Review

Rainer Merkl* and Reinhard Sterner*

Ancestral protein reconstruction: techniques and applications

DOI 10.1515/hsz-2015-0158
Received April 10, 2015; accepted July 30, 2015; previously published online August 7, 2015

Abstract: Ancestral sequence reconstruction (ASR) is the calculation of ancient protein sequences on the basis of extant ones. It is most powerful in combination with the experimental characterization of the corresponding proteins. Such analyses allow for the study of problems that are otherwise intractable. For example, ASR has been used to characterize ancestral enzymes dating back to the Paleoarchean era and to deduce properties of the corresponding habitats. In addition, the historical approach underlying ASR enables the identification of amino acid residues key to protein function, which is often not possible by only comparing extant proteins. Along these lines, residues responsible for the spectroscopic properties of protein pigments were identified as well as residues determining the binding specificity of steroid receptors. Further applications are studies related to the longevity of mutations, the contribution of gene duplications to enzyme functionalization, and the evolution of protein complexes. For these applications of ASR, we discuss recent examples; moreover, we introduce the basic principles of the underlying algorithms and present state-of-the-art protocols.

Keywords: ancestral sequence reconstruction; phylogenetic analysis; protein evolution; vertical analysis of protein function.

Introduction

Starting from ancestral precursors, gene duplication and diversification events have yielded families of homologous proteins with highly variable amino acid sequences. Multiple sequence alignments of such proteins allow for the identification of conserved key amino acids, for example active site residues, that are characteristic of the entire protein family (Brown and Babbitt, 2014). However, such an analysis will rarely uncover the set of residues that are responsible for the functional diversity observed in large protein families (Gerlt and Babbitt, 2009). The reason for this is that many neutral as well as epistatic mutations may have accumulated during the evolution of the proteins under comparison. Neutral mutations produce sequence noise that impedes the identification of the crucial mutations, leading to altered functions. Epistatic mutations, i.e. mutations that have different consequences depending on the genetic background, can be divided into permissive and restrictive mutations. Permissive mutations, for example stabilizing ones, are often the prerequisite for a change in function, when the causative key mutation is destabilizing. In contrast, restrictive mutations will prevent a key mutation from becoming effective, for example by introducing steric clashes. Thus, the exchange of putative key residues by site-directed mutagenesis in the framework of a functional analysis can lead to non-functional proteins, when permissive mutations are missing or restrictive mutations are present in the alternative background. As a consequence, the residues that historically led to a new function can often not be identified by comparing extant sequences (Harms and Thornton, 2010).

From an evolutionary point of view, extant homologs are the leaves of a phylogenetic tree and represent variations observed for one specific point in time. Therefore, a comparison of extant sequences was termed “horizontal approach” (Figure 1). It is easy to accept that a “vertical approach”, which additionally takes into account the evolutionary history of the proteins under study, is a more straightforward strategy to identify crucial but subtle amino acid differences (Harms and Thornton, 2010). Instead of exclusively comparing the leaves, such an approach includes the internal nodes of the tree and thus considers the chronology of mutations (Figure 1).

***Corresponding authors: Rainer Merkl and Reinhard Sterner,** Institute of Biophysics and Physical Biochemistry, University of Regensburg, Universitätsstrasse 31, D-93053 Regensburg, Germany, e-mail: Rainer.Merkl@ur.de (Rainer Merkl), Reinhard.Sterner@ur.de (Reinhard Sterner).
http://orcid.org/0000-0002-3521-2957 (Rainer Merkl)

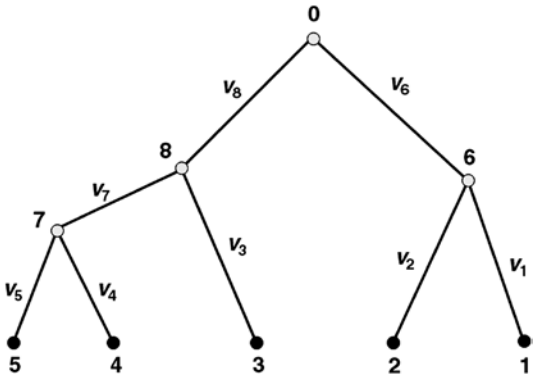

Figure 1: An example of a phylogeny.
Leaves representing extant sequences are labeled 1–5, internal nodes representing reconstructed ancestral sequences are labeled 6–8; 0 represents the root. The exclusive comparison of the leaf sequences is termed a "horizontal" approach; a "vertical" approach additionally takes into account the sequences of the internal nodes. The values v_1–v_8 represent the length of the vertices; example according to Felsenstein (1981).

