35.

Janin, J., R. Bahadur, and P. Chakrabarti (2008). Protein–protein interaction and quaternary structure. *Quarterly reviews of biophysics*, 41(02):133–180.

Jensen, R. A. (1976). Enzyme recruitment in evolution of new function. *Annual Review of Microbiology*, 30:409–25.

Jeong, C. S. and D. Kim (2012). Reliable and robust detection of coevolving protein residues. *Protein Eng Des Sel*, 25(11):705–13.

Joy, J. B., R. H. Liang, R. M. McCloskey, T. Nguyen, and A. F. Poon (2016). Ancestral reconstruction. *PLoS Comput Biol*, 12(7):e1004763.

Kabsch, W. (2010). Xds. *Acta crystallographica. Section D, Biological crystallography*, 66(Pt 2):125–32.

Katoh, K. and D. M. Standley (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, 30(4):772–80.

Krieger, E., K. Joo, J. Lee, S. Raman, J. Thompson, M. Tyka, D. Baker, and K. Karplus (2009). Improving physical realism, stereochemistry, and side-chain accuracy in homology modeling: Four approaches that performed well in CASP8. *Proteins*, 77 Suppl 9:114–22.

Krissinel, E. and K. Henrick (2007). Inference of macromolecular assemblies from crystalline state. *Journal of Molecular Biology*, 372(3):774–97.

Krüger, D. M. and H. Gohlke (2010). DrugScore^{PPI} webserver: fast and accurate *in silico* alanine scanning for scoring protein-protein interactions. *Nucleic Acids Research*, 38(Web Server issue):W480–6.

Kumar, S., G. Stecher, D. Peterson, and K. Tamura (2012). MEGA-CC: computing core of molecular evolutionary genetics analysis program for automated and iterative data analysis. *Bioinformatics*, 28(20):2685–6.

Kupczok, A. (2011). Split-based computation of majority-rule supertrees. *BMC evolutionary biology*, 11(1):205.

La, D., M. Kong, W. Hoffman, Y. I. Choi, and D. Kihara (2013). Predicting permanent and transient protein-protein interfaces. *Proteins*, 81(5):805–18.

Lane, A. N. and K. Kirschner (1991). Mechanism of the physiological reaction catalyzed by tryptophan synthase from *Escherichia coli*. *Biochemistry*, 30(2):479–84.

Lartillot, N., T. Lepage, and S. Blanquart (2009). Phylobayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics*, 25(17):2286–8.

Lartillot, N. and H. Philippe (2004). A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular biology and evolution*, 21(6):1095–1109.

Lee, D., O. Redfern, and C. Orengo (2007). Predicting protein function from sequence and structure. *Nat Rev Mol Cell Biol*, 8(12):995–1005.

Lemoine, F., J.-B. D. Entfellner, E. Wilkinson, D. Correia, M. D. Felipe, T. Oliveira, and O. Gascuel (2018). Renewing felsenstein's phylogenetic bootstrap in the era of big data. *Nature*, P. 1.

Letunic, I. and P. Bork (2016). Interactive tree of life (iTol) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic acids research*, 44(W1):W242–W245.

Li, G., M. Steel, and L. Zhang (2008). More taxa are not necessarily better for the reconstruction of ancestral character states. *Syst Biol*, 57(4):647–53.

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

Liberles, D. A. (2007). *Ancestral sequence reconstruction*. Oxford: Oxford University Press.

Linde, M., K. Heyn, R. Merkl, R. Sterner, and P. Babinger (2018). Hexamerization of geranylgeranylglyceryl phosphate synthase ensures structural integrity and catalytic activity at high temperatures. *Biochemistry*, 57(16):2335–2348.

Lisi, G. P., G. A. Manley, H. Hendrickson, I. Rivalta, V. S. Batista, and J. P. Loria (2016). Dissecting dynamic allosteric pathways using chemically related small-molecule activators. *Structure*, 24(7):1155–66.

