Sequence selection by FitSS4ASR alleviates ancestral sequence reconstruction as exemplified for geranylgeranylglyceryl phosphate synthase

Kristina Straub, Mona Linde, Cosimo Kropp, Samuel Blanquart, Patrick Babinger, Rainer Merkl

To cite this version:

Kristina Straub, Mona Linde, Cosimo Kropp, Samuel Blanquart, Patrick Babinger, et al.. Sequence selection by FitSS4ASR alleviates ancestral sequence reconstruction as exemplified for geranylgeranylglyceryl phosphate synthase. Biological Chemistry, De Gruyter, 2019, 400 (3), pp.367-381. 10.1515/hsz-2018-0344 . hal-02284913

HAL Id: hal-02284913
https://hal.inria.fr/hal-02284913

Submitted on 4 Oct 2019

Sequence selection by `FitSS4ASR` alleviates ancestral sequence reconstruction as exemplified for geranylgeranylglyceryl phosphate synthase

Kristina Straub[1], Mona Linde[1], Cosimo Kropp[1], Samuel Blanquart[2], Patrick Babinger[1], Rainer Merkl*[1]

[1] Institute of Biophysics and Physical Biochemistry, University of Regensburg, D-93040 Regensburg, Germany

[2] University of Rennes, Inria, CNRS, IRISA, Rennes F-35000, France

*Corresponding author: Rainer Merkl, Institute of Biophysics and Physical Biochemistry, University of Regensburg, Universitätsstraße 31, D-93053 Regensburg
e-mail: rainer.merkl@ur.de
Phone: (49)941-3086
Fax: (49)941-2813
https://orcid.org/0000-0002-3521-2957

Running title: Sequence selection for ASR

Abstract

For evolutionary studies, but also for protein engineering, ancestral sequence reconstruction (ASR) has become an indispensable tool. The first step of every ASR protocol is the preparation of a representative sequence set containing at most a few hundred recent homologs whose composition determines decisively the outcome of a reconstruction. A common approach for sequence selection consists of several rounds of manual recompilation that is driven by embedded phylogenetic analyses of the varied sequence sets. For ASR of a geranylgeranylglyceryl phosphate synthase, we additionally utilized `FitSS4ASR`, which replaces this time-consuming protocol with an efficient and more rational approach. `FitSS4ASR` applies orthogonal filters to a set of homologs to eliminate outlier sequences and those bearing only a weak phylogenetic signal. To demonstrate the usefulness of `FitSS4ASR`, we determined experimentally the oligomerization state of eight predecessors, which is a delicate and taxon-specific property. Corresponding ancestors deduced in a manual approach and by means of `FitSS4ASR` had the same dimeric or hexameric conformation; this concordance testifies to the efficiency of `FitSS4ASR` for sequence selection. `FitSS4ASR` based results of two other ASR experiments were added to the Supporting Information. Program and documentation are available at https://gitlab.bioinf.ur.de/hek61586/FitSS4ASR.

Keywords

Common ancestors, oligomerization, protein design, protein evolution, sequence space

Introduction

During the last forty years, ancestral sequence reconstruction (ASR) has become 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; 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 results from 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 last universal common ancestor (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., 2014) 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. Many protein families contain functionally diverse members and it is, e. g., difficult to identify residues that are sufficient to switch function by comparing recent sequences, because many of these variations are irrelevant for functional differences (Harms and Thornton, 2010). In contrast, ASR is a "vertical approach", as it takes into account the evolutionary history of the proteins under study. Focusing on the specific substitutions that occurred along the branches leading to different functions is more straightforward. Thus, vertical approaches can drastically reduce experimental efforts to identify key residues as demonstrated for the specificity of hormone receptors (Ortlund et al., 2007), the fluorescence properties of GFP variants (Field and Matz, 2010), or hotspots of protein-protein interfaces (Holinski et al., 2017).

The insight that ancestral proteins are generally more robust and often more versatile (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; for a review see (Gumulya and Gillam, 2017). For example, predecessors of serum paraoxonases and cytosolic sulfotransferases are highly active and functionally diverse and few mutations have been sufficient to introduce a new specificity (Alcolombri et al., 2011). The replacement of residues that change along the branches leading to a functional switch of DNA polymerases has led to a broadened substrate spectrum (Chen et al., 2010). Generally, a stable and robust template is beneficial for directed evolution studies, as mutations that lead to a new function are often destabilizing (Tokuriki et al., 2008). Using ASR, a 30 - 40 °C increase in denaturation temperature has been obtained, which is much larger than that typically gained with alternative protein engineering protocols; for a review see (Wijma et al., 2013).

A crucial element of ASR is the computation of a phylogenetic tree, whose topology is determined by the set of recent sequences and the phylogenetic model, which are both chosen by the user. Once a tree is available, the composition of ancestral sequences that correspond to internal nodes of the tree can be deduced from the leaves, i. e., the recent sequences (Ashkenazy et al., 2012). If the topology of this tree is wrong, some internal nodes correspond to ancestors that never existed and the relative order of mutations can be biased as well. Generally, the quality of the phylogenetic tree can be affected by systematic errors related to the evolutionary model and by stochastic errors, caused by sequences that do not contain enough phylogenetic signal to support a robust tree. One source of systematic errors could be the second step of each ASR (Merkl and Sterner, 2016), which is the creation of an MSA. Often, MSA algorithm rely on a simplistic evolutionary model, whose assumptions specify the data for the subsequent reconstruction steps. Thus, for highly divergent data sets with less than 30% sequence identity (SeqId), the generation of MSAs is a crucial step of each phylogenetic analysis (Essoussi et al., 2008). Recent comparisons of several state-of-the-art methods indicated that all performed equally well for more homogeneous sequence sets (Essoussi et al., 2008; Le et al., 2017) and `MAFFT` (Katoh and Standley, 2013) and `PRANK` (Löytynoja and Goldman, 2008) exhibited best performance for ASR (Vialle et al., 2018).

The protocols implemented for ASR are based on well-proven algorithms and evolutionary models and for each step of the reconstruction process, probability measures allow for the assessment of their outcome; see e. g. (Straub and Merkl, 2019). 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). Thus, by carefully examining the quality of all phases of ASR, one can avoid systematic errors e.g. those caused by fast-evolving sites, rate-variation among sites, or compositional heterogeneity. Removing columns from the MSA and choosing an adequate phylogenetic models are effective means to avoid such errors (Rodríguez-Ezpeleta et al., 2007).To minimize stochastic errors, the set of recent sequences selected as input has to be chosen diligently. ASR can be carried out for DNA, codon, and amino acid sequences. In the following, we exclusively focus on the compilation of a set of amino acid sequences belonging to one protein isoform, which is a key issue of ASR.

Due to their chronological order, mutations of early evolutionary phases have been manifested in many recent members of a family; therefore, it is not necessary to consider all of them for ASR. Indeed, huge numbers of homologs do not necessarily improve the reconstruction of ancestral states (Li et al., 2008); thus not more than 150 to 200 input sequences are commonly picked by the user. However, the current databases offer for many 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, because sequence selection greatly affects the quality of the phylogenetic tree. Quality must meet high standards for ASR (Pagel et al., 2004) and trees showing the strongest phylogenetic signal give more accurate reconstructions (Litsios and Salamin, 2012). To allow for the reconstruction of a long evolutionary time-span and early predecessors, the sequences must represent phylogenetically diverse species. On the other hand, closely related sequences that diverge by few mutations introduce redundancy only and do not support a deep reconstruction. Thus, the user has to identify a set of homologs that support a sufficiently deep but also highly robust tree.