Moreover, the comparison of the more similar sequences that are related to adjacent nodes reduces the number of neutral mutations and could help to identify epistatic mutations. However, internal nodes represent extinct proteins, whose properties cannot easily be determined, due to the lack of macromolecular fossils. Fortunately, novel computational techniques allow us to reconstruct the sequences of such proteins and to travel back in time (Thornton, 2004; Hanson-Smith et al., 2010). The outcome of these *in silico* approaches, termed ancestral sequence reconstruction (ASR), is a value in itself. Furthermore, in combination with modern gene synthesis technology, these proteins can be produced in recombinant form and characterized by means of all biochemical and biophysical methods at hand.

Sorting out neutral and epistatic mutations is an important but by no means the only application of ASR. Driven to extremes, the most ancient sequences that can be reconstructed are related to the era of the last universal common ancestor (LUCA), which preceded the diversification of life and existed in the Paleoarchean era, i.e. at least 3.8 billion years (Gyr) and presumably 4.5 Gyr ago (Nisbet and Sleep, 2001). Thus, due to the enormous number of known sequences, ASR makes it possible to follow changes in properties like substrate specificity and to reproduce the advent of novel or more specialized functions over this long evolutionary time span. Moreover, certain features of reconstructed macromolecules like stability are correlated with important characteristics of the corresponding habitats. Thus, ASR can implicitly reproduce the adaptation of extinct life to climatic, ecological and physiological alterations (see, for example, Boussau et al., 2008).

ASR was also utilized to characterize the promiscuity of ancestral enzymes (Perez-Jimenez et al., 2011) and to determine the longevity of mutations (Risso et al., 2015). Moreover, the contribution of gene duplications to the evolution of modern enzymes (Voordeckers et al., 2012) and the sophistication of enzyme complexes were studied by means of ASR (Bridgham et al., 2006; Perica et al., 2014).

In the following sections, we will first review *in silico* techniques that have been developed for ASR. Then we will discuss how ASR was used to address the applications mentioned above.

Ancestral sequence reconstruction: history, theoretical background, current protocols

In the following paragraphs, we will survey pioneering experiments, give an introduction to the theory of phylogenetic algorithms, and present state-of-the-art protocols that have been used for ASR.

Pioneering methods of ASR

The idea of reconstructing ancestral amino acid sequences based on a comparison of extant sequences was put forward by E. Zuckerkandl and L. Pauling in 1963 (Pauling and Zuckerkandl, 1963). However, the technology needed for ASR is borrowed from phylogenetic analyses and the first algorithm was not developed until 1971 (Fitch, 1971). Fitch used the principles of maximum parsimony (MP), which is a non-parametric statistical method, to deduce a phylogenetic tree for a given set of extant sequences. Parsimony aims at constructing a tree that minimizes the number of mutations needed to explain the observed data. Thus, the optimality criterion (Swofford et al., 1996) is total tree length *len*(*tree*), which is given by

$$len(tree) = \sum_{k=1}^{B} \sum_{j=1}^{N} w_j diff(x_{k'j}, x_{k''j}) \quad (1)$$

B is the number of branches, N the number of sites (nucleotide or amino acid positions), k' and k'' are the two nodes connected by branch k, and $x_{k'j}$, $x_{k''j}$ are the corresponding nucleotides or amino acids observed in the leaves or inferred for internal nodes. The function *diff*(*y*, *z*)

specifies the cost of a mutation from y to z and w_j assigns a weight to each site. This concept was attractive, because the tree provided the minimal number of mutations required to explain the variations observed in the given extant sequences. The corresponding implementation, named PAUP (Swofford, 1984), has proved popular and the algorithm proposed by Sankoff (Sankoff, 1975) further improved this principle by adding costs to mutations.