List, F., M. C. Vega, A. Razeto, M. C. Häger, R. Sterner, and M. Wilmanns (2012). Catalysis uncoupling in a glutamine amidotransferase bienzyme by unblocking the glutaminase active site. *Chemistry and Biology*, 19(12):1589–99.

López-García, P. and D. Moreira (2015). Open questions on the origin of eukaryotes. *Trends in ecology & evolution*, 30(11):697–708.

Löytynoja, A. and N. Goldman (2008). Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science*, 320(5883):1632–5.

Massiere, F. and M. A. Badet-Denisot (1998). The mechanism of glutamine-dependent amidotransferases. *Cell Mol Life Sci*, 54(3):205–22.

Merkl, R. (2007). Modelling the evolution of the archaeal tryptophan synthase. *BMC Evolutionary Biology*, 7:59.

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

Miles, E. W. (2001). Tryptophan synthase: a multienzyme complex with an intramolecular tunnel. *Chem Rec*, 1(2):140–51.

Mitchell, A., H.-Y. Chang, L. Daugherty, M. Fraser, S. Hunter, R. Lopez, C. McAnulla, C. McMenamin, G. Nuka, S. Pesseat, et al. (2014). The InterPro protein families database: the classification resource after 15 years. *Nucleic acids research*, 43(D1):D213–D221.

Murshudov, G. N., A. A. Vagin, and E. J. Dodson (1997). Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallographica. Section D: Biological Crystallography*, 53(Pt 3):240–55.

Myers, R. S., J. R. Jensen, I. L. Deras, J. L. Smith, and V. J. Davisson (2003). Substrate-induced changes in the ammonia channel for imidazole glycerol phosphate synthase. *Biochemistry*, 42(23):7013–22.

Needleman, S. B. and C. D. Wunsch (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–53.

Nisbet, E. G. and N. H. Sleep (2001). The habitat and nature of early life. *Nature*, 409(6823):1083–91.

Ochman, H., J. G. Lawrence, and E. A. Groisman (2000). Lateral gene transfer and the nature of bacterial innovation. *nature*, 405(6784):299.

Ortlund, E. A., J. T. Bridgham, M. R. Redinbo, and J. W. Thornton (2007). Crystal structure of an ancient protein: evolution by conformational epistasis. *Science*, 317(5844):1544–8.

Pagel, M., A. Meade, and D. Barker (2004). Bayesian estimation of ancestral character states on phylogenies. *Systematic biology*, 53(5):673–684.

Pal, G., J. L. Kouadio, D. R. Artis, A. A. Kossiakoff, and S. S. Sidhu (2006). Comprehensive and quantitative mapping of energy landscapes for protein-protein interactions by rapid combinatorial scanning. *J Biol Chem*, 281(31):22378–85.

Pauling, L. and E. Zuckerkandl (1963). Chemical paleogenetics: molecular "restoration studies" of extinct forms of life. *Acta Chemica Scandinavica*, 17:9–16.

Perez-Jimenez, R., A. Inglés-Prieto, Z. M. Zhao, I. Sanchez-Romero, J. Alegre-Cebollada, P. Kosuri, S. Garcia-Manyes, T. J. Kappock, M. Tanokura, A. Holmgren, J. M. Sanchez-Ruiz, E. A. Gaucher, and J. M. Fernandez (2011). Single-molecule paleoenzymology probes the chemistry of resurrected enzymes. *Nature Structural & Molecular Biology*, 18(5):592–6.

Perica, T., Y. Kondo, S. P. Tiwari, S. H. McLaughlin, K. R. Kemplen, X. Zhang, A. Steward, N. Reuter, J. Clarke, and S. A. Teichmann (2014). Evolution of oligomeric state through allosteric pathways that mimic ligand binding. *Science*, 346(6216):1254346.

Perriere, G. and M. Gouy (1996). WWW-query: an on-line retrieval system for biological sequence banks. *Biochimie*, 78(5):364–9.

Peterhoff, D., B. Beer, C. Rajendran, E. P. Kumpula, E. Kapetaniou, H. Guldan, R. K. Wierenga, R. Sterner, and P. Babinger (2014). A comprehensive analysis of the geranylgeranylglyceryl phosphate synthase enzyme family identifies novel members and reveals mechanisms of substrate specificity and quaternary structure organization. *Molecular Microbiology*, 92(4):885–99.