Often, users initially create for ASR a large set of homologs by means of a `BLAST` search, which needs further processing. To reduce the number of sequences and redundancy, one can start by picking representatives from the sequence clusters created with the help of tools like `cd-hit` (Li and Godzik, 2006) or use more specialized approaches (Frickey and Lupas, 2004; Fuellen et al., 2005; Dereeper et al., 2008; Tamura et al., 2011). In order to combine several orthogonal methods for sequence selection, we designed `FitSS4ASR`, which **fil**t**ers s**equence **s**ets **for ASR**. This tool draws upon well-proven concepts applied to iteratively refine sets and can be used in a semi-automatic manner with minimal user interaction. Our tool selects sequences that i) agree best with current evolutionary models in order to avoid systematic errors and ii) have retained a strong phylogenetic signal in order to avoid stochastic errors.

We confirmed the validity of this kind of sequence selection by means of a biochemical characterization of ancestral geranylgeranylglyceryl phosphate synthases (GGGPSs). This enzyme is involved in the biosynthesis of ether membrane lipids that are prototypical for

Archaea and catalyzes the formation of an ether bond between glycerol 1-phosphate and geranylgeranyl diphosphate (Chen et al., 1993; Peterhoff et al., 2014). A characteristic, taxon-specific property of GGGPS is the oligomerization state, which can be dimeric or hexameric (Peterhoff et al., 2014; Linde et al., 2018). We used the same ASR protocol, but two different sets of recent GGGPS homologs to compute ancestors. The first set of sequences was compiled in a 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 ancestral sequences were computed; the corresponding proteins were heterologously expressed and characterized. The accordance of the experimentally determined oligomerization states confirmed the equivalence of the resurrected enzymes with respect to this delicate property and testified to the applicability of `FitSS4ASR`.

Results

General criteria guiding sequence selection for ASR

Commonly, the first step of sequence selection is the generation of a set by means of `BLAST` (Altschul et al., 1997) or the choice of a precompiled dataset as offered by InterPro (Mitchell et al., 2015) 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, the distribution of species in the comprehensive tree of life (Hug et al., 2016) suggests for the reconstruction of LUCA proteins a sequence set representing species from the six dominating bacterial and two archaeal clades. Usually, it is easy to provide a broad phylogenetic representation for a given protein due to the wide coverage of extant sequences deposited in databases. Thus, the crucial task of sequence selection is a rigorous but specific filtering, and the appropriate combination of filters might advantageously be exploited to increase the robustness of the ASR process. As a first step, a single representative can be chosen for each subset of highly similar sequences to reduce redundancy. In order to eliminate splice variants and flawed sequences caused by misassembly or gene-prediction errors, non-canonical outliers whose length differs significantly, i. e., by more than *len_dev* = 2 standard deviations from the mean can be eliminated as well (Figure 1A). Note that we use the term "*param* = value" to specify parameters of FitSS4ASR and their default values that can be altered by the user. Sequences without indels have more likely retained their ancestral length (Akiva et al., 2017) and the evolutionary correct modeling of indels is still difficult. Thus, it is appropriate to ignore also sequences with internal insertions (Dereeper et al., 2008) as indicated by a multiple sequence alignment (MSA) (Figure 1B).

Other filter criteria (Merkl and Sterner, 2016) are only applicable after a phylogenetic tree has been computed for the input. Horizontal gene transfer (HGT) (Figure 1C) is a frequent 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, which requires to compare the taxonomy of nodes and subtrees. Moreover, to support a reliable reconstruction of subsequent ancestral states, the length of each branch has to indicate a sufficiently low rate *subs_r* = 1.0 of substitutions per site (Figure 1D). The reconstruction of subsequent states can be unreliable, if more than one substitution occurred during the time span represented by the length of a given branch. FitSS4ASR considers substitution rates exceeding *subs_r* as critical. Note that low substitution rates reduce also the risk of long-branch attraction, which is a systematic error that may occur if a tree contains long and short branches. As a consequence, two or more long branches can be grouped as sisters (nodes that share the same parent node) and distantly related species seem to be closely related (Bergsten, 2005). A further criterion for the robustness of subtree topology are the local bootstrap values/posterior probabilities that are compared to *loc_qual* = 0.75. In agreement with the literature (Soltis and Soltis, 2003), FitSS4ASR considers *loc_qual* values ≥ 0.75 as an indicator of sufficient support to reconstruct the corresponding sequences.

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, a manual sequence selection may turn into a tedious and time-consuming task.

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 2A). 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 initial 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.

During the sequence elimination phase of `FitSS4ASR`, 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., 2013) detects “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. Solitary are the sequences seq_k^r of a given set SEQ_k, for which each of the sisters *s* occurs in a fraction below *min_sis* = 0.75 of the *k*-specific trees tr_k^i; see Methods. `FitSS4ASR` allows the user to choose one of three alternatives rules that eliminate i) rogue, ii) solitary, iii) rogue and solitary sequences that may cause the computation of incorrect ancestors. The removal of these sequences can create branches of undesired length. Thus, all other sequences inducing branches longer than *subs_r* substitution/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` stops, if SEQ_k contains not more than *min_seq* = 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 $u = \left|SEQ_{last}\right|$ sequences.

Upon completion of sequence elimination, `FitSS4ASR` assesses the robustness of the generated datasets to offer alternatives from which the user can choose (Figure 2B). Initially, up to 15 datasets $Iter_k^+$ are chosen from the last rounds of sequence selection so that the sets

$Iter_k^+$ contain approximately evenly distributed between u and maximally 500 sequences. 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 filter out such sets, `FitSS4ASR` deduces a supertree and discards trees tr_k^+ that are not compatible with the core topology; see Methods. A supertree is a single phylogenetic tree resulting from a combination of trees; here it is expected that the trees tr_k^+ overlap largely. 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 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 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 a further characterization of the sets is needed to facilitate a selection.

Choosing a dataset for ASR

After program termination, the user has to select one of 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*(*Alts*) (Formula 1), the user can survey the phylogenetic coverage of the sequence set SEQ_s. Two scores assess the quality of tr_s: *branch_distr*(*Alts*) (Formula 2) is a measure for the existence of exceedingly long branches longer than *subs_r* substitution per site and *pp_distr*(*Alts*) (Formula 3) indicates the "reliability" of branches near the root of the tree. *tr_rob*(*Alts*) (Formula 4), summarizes the phylogenetic robustness of tr_s with respect to perturbations and *tr_mf*(*Alts*) (Formula 5) penalizes the existence of multifurcations. We consider a dataset SEQ_s a good choice for ASR, if 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 (Formula 6).

Validation of FitSS4ASR by means of an ASR of GGGPS

In order to confirm the efficacy of our approach, we performed in parallel a conventional and a FitSS4ASR-assisted ASR of the enzyme GGGPS. GGGPS is a key enzyme in the evolution of Archaea (Payandeh and Pai, 2007), but also occurs in bacterial species, albeit with unknown physiological function. In a previous analysis, all enzymes have been assigned to one of two groups (Peterhoff et al., 2014): Group I enzymes occur in Euryarchaeota and Firmicutes, however, most Archaea possess a group II GGGPS, which also occurs in some bacterial species like Bacteroidetes. The reconstruction of a common ancestor of both groups is not feasible due to the length of the edge (> 4 substitutions per site) that interconnects the nodes representing the ancestors of group I (AncGGGPS1) and group II (AncGGGPS2) enzymes. Whereas all group I enzymes oligomerize to dimers, group II enzymes form dimers or hexamers that can be clustered based on sequence similarity (Peterhoff et al., 2014). Thus, we expected differing oligomerization states for predecessors of group II enzymes and a dimeric last common ancestor due to the principle of parsimony. The deliberately chosen single point mutation W141A in GGGPS2 from the archaeon *Methanothermobacter thermautotrophicus* turns the hexameric wild-type into a dimeric complex (Peterhoff et al., 2014). This finding indicates that minor sequence alterations can affect the oligomerization state of GGGPS2, which makes its analysis to an ideal testbed for the performance and robustness of ASR protocols.