This parsimonious principle has been the basis for pioneering work on ribonucleases (RNases) that hydrolyze single- and double-stranded RNA (Stackhouse et al., 1990). The reconstruction required to infer from five extant homologs the protein sequence of a highly conserved RNase of a ruminant that lived 5–10 million years ago. In a follow-up study, 13 artiodactyl RNases were reconstructed and characterized. Among them was the RNase of the founding ancestor of this lineage, i.e. the ancestor of pig, camel, deer, sheep, and ox. The ancestor lived about 40 million years ago, i.e. in a period where ruminant digestion arose. This finding suggests that recent digestive RNases evolved from a non-digestive ancestor; the activity of the reconstructed RNase was fivefold increased against double-stranded RNA (Jermann et al., 1995).

Meanwhile, the drastic increase of computing power allowed for the implementation of considerably more complex algorithms and the limitations of the Fitch approach became evident (see, for example, Frumhoff and Reeve, 1994; Cunningham, 1999). For instance, MP approaches overestimate the number of common to rare changes (Eyre-Walker, 1998). Concurrently, maximum likelihood (ML) approaches have been developed (Yang et al., 1995; Koshi and Goldstein, 1996; Pupko et al., 2000), as well as Bayesian algorithms (Schultz et al., 1996; Huelsenbeck and Bollback, 2001). An ML approach searches for the tree with highest probability (likelihood) given the extant sequences and the parameters of the phylogenetic model used for computation. A Bayesian approach searches for trees with the highest posterior probability. This value results from the prior probabilities of the trees and the likelihood of the data under the given evolutionary model. The current ASR protocols, which are based on these ideas, model the evolution of proteins more precisely than MP. For a detailed history of ASR, see Liberles (2007).

State-of-the-art methods of ASR

In the following section, we briefly introduce some of the stochastic concepts and phylogenetic models that are required to understand modern ASR methods. We give a short description of evolutionary models and trees, which summarizes two recent publications (Liò and Bishop, 2008; Whelan, 2008); for a more detailed presentation, see Liberles (2007). The reader who is familiar with or is not interested in the theoretical background can skip the next six paragraphs.

Assessing mutations by means of Markov models

Phylogenetic trees can be built by means of parametric ML and non-parametric MP, which are *cladistic* approaches, but also with *phenetic* methods, which construct a tree based on a matrix of pairwise distances for the sequences under study. Among the latter is the neighbor-joining algorithm (Saitou and Nei, 1987), which is frequently used to illustrate phylogenetic relationships, because it is a fast and robust method. However, phenetic approaches lack an evolutionary model and therefore these trees can only serve as an approximation, when more complex methods are too computationally expensive. In contrast, protocols for ASR require a precise model of evolutionary processes, which we introduce now.

For the sake of simplicity, we will concentrate on a stochastic model of DNA, which however can easily be extended to codons and proteins. Additionally, we assume independency for the different sites (sequence positions) and therefore, the probability of a set of sequences for a given tree is the product of the probabilities of each site in the sequences.

For each site, $p_{ij}(t)$ is the probability that nucleotide $i \in \{A, C, G, T\}$ will mutate to nucleotide j during the time interval t. Thus, a Markov chain with the state space $S_{DNA}=\{A, C, G, T\}$ and a random variable $X(t) \in S_{DNA}$ describe the substitution process. The homogeneous Markov process, which is used for modeling, assumes that $p(X(s+t)=j|X(s)=i)$ holds, which states that the probability for a replacement of nucleotide i with j within the time interval t is independent of the actual time point $s \geq 0$. We presume now i) a constant rate μ of mutations per unit time (e.g. generation) and ii) a constant prior probability π_j for a mutation leading to nucleotide j. Consequently, the probability that we observe no mutations at the considered site after t generations is $(1-\mu)^t$. Thus, the probability for a mutation is

$$p_{Mut} = 1-(1-\mu)^t \approx 1-e^{-\mu t} \quad (2)$$

and the probability that we observe a change from nucleotide i to nucleotide j within the time interval t is then

$$p_{ij}(t)=\begin{cases}(1-p_{Mut})+p_{Mut}\pi_j & i=j\\ p_{Mut}\pi_j & i\neq j\end{cases} \quad (3)$$