Plach, M., B. Reisinger, R. Sterner, and R. Merkl (2016). Long-term persistence of bifunctionality contributes to the robustness of microbial life through exaptation. *PLoS Genet*, 12(1):e1005836.

Plach, M. G., P. Löffler, R. Merkl, and R. Sterner (2015). Conversion of anthranilate synthase into isochorismate synthase: implications for the evolution of chorismate-utilizing enzymes. *Angew Chem Int Ed Engl*, 54(38):11270–4.

Plach, M. G., F. Semmelmann, F. Busch, M. Busch, L. Heizinger, V. H. Wysocki, R. Merkl, and R. Sterner (2017). Evolutionary diversification of protein–protein interactions by interface add-ons. *Proceedings of the National Academy of Sciences*, 114(40):E8333–E8342.

Pollock, D. D., G. Thiltgen, and R. A. Goldstein (2012). Amino acid coevolution induces an evolutionary stokes shift. *Proceedings of the National Academy of Sciences*, 109(21):E1352–E1359.

Potterton, L., S. McNicholas, E. Krissinel, J. Gruber, K. Cowtan, P. Emsley, G. N. Murshudov, S. Cohen, A. Perrakis, and M. Noble (2004). Developments in the CCP4 molecular-graphics project. *Acta Crystallogr. Section D, Biol. Crystallogr.*, 60(Pt 12 Pt 1):2288–94.

Pruitt, K. D., T. Tatusova, W. Klimke, and D. R. Maglott (2009). NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic Acids Research*, 37(suppl_1):D32–6.

Puigbo, P., Y. I. Wolf, and E. V. Koonin (2009). Search for a 'tree of life' in the thicket of the phylogenetic forest. *J Biol*, 8(6):59.

Pupko, T., I. Pe'er, R. Shamir, and D. Graur (2000). A fast algorithm for joint reconstruction of ancestral amino acid sequences. *Mol Biol Evol*, 17(6):890–6.

Pürzer, A., F. Grassmann, D. Birzer, and R. Merkl (2011). Key2ann: a tool to process sequence sets by replacing database identifiers with a human-readable annotation. *Journal of Integrative Bioinformatics (JIB)*, 8(1):35–46.

Raboni, S., S. Bettati, and A. Mozzarelli (2005). Identification of the geometric requirements for allosteric communication between the α- and β-subunits of tryptophan synthase. *Journal of Biological Chemistry*, 280(14):13450–6.

Rambaut, A. (2012). Figtree v1.4.

Randall, R. N., C. E. Radford, K. A. Roof, D. K. Natarajan, and E. A. Gaucher (2016). An experimental phylogeny to benchmark ancestral sequence reconstruction. *Nat Commun*, 7:12847.

Rannala, B. and Z. Yang (1996). Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. *Journal of molecular evolution*, 43(3):304–311.

Raushel, F. M., J. B. Thoden, and H. M. Holden (2003). Enzymes with molecular tunnels. *Accounts of chemical research*, 36(7):539–548.

Reichmann, D., O. Rahat, M. Cohen, H. Neuvirth, and G. Schreiber (2007). The molecular architecture of protein-protein binding sites. *Current Opinion in Structural Biology*, 17(1):67–76.

Reinstein, J., I. R. Vetter, I. Schlichting, P. Roesch, A. Wittinghofer, and R. S. Goody (1990). Fluorescence and NMR investigations on the ligand binding properties of adenylate kinases. *Biochemistry*, 29(32):7440–50.

Reisinger, B., N. Kuzmanovic, P. Löffler, R. Merkl, B. König, and R. Sterner (2014a). Exploiting protein symmetry to design light-controllable enzyme inhibitors. *Angew Chem Int Ed Engl*, 53(2):595–8.