Conventional sequence selection for ASR of GGGPS2

We started sequence selection with the analysis of a comprehensive and precompiled set *$GGGPS2_{initial}$*, which consisted of 217 entries from InterPro family IPR008205 (version 67.0). To generate this set, the above-mentioned sequence length 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 an easy to understand string representing the phylogenetic lineage of the contributing species, which is deduced from the NCBI taxonomy database (Ashkenazy et al., 2009). We used these annotations to determine the phylogenetic diversity of this and all other sequence sets.

An MSA was computed by means of MAFFT (Katoh and Standley, 2013) and exclusively for the computation of a phylogenetic tree, 75 of the 287 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 criteria, 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* consisted of 87 sequences and 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 0.00024, which indicates high MCMC convergence. The MSA consisting of the full-length sequences that were used to deduce ancestral sequences is listed in Table S1 (Supporting Information), the resulting phylogenetic tree is given in Table S3 (Supporting Information) and shown in Figure 3.

Sequence selection for an ASR of GGGPS2 by means of FitSS4ASR

The sequences *GGGPS2*$_{initial}$ used above for a conventional sequence selection were also subjected to FitSS4ASR. The program converged after two rounds of iteration and the *ASR_score*(*Alts*) of the alternatives suggested to consider a specific set $Alt_s = \{tr_s, SEQ_s\}$. However, it 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 *GGGPS2*$_{initial}$ was added to SEQ_s and a further FitSS4ASR run was performed, which resulted in a final set *GGGPS2_auto* consisting of 61 sequences. The MSA consisting of the full-length sequences used to deduce ancestral sequences is listed in Table S2 (Supporting Information), the resulting phylogenetic tree is given in Table S4 (Supporting Information) and shown in Figure 4. This final dataset had an *ASR_score* (Formula 6) of 0.17 due to two multifurcations within Thermococcales and Methanosarcinales (compare Figure 4); *tr_mf* (Formula 5) was 0.33. An assessment of the other scores confirmed that it fulfills all criteria for ASR: *branch_distr* (Formula 2) was 0.99, *pp_distr* (Formula 3) was 0.85, and *tr_rob* (Formula 4) was 0.61. Moreover, a comparison of Figures 3 and 4 made clear that both trees possess a highly similar topology. 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.

Reconstruction and characterization of AncGGGPS2 predecessors

Extant GGGPS group II enzymes form dimers or hexamers and it is unknown when these oligomerization states arose. All euryarchaeal and thaumarchaeal proteins shown in Figures 3 and 4 form hexamers and for these, we wanted to elucidate oligomerization for the evolutionary interval dating back to the last common ancestor AncGGGPS2. Thus, sequences representing this LCA sequence and intermediates AncGGGPS2_N* were reconstructed by means of FastML (Ashkenazy et al., 2012). 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 predecessors by comparing the topology of the trees; see Figures 3 and 4.

We first analyzed the oligomerization states of the last common ancestors AncGGGPS2_N1_man and AncGGGPS2_N1_auto. Using synthetic genes, proteins were heterologously expressed in *Escherichia coli* and purified by means of metal chelate affinity chromatography. Their oligomerization states were determined by using analytical size exclusion chromatography (SEC) and well-characterized hexameric (mtGGGPS_wt) or dimeric variants (mtGGGPS_W141A) (Peterhoff et al., 2014; Linde et al., 2018) served as references.

The finding that both common ancestors form dimers (Figure 5A) suggests that hexamerization arose at later evolutionary phases. For euryarchaeal group II GGGPS, this transition must occur at one of the internal nodes of the trees used for ASR. Almost all recent euryarchaeal and thaumarchaeal GGGPS are hexamers (Peterhoff et al., 2014; Linde et al., 2018), thus we assumed that this transition occurs at the furcation AncGGGPS2_N4_man → (AncGGGPS2_N5_man, AncGGGPS2_N12_man); see Figure 3. The corresponding furcation in the FitSS4ASR-based tree is AncGGGPS2_N17_auto → (AncGGGPS2_N18_auto, AncGGGPS2_N22_auto); see Figure 4.

Synthetic genes encoding those six proteins were expressed heterologously in *E. coli*, and the purified proteins were analyzed by SEC; the chromatograms are shown in Figure 5B – D. While both AncGGGPS2_N4_man and AncGGGPS2_N17_auto eluted as dimers (Figure 5B), all four successor-proteins eluted as hexamers (Figure 5C, D). This result confirms that we have successfully predicted the evolutionary phase related to the alteration of the oligomerization state. More importantly, the concordant transition of oligomerization states strongly supports the validity of the FitSS4ASR approach, although the input and the output of the two ASR protocols varied: The sequence sets *GGGPS2_man* and *GGGPS2_auto* overlapped by not more than 34 sequences. The ancestor AncGGGPS2_N1_man consists of 246 amino acids and a BLAST search made clear that it shares 65% SeqId with the most similar extant GGGPS2 sequence from the Thermoproteales *Thermoproteus uzoniensis* abbreviated as ACrThThTh_Tuz in Figure 3. The ancestor AncGGGPS2_N1_auto possesses 242 amino acids and a BLAST search indicated that it shares 72% SeqId with the most similar extant sequence from the Desulfurococcaceae *Staphylothermus hellenicus* abbreviated as ACrThDeDe_She in Figure 4. AncGGGPS2_N1_man and AncGGGPS2_N1_auto share 75% SeqId and 36 of the 45 differences are similar residues that possess a positive pairwise BLOSUM (Henikoff and Henikoff, 1992) score. Similarity increases for less ancestral predecessors: The corresponding intermediates AncGGGPS2_N5_man and AncGGGPS2_N18_auto share 93%, and AncGGGPS2_N12_man and AncGGGPS2_N22_auto 91% SeqId. All sequences are listed in Table S1 – Table S5 (Supporting Information). Despite these differences, the phenotypes, i. e., the oligomerization state of the ancestors, as well as the location of the transition from a dimer to a hexamer within the trees derived from the *GGGPS2_man* and *GGGPS2_auto* sequence sets are identical. Thus, our experiments provide a further example for the robustness of ASR against uncertainty (Hanson-Smith et al., 2010) and point to a strategy to strengthen the support for conclusions of ASR experiments: In addition to ensemble methods used to sample and characterize less likely ancestors (Bar-Rogovsky et al., 2015), the support for a specific phenotype can also be increased by a co-validation of experimental findings based on two or more distinct sequence samplings.

Selecting sequences for ASR of another two enzymes

We reconstructed ancestral GGGPS2 enzymes because we wanted to retrace the advent of their different oligomerization states. In this case, sequence selection was initiated with a precompiled set of *bona fide* group II enzymes, which was comprehensive (Peterhoff et al., 2014) but contained only 217 sequences. In order to demonstrate the usefulness of `FitSS4ASR` for the filtering of larger sequence sets, we applied it to a dataset of TrpD2 sequences (16820 sequences) and a dataset of concatenated HisH/HisF sequences (1309 sequences, (Richter et al., 2010)), which are or have been in the focus of other ASR experiments. In both cases, `FitSS4ASR` was run with default parameters. For each test set, we could create with minimal additional effort at least one tree that strongly supports the reconstruction of early predecessors. These trees are plotted in Figures S2, S3 and detailed in Tables S6, S7 (Supporting Information).

Discussion

ASR requires a strong phylogenetic signal necessitating a rigorous preselection of sequences

Often, the analysis of large datasets is regarded 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 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 has been 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 the trees required for the intended ASR. The inspection of suboptimal trees provides insight into the interplay among conflicting noise versus phylogenetic signal (Swofford et al., 1996) and to reduce noise, 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).

A limitation of `FitSS4ASR` is its blindness against the phylogenetic diversity of the chosen sequences. As demonstrated for the GGGPS2 and the HisF/HisH reconstruction, it might be necessary to add sequences manually to broaden the phylogenetic basis. Additionally, some HGT events have to be resolved manually by comparing `key2ann` annotations. Moreover, `FitSS4ASR` is less suitable for proteins, which encountered frequently length variations causing many indels or possess high evolutionary rates. For these cases, the filter routines might eliminate too many sequences of the input.

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 assessing their robustness (Lemoine et al., 2018).

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 (Monit and Goldstein, 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 can help to the speed up any ASR protocol.

Materials and methods

Software and parameters used for the 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 site-heterogeneous CAT model and by launching 4 independent MCMC samplings of length 50,000 to ensure convergence. For the final dataset, the first 6000 trees of each chain were discarded as burn-in and the remaining trees of two chains were concatenated to deduce a consensus tree.

Software and parameters used for **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 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 2 standard deviations from the mean were also 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 *min_len* = 5 in at least 90% of the sequences were removed; the resulting sequence set $SEQ_{k=1}$ contained $m = \left|SEQ_k\right|$ sequences. For subsequent analyses aimed at the computation of robust trees, the sequences were realigned and columns containing more than 50% gap symbols were eliminated by utilizing Gblocks (Castresana, 2000). Two phylogenetic analyses were performed by means of RAxML or MrBayes. The RAxML option -f a and the substitution model

PROTGAMMAAUTO were utilized 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 2 chains. During each iteration, the dataset $Iter_k = \{tr_k, SEQ_k\}$ consisting of the consensus tree and the sequences was saved 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 ≤ *min_seq* = 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., 2013) 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 $sis_k[r,s] < min\_sis = 0.75 \ \forall s$. 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 dataset SEQ_1. These 100 trees were pruned to the dataset 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).

Indicators of ASR suitability

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

$$tax\_num(Alt_s) = \#\text{ species to be found in } SEQ_s \quad (1)$$

tax_num(*Alts*) is the number of species contributing to the respective sequence set.

$$branch\_distr(Alt_s) = \#\text{ branches shorter than } subs\_r \text{ substitution per site / \# all branches} \quad (2)$$

The value of *branch_distr*(*Alts*) is 1.0, if all branches are shorter than *subs_r* substitution per site and decreases with the number of exceedingly long branches.

$$pp\_distr(Alt_s) = \sum_{v, pp(v) \geq loc\_qual} pp(v) \cdot dist(root, v) \Big/ \sum_{v} pp(v) \cdot dist(root, v) \quad (3)$$

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*(*Alts*) reaches 1.0, if all branches near the root are well supported by posterior probabilities of at least *loc_qual*.

$$tr\_rob(Alt_s) = \text{the fraction of shared bipartitions} \quad (4)$$

The value of *tr_rob*(*Alts*) 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.

$$tr\_mf(Alt_s) = 1/(1 + \text{the number of multifurcations}) \quad (5)$$

The value of *tr_mf*(*Alts*) is 1.0, if tr_s does not contain multifurcations, which impede ASR and are introduced by some programs during the computation of consensus trees.

$$ASR\_score(Alt_s) = branch\_distr(Alt_s) \cdot pp\_distr(Alt_s) \cdot tr\_rob(Alt_s) \cdot tr\_mf(Alt_s) \quad (6)$$

ASR_score(*Alts*) is close to 1.0, if a tree fulfils all stability parameters.

Ancestral sequence reconstruction

NJplot (Gouy, 1995) 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.

Cloning

Gene sequences for the reconstructed proteins (AncGGGPS2_N1_man, Anc-GGGPS2_N4_man, AncGGGPS2_N5_man, AncGGGPS2_N12_man, Anc-GGGPS2_N1_auto, AncGGGPS2_N17_auto, AncGGGPS2_N18_auto and Anc-GGGPS2_N22_auto) were optimized in their codon usage for expression in *E. coli* and ordered as GeneArt™ Strings™ DNA fragments from Thermo Fisher Scientific. Their nucleotide sequences are given in Table S5 (Supporting Information). The DNA fragments had *Bsa*I restriction sites at the 5' and 3'-end and were cloned into a modified pET21a expression vector (Rohweder et al., 2018), providing a C-terminal hexahistidine $(His)_6$ tag. The constructs were verified by sequencing. The wild-type *Methanothermobacter thermautotrophicus* GGGPS (mtGGGPS_wt) and the mtGGGPS_W141A variant have been cloned previously into pET21a (Peterhoff et al., 2014). The sequence numbering of mtGGGPS used in this study refers to EMBL ENA entry AAB85058 and PDB-ID 4mm1, which have an N-terminal three amino acid extension compared to UniProt entry O26652.

Production and purification of recombinant proteins

Heterologous gene expression was performed in *E. coli* strain BL21-CodonPlus(DE3)-RIPL (Agilent Technologies). The transformed cells were grown at 37 °C in LB medium containing ampicillin (150 µg ml^{-1}) and chloramphenicol (30 µg ml^{-1}). When OD_{600} reached 0.6 - 0.8, expression was induced by adding 1 mM isopropyl-β-D-1-thiogalactopyranoside (IPTG), and growth was continued overnight. After harvesting by centrifugation, cells were resuspended in 50 mM potassium phosphate, pH 7.5, 300 mM KCl, 10 mM imidazole and disrupted by sonication. The His_6-tagged proteins were purified from the clarified cell extract by metal chelate affinity chromatography. An ÄKTApurifier system with a HisTrap FF crude column (5 ml, GE Healthcare) was used, and a linear gradient of imidazole (10 - 500 mM) in 50 mM potassium phosphate, pH 7.5, 300 mM KCl was applied to elute the protein. Interfering imidazole and salt were removed from the purified proteins, except for AncGGGPS2_N17_auto, by dialysis against 50 mM potassium phosphate, pH 7.5 at 4 °C. Because the AncGGGPS2_N17_auto preparation contained higher-oligomeric soluble aggregates, it was further purified by subsequent preparative SEC on a Highload^{TM} 26/600

Superdex^{TM} S75 pg column (GE Healthcare). The column was equilibrated with 50 mM potassium phosphate, pH 7.5 and was run at a flow rate of 0.5 ml min^{-1}. Protein concentrations were determined by absorbance spectroscopy. The molar extinction coefficients ε_{280} and the molecular weight were calculated from the amino acid sequence by means of `ProtParam` (Gasteiger et al., 2005). The purity of all proteins was verified by SDS-PAGE; see Figure S1 (Supporting Information). The proteins were dropped into liquid nitrogen and stored at -80 °C until further analysis.

Characterization of oligomerization states of GGGPS variants

All GGGPS variants were characterized by analytical SEC experiments. Analysis was performed on a calibrated Superdex^{TM} S75 10/300 GL column (GE Healthcare), which was operated with 50 mM potassium phosphate, pH 7.5, 300 mM KCl at a flow rate of 0.5 ml min^{-1}. 100 µl of protein with a subunit concentration of 40 µM was applied. mtGGGPS_wt and mtGGGPS_W141 served as references for the hexameric and dimeric oligomerization state, respectively (Peterhoff et al., 2014).

Acknowledgement

This work was supported by the Deutsche Forschungsgemeinschaft (ME2259/2-1) and calculations were facilitated through the use of advanced computational infrastructure provided for account pr48fu by the Leibniz Supercomputing Center of the Bavarian Academy of Sciences and Humanities (LRZ).