Reisinger, B., J. Sperl, A. Holinski, V. Schmid, C. Rajendran, L. Carstensen, S. Schlee, S. Blanquart, R. Merkl, and R. Sterner (2014b). Evidence for the existence of elaborate enzyme complexes in the paleoarchean era. *J Am Chem Soc*, 136(1):122–9.

Richter, M., M. Bosnali, L. Carstensen, T. Seitz, H. Durchschlag, S. Blanquart, R. Merkl, and R. Sterner (2010). Computational and experimental evidence for the evolution of a $(\alpha\beta)_8$-barrel protein from an ancestral quarter-barrel stabilised by disulfide bonds. *Journal of Molecular Biology*, 398(5):763–773.

Risso, V. A., J. A. Gavira, D. F. Mejia-Carmona, E. A. Gaucher, and J. M. Sanchez-Ruiz (2013). Hyperstability and substrate promiscuity in laboratory resurrections of precambrian β-lactamases. *Journal of the American Chemical Society*, 135(8):2899–902.

Risso, V. A., J. A. Gavira, and J. M. Sanchez-Ruiz (2014). Thermostable and promiscuous precambrian proteins. *Environ Microbiol*, 16(6):1485–9.

Rivera-Rivera, C. J. and J. I. Montoya-Burgos (2016). LS^3: A method for improving phylogenomic inferences when evolutionary rates are heterogeneous among taxa. *Molecular biology and evolution*, 33(6):1625–1634.

Rodríguez-Ezpeleta, N., H. Brinkmann, B. Roure, N. Lartillot, B. F. Lang, and H. Philippe (2007). Detecting and overcoming systematic errors in genome-scale phylogenies. *Systematic Biology*, 56(3):389–399.

Romero-Romero, M. L., V. A. Risso, S. Martinez-Rodriguez, B. Ibarra-Molero, and J. M. Sanchez-Ruiz (2016). Engineering ancestral protein hyperstability. *Biochem J*, 473(20):3611–3620.

Ronquist, F. and J. P. Huelsenbeck (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, 19(12):1572–4.

Saitou, N. and M. Nei (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4):406–425.

Sali, A. and T. L. Blundell (1993). Comparative protein modelling by satisfaction of spatial restraints. *Journal of Molecular Biology*, 234(3):779–815.

Salichos, L. and A. Rokas (2013). Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature*, 497(7449):327.

Sanderson, M. J. and H. B. Shaffer (2002). Troubleshooting molecular phylogenetic analyses. *Annual review of ecology and Systematics*, 33(1):49–72.

Savitzky, A. and M. J. E. Golay (1964). Smoothing and differentiation of data by simplified least squares procedures. *Anal. Cham.*, 36:1627–1639.

Schiaretti, F., S. Bettati, C. Viappiani, and A. Mozzarelli (2004). pH dependence of tryptophan synthase catalytic mechanism: I. The first stage, the β-elimination reaction. *The Journal of biological chemistry*, 279(28):29572–82.

Schneider, T. R., E. Gerhardt, M. Lee, P. H. Liang, K. S. Anderson, and I. Schlichting (1998). Loop closure and intersubunit communication in tryptophan synthase. *Biochemistry*, 37(16):5394–406.

Schrödinger, LLC (2015). The PyMOL molecular graphics system, version 1.7.

Schymkowitz, J., J. Borg, F. Stricher, R. Nys, F. Rousseau, and L. Serrano (2005). The FoldX web server: an online force field. *Nucleic Acids Res*, 33(Web Server issue):W382–8.

Sehnal, D., R. S. Vařeková, K. Berka, L. Pravda, V. Navrátilová, P. Banáš, C.-M. Ionescu, M. Otyepka, and J. Koča (2013). MOLE 2.0: advanced approach for analysis of biomacromolecular channels. *Journal of Cheminformatics*, 5(1):39.

Soltis, P. S. and D. E. Soltis (2003). Applying the bootstrap in phylogeny reconstruction. *Statistical Science*, Pp. 256–267.

Stackhouse, J., S. R. Presnell, G. M. McGeehan, K. P. Nambiar, and S. A. Benner (1990). The ribonuclease from an extinct bovid ruminant. *FEBS Letters*, 262(1):104–6.