References

Aberer A.J., Krompass D., and Stamatakis A. (2013). Pruning rogue taxa improves phylogenetic accuracy: an efficient algorithm and webservice. Syst. Biol. *62*, 162-166.

Akanuma S., Nakajima Y., Yokobori S., Kimura M., Nemoto N., Mase T., Miyazono K., Tanokura M., and Yamagishi A. (2013). Experimental evidence for the thermophilicity of ancestral life. Proc. Natl. Acad. Sci. U S A *110*, 11067-11072.

Akiva E., Copp J.N., Tokuriki N., and Babbitt P.C. (2017). Evolutionary and molecular foundations of multiple contemporary functions of the nitroreductase superfamily. Proc. Natl. Acad. Sci. U S A *114*, E9549-E9558.

Alcolombri U., Elias M., and Tawfik D.S. (2011). Directed evolution of sulfotransferases and paraoxonases by ancestral libraries. J. Mol. Biol. *411*, 837-853.

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

Ashkenazy H., Penn O., Doron-Faigenboim A., Cohen O., Cannarozzi G., Zomer O., and Pupko T. (2012). FastML: a web server for probabilistic reconstruction of ancestral sequences. Nucleic Acids Res. *40*, W580-584.

Ashkenazy H., Unger R., and Kliger Y. (2009). Optimal data collection for correlated mutation analysis. Proteins *74*, 545-555.

Bar-Rogovsky H., Stern A., Penn O., Kobl I., Pupko T., and Tawfik D.S. (2015). Assessing the prediction fidelity of ancestral reconstruction by a library approach. Protein. Eng. Des. Sel. *28*, 507-518.

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

Boussau B., Blanquart S., Necsulea A., Lartillot N., and Gouy M. (2008). Parallel adaptations to high temperatures in the Archaean eon. Nature *456*, 942-945.

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

Castresana J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol. Biol. Evol. *17*, 540-552.

Chen A., Zhang D., and Poulter C.D. (1993). (S)-geranylgeranylglyceryl phosphate synthase. Purification and characterization of the first pathway-specific enzyme in archaebacterial membrane lipid biosynthesis. J. Biol. Chem. *268*, 21701-21705.

Chen F., Gaucher E.A., Leal N.A., Hutter D., Havemann S.A., Govindarajan S., Ortlund E.A., and Benner S.A. (2010). Reconstructed evolutionary adaptive paths give polymerases accepting reversible terminators for sequencing and SNP detection. Proc. Natl. Acad. Sci. U S A *107*, 1948-1953.

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

Dereeper A., Guignon V., Blanc G., Audic S., Buffet S., Chevenet F., Dufayard J.F., Guindon S., Lefort V., Lescot M., Claverie J.M., and Gascuel O. (2008). Phylogeny.fr: robust phylogenetic analysis for the non-specialist. Nucleic Acids Res. *36*, W465-469.

Essoussi N., Boujenfa K., and Limam M. (2008). A comparison of MSA tools. Bioinformation *2*, 452-455.

Field S.F., and Matz M.V. (2010). Retracing evolution of red fluorescence in GFP-like proteins from Faviina corals. Mol. Biol. Evol. *27*, 225-233.

Frickey T., and Lupas A.N. (2004). PhyloGenie: automated phylome generation and analysis. Nucleic Acids Res. *32*, 5231-5238.

Fuellen G., Spitzer M., Cullen P., and Lorkowski S. (2005). Correspondence of function and phylogeny of ABC proteins based on an automated analysis of 20 model protein data sets. Proteins *61*, 888-899.

Gasteiger E., Hoogland C., Gattiker A., Duvaud S., Wilkins M.R., Appel R.D., and Bairoch A. (2005). Protein identification and analysis tools on the ExPASy Server. In: The proteomics protocols handbook, Walker JM, editor (Totowa, N. J.: Humana Press), pp. 571-607.

Gouy M. (1995). NJplot. University of Lyon. URL: http://pbil.univ-lyon1.fr/software/njplot.html.

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

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

Hanson-Smith V., Kolaczkowski B., and Thornton J.W. (2010). Robustness of ancestral sequence reconstruction to phylogenetic uncertainty. Mol. Biol. Evol. *27*, 1988-1999.

Harms M.J., and Thornton J.W. (2010). Analyzing protein structure and function using ancestral gene reconstruction. Curr. Opin. Struct. Biol. *20*, 360-366.

Henikoff S., and Henikoff J.G. (1992). Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. U S A *89*, 10915-10919.

Ho S.Y., and Jermiin L. (2004). Tracing the decay of the historical signal in biological sequence data. Syst. Biol. *53*, 623-637.

Hobbs J.K., Shepherd C., Saul D.J., Demetras N.J., Haaning S., Monk C.R., Daniel R.M., and Arcus V.L. (2012). On the origin and evolution of thermophily: reconstruction of functional precambrian enzymes from ancestors of *Bacillus*. Mol. Biol. Evol. *29*, 825-835.

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

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

Hug L.A., Baker B.J., Anantharaman K., Brown C.T., Probst A.J., Castelle C.J., Butterfield C.N., Hernsdorf A.W., Amano Y., Ise K., Suzuki Y., Dudek N., Relman D.A., Finstad K.M., Amundson R., Thomas B.C., and Banfield J.F. (2016). A new view of the tree of life. Nat Microbiol *1*, 16048.

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

Katoh K., and Standley D.M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. *30*, 772-780.

Kumar S., Stecher G., Peterson D., and Tamura K. (2012). MEGA-CC: computing core of molecular evolutionary genetics analysis program for automated and iterative data analysis. Bioinformatics *28*, 2685-2686.

Kupczok A. (2011). Split-based computation of majority-rule supertrees. BMC Evol. Biol. *11*, 205.

Lartillot N., Lepage T., and Blanquart S. (2009). PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics *25*, 2286-2288.

Lartillot N., and Philippe H. (2004). A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol. Biol. Evol. *21*, 1095-1109.

Le Q., Sievers F., and Higgins D.G. (2017). Protein multiple sequence alignment benchmarking through secondary structure prediction. Bioinformatics *33*, 1331-1337.

Lemoine F., Domelevo Entfellner J.B., Wilkinson E., Correia D., Davila Felipe M., De Oliveira T., and Gascuel O. (2018). Renewing Felsenstein's phylogenetic bootstrap in the era of big data. Nature *556*, 452-456.

Li G., Steel M., and Zhang L. (2008). More taxa are not necessarily better for the reconstruction of ancestral character states. Syst. Biol. *57*, 647-653.

Li W., and Godzik A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics *22*, 1658-1659.

Liberles D.A. 2007. Ancestral sequence reconstruction. Oxford: Oxford University Press.

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

Litsios G., and Salamin N. (2012). Effects of phylogenetic signal on ancestral state reconstruction. Syst. Biol. *61*, 533-538.

Löytynoja A., and Goldman N. (2008). Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. Science *320*, 1632-1635.

Merkl R., and Sterner R. (2016). Ancestral protein reconstruction: techniques and applications. Biol. Chem. *397*, 1-21.

Mitchell A., Chang H.Y., Daugherty L., Fraser M., Hunter S., Lopez R., McAnulla C., McMenamin C., Nuka G., Pesseat S., Sangrador-Vegas A., Scheremetjew M., Rato C., Yong S.Y., Bateman A., Punta M., Attwood T.K., Sigrist C.J., Redaschi N., Rivoire C., Xenarios I., Kahn D., Guyot D., Bork P., Letunic I., Gough J., Oates M., Haft D., Huang H., Natale D.A., Wu C.H., Orengo C., Sillitoe I., Mi H., Thomas P.D., and Finn R.D. (2015). The InterPro protein families database: the classification resource after 15 years. Nucleic Acids Res. *43*, D213-221.

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

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

Ochman H., Lawrence J.G., and Groisman E.A. (2000). Lateral gene transfer and the nature of bacterial innovation. Nature *405*, 299-304.

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

Pagel M., Meade A., and Barker D. (2004). Bayesian estimation of ancestral character states on phylogenies. Syst. Biol. *53*, 673-684.

Payandeh J., and Pai E.F. (2007). Enzyme-driven speciation: crystallizing Archaea via lipid capture. J. Mol. Evol. *64*, 364-374.

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

Peterhoff D., Beer B., Rajendran C., Kumpula E.P., Kapetaniou E., Guldan H., Wierenga R.K., Sterner R., and Babinger P. (2014). A comprehensive analysis of the geranylgeranylglyceryl phosphate synthase enzyme family identifies novel members and reveals mechanisms of substrate specificity and quaternary structure organization. Mol. Microbiol. *92*, 885-899.

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

Reisinger B., Sperl J., Holinski A., Schmid V., Rajendran C., Carstensen L., Schlee S., Blanquart S., Merkl R., and Sterner R. (2014). Evidence for the existence of elaborate enzyme complexes in the Paleoarchean era. J. Am. Chem. Soc. *136*, 122-129.

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

Rivera-Rivera C.J., and Montoya-Burgos J.I. (2016). LS^3: a method for improving phylogenomic inferences when evolutionary rates are heterogeneous among taxa. Mol. Biol. Evol. *33*, 1625-1634.

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

Rohweder B., Semmelmann F., Endres C., and Sterner R. (2018). Standardized cloning vectors for protein production and generation of large gene libraries in *Escherichia coli*. BioTechniques *64*, 24-26.

Ronquist F., and Huelsenbeck J.P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics *19*, 1572-1574.

Salichos L., and Rokas A. (2013). Inferring ancient divergences requires genes with strong phylogenetic signals. Nature *497*, 327-331.

Sanderson M.J., and Shaffer H.B. (2002). Troubleshooting molecular phylogenetic analyses. Annu. Rev. Ecol. Syst. *33*, 49-72.

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

Stamatakis A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics *22*, 2688-2690.

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

Straub K., and Merkl R. (2019). Ancestral sequence reconstruction as a tool for the elucidation of a stepwise evolutionary adaptation. In: Computational methods in protein evolution, Sikosek T, editor (New York City: Humana Press), pp. 171-182.

Swofford D.L., Olsen G.J., Waddell P.J., and Hillis D.M. (1996). Phylogenetic inference. In: Molecular Systematics, Hillis DM, Moritz C, Mable BK, eds. (Sunderland, MA: Sinauer and Associates), pp. 407-514.

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

Tamura K., Peterson D., Peterson N., Stecher G., Nei M., and Kumar S. (2011). MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol. Biol. Evol. *28*, 2731-2739.

Thornton J.W., Need E., and Crews D. (2003). Resurrecting the ancestral steroid receptor: ancient origin of estrogen signaling. Science *301*, 1714-1717.

Tokuriki N., Stricher F., Serrano L., and Tawfik D.S. (2008). How protein stability and new functions trade off. PLoS Comp. Biol. *4*, e1000002.

Vialle R.A., Tamuri A.U., and Goldman N. (2018). Alignment modulates ancestral sequence reconstruction accuracy. Mol. Biol. Evol. *37*, 1783-1797.

Waterhouse A.M., Procter J.B., Martin D.M., Clamp M., and Barton G.J. (2009). Jalview Version 2–a multiple sequence alignment editor and analysis workbench. Bioinformatics *25*, 1189-1191.

Wheeler L.C., Lim S.A., Marqusee S., and Harms M.J. (2016). The thermostability and specificity of ancient proteins. Curr. Opin. Struct. Biol. *38*, 37-43.

Wiens J.J. (2005). Can incomplete taxa rescue phylogenetic analyses from long-branch attraction? Syst. Biol. *54*, 731-742.

Wijma H.J., Floor R.J., and Janssen D.B. (2013). Structure- and sequence-analysis inspired engineering of proteins for enhanced thermostability. Curr. Opin. Struct. Biol. *23*, 588-594.

Wilkinson M., and Crotti M. (2017). Comments on detecting rogue taxa using RogueNaRok. Syst. Biodivers. *15*, 291-295.

Wouters M.A., Liu K., Riek P., and Husain A. (2003). A despecialization step underlying evolution of a family of serine proteases. Mol. Cell *12*, 343-354.

Figures

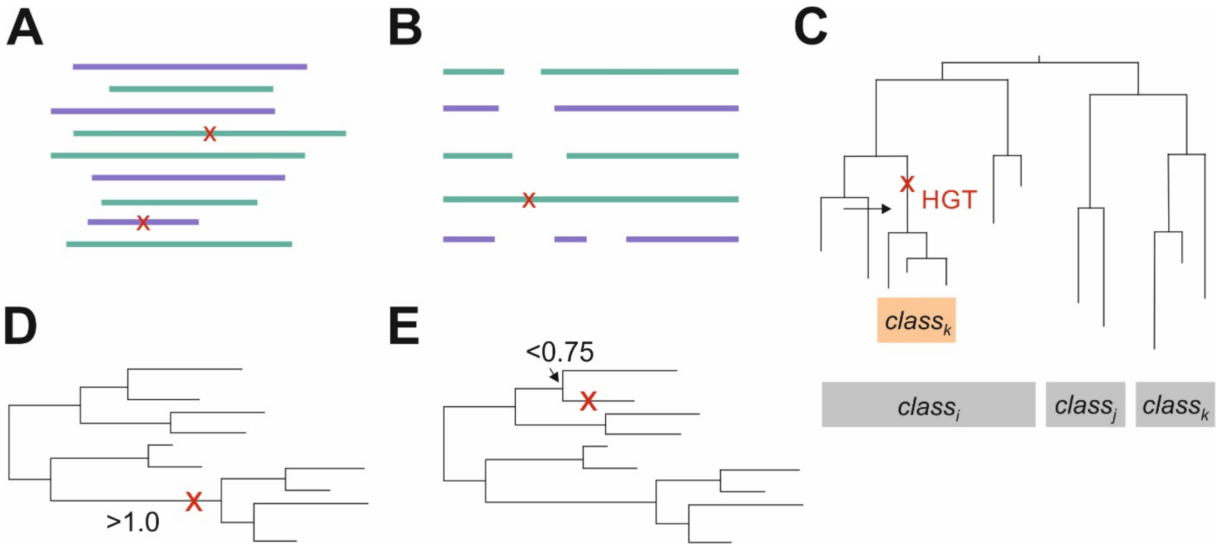

Figure 1. Criteria applied in ASR to eliminate sequences.

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 (*class$_i$* → *class$_k$*) by means of horizontal gene transfer (HGT) as exemplified for species from three distinct phylogenetic classes *class$_i$*, *class$_j$*, and *class$_k$*. (D) Sequences inducing subtrees with long branches. (E) Sequences causing a weakly supported subtree topology.