Stamatakis, A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, 22(21):2688–90.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313.

Stefanović, S., D. W. Rice, and J. D. Palmer (2004). Long branch attraction, taxon sampling, and the earliest angiosperms: *Amborella* or monocots? *BMC Evolutionary Biology*, 4(1):35.

Swofford, D., G. Olsen, P. Waddell, and D. Hillis (1996). *Phylogenetic inference*, book section 5, Pp. 407–514. Sunderland, MA: Sinauer and Associates, 2nd edition.

Söding, J. (2005). Protein homology detection by HMM-HMM comparison. *Bioinformatics*, 21(7):951–60.

Talevich, E., B. M. Invergo, P. J. Cock, and B. A. Chapman (2012). Bio.Phylo: a unified toolkit for processing, analyzing and visualizing phylogenetic trees in Biopython. *BMC bioinformatics*, 13(1):209.

Tamura, K., D. Peterson, N. Peterson, G. Stecher, M. Nei, and S. Kumar (2011). MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol*, 28(10):2731–9.

Thornton, J. W. (2004). Resurrecting ancient genes: experimental analysis of extinct molecules. *Nature Reviews: Genetics*, 5(5):366–75.

Thornton, J. W., E. Need, and D. Crews (2003). Resurrecting the ancestral steroid receptor: ancient origin of estrogen signaling. *Science*, 301(5640):1714–7.

Ugalde, J. A., B. S. Chang, and M. V. Matz (2004). Evolution of coral pigments recreated. *Science*, 305(5689):1433.

UniProt, C. (2013). Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic Acids Res*, 41(Database issue):D43–7.

Vialle, R. A., A. U. Tamuri, and N. Goldman (2018). Alignment modulates ancestral sequence reconstruction accuracy. *Molecular biology and evolution*, P. msy055.

Wang, W. and B. A. Malcolm (1999). Two-stage PCR protocol allowing introduction of multiple mutations, deletions and insertions using QuikChange Site-Directed Mutagenesis. *BioTechniques*, 26(4):680–2.

Watanabe, K., T. Ohkuri, S. Yokobori, and A. Yamagishi (2006). Designing thermostable proteins: ancestral mutants of 3-isopropylmalate dehydrogenase designed by using a phylogenetic tree. *Journal of Molecular Biology*, 355(4):664–74.

Watanabe, N. and H. Osada (2016). Small molecules that target phosphorylation dependent protein-protein interaction. *Bioorg Med Chem*, 24(15):3246–54.

Waterhouse, A. M., J. B. Procter, D. M. Martin, M. Clamp, and G. J. Barton (2009). Jalview version 2 – a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, 25(9):1189–91.

Webb, B. and A. Sali (2014). Protein structure modeling with MODELLER. *Methods Mol Biol*, 1137:1–15.

Wheeler, L. C., S. A. Lim, S. Marqusee, and M. J. Harms (2016). The thermostability and specificity of ancient proteins. *Current Opinion in Structural Biology*, 38:37–43.

Wiens, J. J. (2005). Can incomplete taxa rescue phylogenetic analyses from long-branch attraction? *Systematic Biology*, 54(5):731–742.

Wijma, H. J., R. J. Floor, and D. B. Janssen (2013). Structure- and sequence-analysis inspired engineering of proteins for enhanced thermostability. *Current opinion in structural biology*, 23(4):588–594.

Wilkinson, M. and M. Crotti (2017). Comments on detecting rogue taxa using RogueNaRok. *Systematics and Biodiversity*, 15(4):291–295.

Williams, P. D., D. D. Pollock, B. P. Blackburne, and R. A. Goldstein (2006). Assessing the accuracy of ancestral protein reconstruction methods. *PLoS Comput Biol*, 2(6):e69.

Wouters, M. A., K. Liu, P. Riek, and A. Husain (2003). A despecialization step underlying evolution of a family of serine proteases. *Molecular cell*, 12(2):343–354.

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*, 24(8):1586–91.