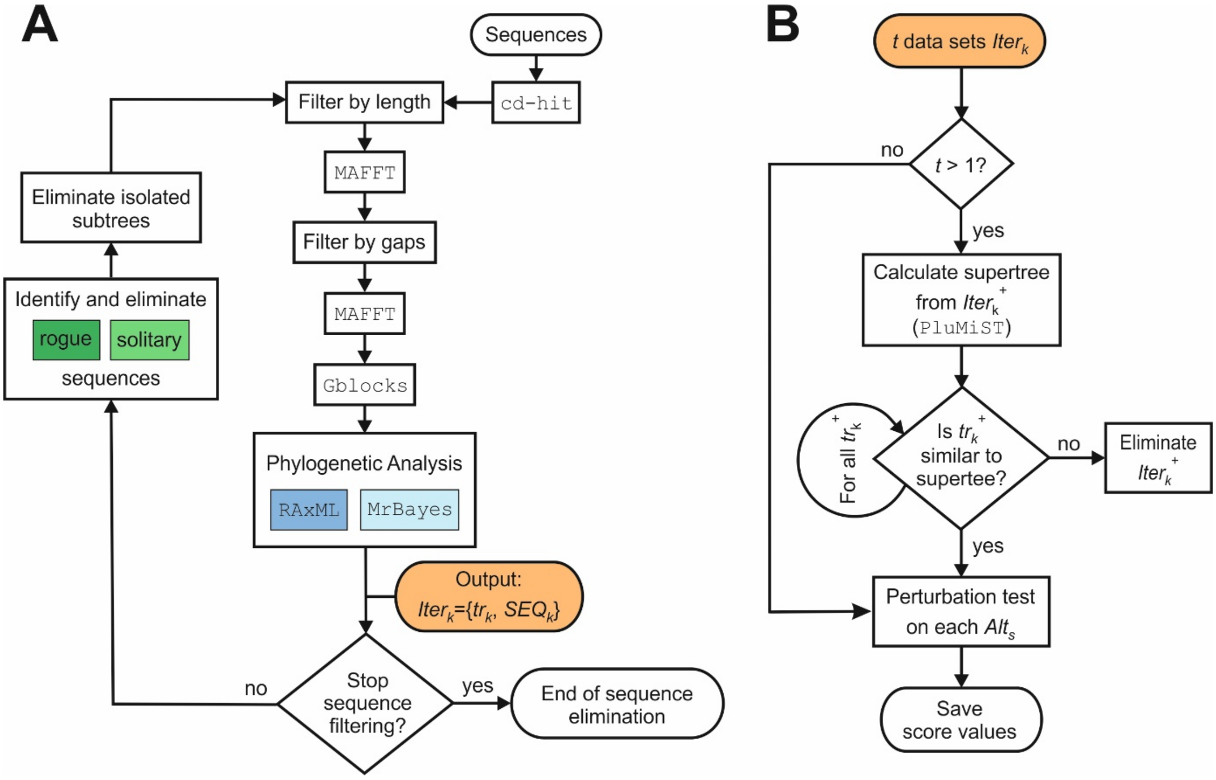

Figure 2. Workflow of FitSS4ASR.

(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 *k*, FitSS4ASR performs a phylogenetic analysis by means of RAxML or MrBayes based on an MSA created by means of MAFFT and Gblocks. The iteration-specific 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 substitution/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_s related to datasets *Alts* 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.

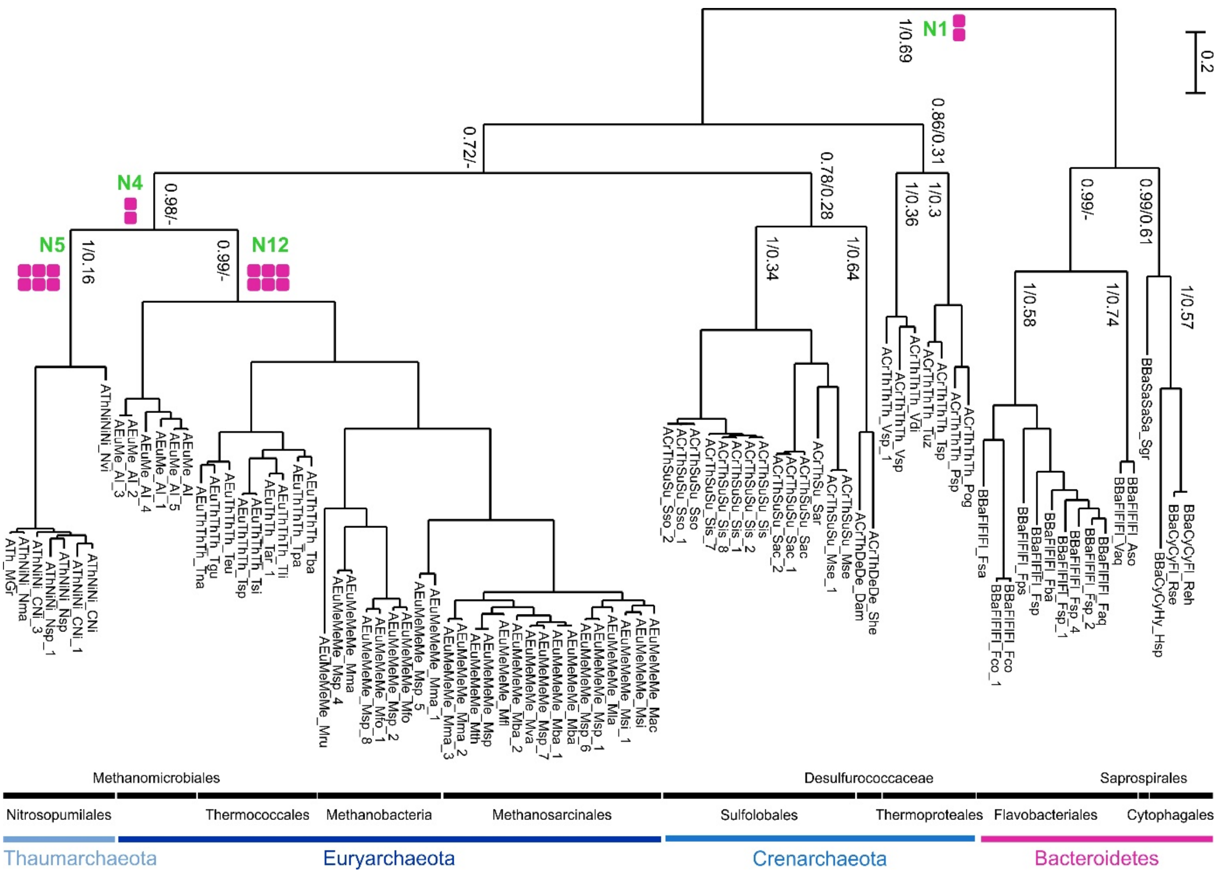

Figure 3. Phylogeny of the manually curated sequence set used for ASR of GGGPS2 predecessors.

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 `MrBayes`. For central nodes, the posterior probability and (if available) a bipartition score is given. The length of the horizontal bar corresponds to 0.2 substitutions per site. Nodes "N*_man" labelled in green correspond to evolutionary intermediates that were characterized biochemically. Their dimeric or hexameric oligomerization states, as determined by size exclusion chromatography, are indicated by red symbols.

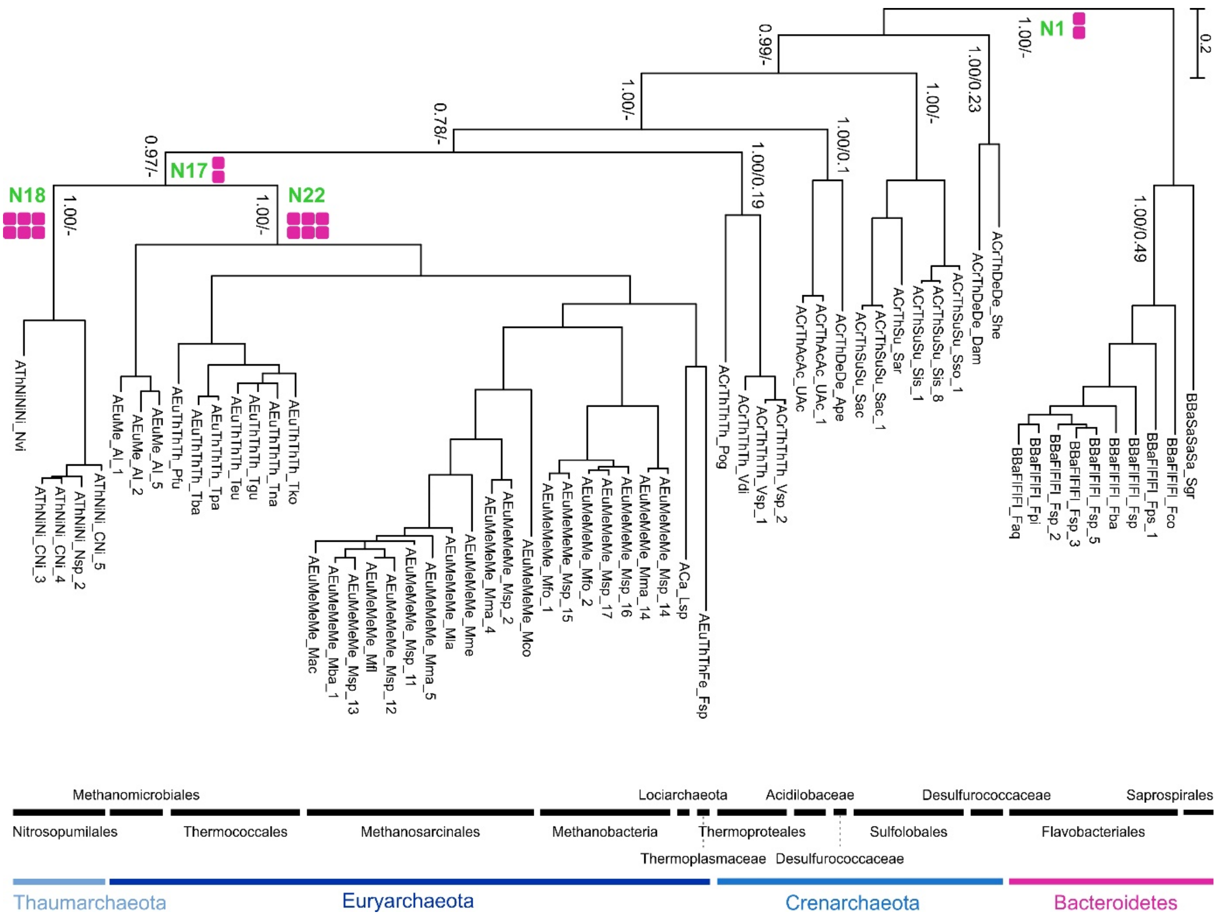

Figure 4. Phylogeny of the sequence set generated by means of `FitSS4ASR` for ASR of GGGPS2 predecessors.

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 Figure 3. The tree was computed by means of `MrBayes`. For central nodes, the posterior probability and (if available) a bipartition score is given. The length of the horizontal bar corresponds to 0.2 substitutions per site. Nodes "N*_auto" labelled in green correspond to evolutionary intermediates that were characterized biochemically. Their dimeric or hexameric oligomerization states, as determined by size exclusion chromatography, are indicated by red symbols.

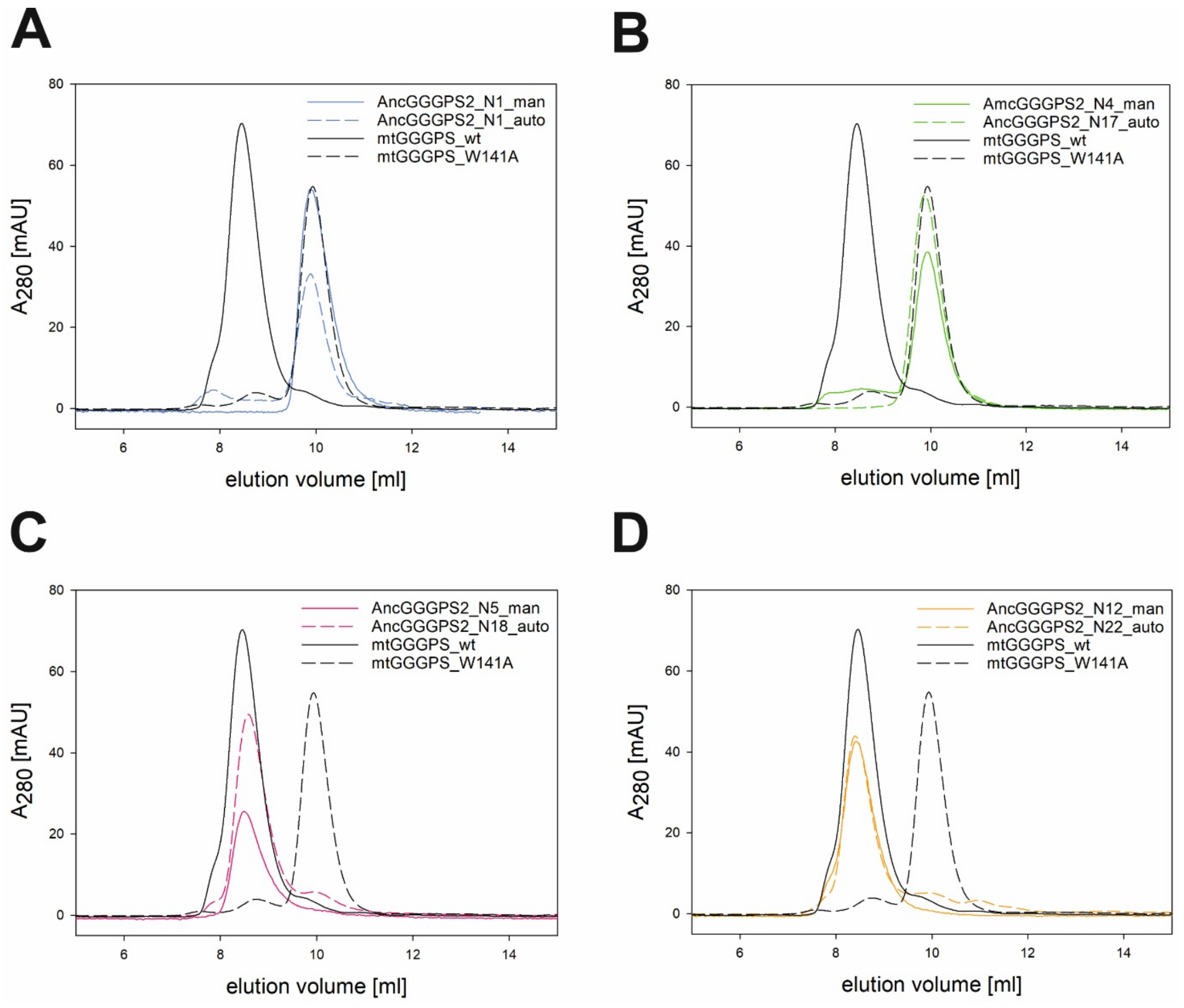

Figure 5. Analytical size exclusion chromatography of ancestral sequences.

The proteins (40 µM subunit concentration each) were applied to a S75 analytical gel filtration column equilibrated with 50 mM potassium phosphate, pH 7.5, 300 mM KCl. Elution was performed at a flow rate of 0.5 ml/min, followed by measuring the absorption at 280 nm (A_{280}) and plotted against the elution volume. mtGGGPS_wt and mtGGGPS_W141A served as references for the hexameric or dimeric oligomerization state, respectively (Peterhoff et al., 2014; Linde et al., 2018). (A) SEC of AncGGGPS2_N1_man and AncGGGPS2_N1_auto, (B) SEC of AncGGGPS2_N4_man and AncGGGPS2_N17_auto, (C) SEC of AncGGGPS2_N5_man and AncGGGPS2_N18_auto, (D) SEC of AncGGGPS2_N12_man and AncGGGPS2_N22_auto.