Yokoyama, S., H. Yang, and W. T. Starmer (2008). Molecular basis of spectral tuning in the red- and green-sensitive (M/LWS) pigments in vertebrates. *Genetics*, 179(4):2037–43.

Zakas, P. M., H. C. Brown, K. Knight, S. L. Meeks, H. T. Spencer, E. A. Gaucher, and C. B. Doering (2017). Enhancing the pharmaceutical properties of protein drugs by ancestral sequence reconstruction. *Nat Biotechnol*, 35(1):35–37.

Zalkin, H. and J. L. Smith (1998). Enzymes utilizing glutamine as an amide donor. *Advances in Enzymology and Related Areas of Molecular Biology*, 72:87–144.

Zellner, H., M. Staudigel, T. Trenner, M. Bittkowski, V. Wolowski, C. Icking, and R. Merkl (2012). PresCont: predicting protein-protein interfaces utilizing four residue properties. *Proteins*, 80(1):154–68.

Zhang, X., T. Perica, and S. A. Teichmann (2013). Evolution of protein structures and interactions from the perspective of residue contact networks. *Curr Opin Struct Biol*, 23(6):954–63.

Zhang, Y. (2008). I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics*, 9:40.

Zhu, X. and J. C. Mitchell (2011). KFC2: a knowledge-based hot spot prediction method based on interface solvation, atomic density, and plasticity features. *Proteins*, 79(9):2671–83.

Acknowledgement

First and foremost, I would like to thank my supervisor, Prof. Dr. Rainer Merkl for his valuable support, guidance, and advice throughout my work. It was a great pleasure to discuss all problems face to face but also to puzzle out the research problems in a self-sufficient way. I appreciate his ingenuous nature and the excellent work atmosphere on the professional as well as on the personal level.

I am no less grateful to Dr. Samuel Blanquart. I profoundly appreciate the far-reaching discussions and conversations on all topics regarding phylogeny. He made it possible to visit him for one week, which was a very intensive and productive week leading to a great jump in my research. Furthermore, his mentoring throughout my work has been very fruitful.

I would like to thank Prof. Dr. Christoph Oberprieler for his mentoring throughout my work. I appreciate his slightly different view on phylogeny, which led to interesting questions on my work. In addition, I want to thank him for reviewing this thesis.

I also would like to thank Prof. Dr. Reinhard Sterner for all the biochemical support on my work. He sparked my interest in biochemistry and was always available for advice or discussion.

Financing by the *Deutsche Forschungsgemeinschaft* is gratefully acknowledged.

I sincerely thank my collaboration partners for their valuable contributions to this work: Dr. Florian Busch, who performed the biochemical experiments on the tryptophan synthase and gave me an introduction to the standard procedures in the laboratory; Dr. Alexandra Holinski for the joint work on the identification of hotspots in protein-protein interaction and the many useful discussions on all kinds of experimental and scientific questions; Dr. Mona Linde for the collaboration on the geranylgeranylglyceryl phosphate synthase. A thank you goes also to my colleagues Dr. Andrea Kneuttinger, Enrico Hupfeld, and Michael Schupfner for the intensive and collaborative work on ImGPS and TS.

A very special thank you goes to my colleagues, Leonhard Heizinger and Julian Nazet, who gave me technical assistance and supported me with their manpower and knowledge; Dr. Patrick Löf-

fler for his patience introducing me to the field of bioinformatic and solving problems regarding coding; Dr. Maximilian Plach for answering all kind of questions regarding biochemistry.

My heartfelt thanks to all current and former members of the Merkl and Sterner groups. I enjoyed the excellent working and encouraging atmosphere within all members. I would like to thank all, for advice or discussions, whenever it was necessary. Also thank you to all for an exciting and fun time I had.

I would also like to thank my former colleagues, Robert Greipl and Thomas Siegert, who supported me during my thesis.

Above all, I am deeply grateful to my family for their encouragement to keep on dreaming. Also their unlimited support in all circumstances was priceless. Special thank to my husband Alex: thank you for all your patience, assistance, and for giving me encouragement for everything in my thesis and beyond my work.