Instead of utilizing discrete generations, the probability can also be determined in continuous time. It follows for the set of all mutations in analogy to equation (3):

$$p(t+dt)=p(t)+p(t)\mathbf{Q}dt=p(t)(\mathbf{I}+\mathbf{Q}dt) \quad (4)$$

Here, **I** is the unit matrix and **Q** is a rate matrix of transition probabilities and we get

$$p(t)=e^{t\mathbf{Q}} \quad (5)$$

Substitution models

For DNA, 16 π_j values are needed to assess all possible mutations and several matrices **Q** have been proposed. A first model was introduced by Jukes and Cantor (1969); more frequently used are the models introduced later by M. Kimura (1981) and J. Felsenstein (Felsenstein, 1981).

For a more precise assessment of protein evolution, a codon-based model has been designed by Z. Yang (Yang and Nielsen, 2000). Here, the elements of **Q** describe the rate of change of codon $i=i_1i_2i_3$ to $j=j_1j_2j_3$ depending on rates for transitions and transversions. This approach has been improved (Yang et al., 2000) and some of the models M0 to M13 are popular choices for evolutionary analyses based on codons.

Alternatively, mutational events can also be studied on the level of amino acid sequences. Early amino acid substitution models are related to the PAM-matrices of M. Dayhoff (Dayhoff et al., 1978). However, due to the small numbers of sequences available at that time, several of these substitution frequencies are crude approximations. The more recent JTT-matrix (Jones et al., 1992) and the WAG-matrix (Whelan and Goldman, 2001) have been deduced from much larger data sets and are thus more realistic models. It is known that not all domains or regions of a protein evolve under the same evolutionary constraints. Thus, specific matrices **Q** have been deduced for transmembrane and non-transmembrane regions, for α-helices, β-strands, or for buried and exposed residues (Koshi and Goldstein, 1996). However, these matrices are seldom used for ASR.

Meanwhile, homogeneous substitution models have been replaced with more complex ones. A continuous distribution, which provides every site with a specific rate, seems most plausible. Often, a gamma distribution was used (Liò and Bishop, 2008), which represents a full family of probability distributions, whose shape depend on parameters α and β. However, it has been shown that a discrete "gamma model" performs well and is computationally efficient (Yang, 1994). It consists of only four, equally probable categories of rates, which were chosen to approximate a gamma distribution. The density of the gamma distribution $G(\alpha, \beta)$ is

$$g(r;\alpha,\beta)=\frac{\beta^\alpha}{\Gamma(\alpha)}\exp(-\beta r)\cdot r^{\alpha-1},\ 0<r<\infty \quad (6)$$

In this context, α is a given or estimated shape parameter and the scale parameter β is redundant and can be set equal to α, so that the mean of the distribution is 1. The range of r $(0, \infty)$ is divided into $k=4$ categories by means of cutting points, and each category is characterized by a rate r_i that indicates the mean of the portion of the gamma distribution falling in the category; see Yang (1994). Consequently, the unconditional probability $p(x)$ for observing symbol x at a site is related to the rate-specific conditional probabilities through

$$p(x)=\int p(x|r)g(r)dr\approx\sum_{i=1}^{k}\frac{1}{k}p(x|r=r_i) \quad (7)$$

Here, $g(r)$ is the gamma density with parameter α, which is chosen so that $r_1, \ldots, r_k$ give the largest approximate likelihood $\prod_i p(x_i)$ and $p(x|r)$ is the conditional probability of x given the rate r at a site (Susko et al., 2003).

The likelihood of a phylogenetic tree

Using the above model of evolution, the likelihood of a tree can be computed. Likelihood is the probability for observing the data (sequences) given the parameters of the chosen evolutionary model and the topology of the tree under study. If all sites mutate independently, then this likelihood is the product of all site-specific values. Thus, to explain the principle, it is sufficient to consider one site $S(j)$ of a sequence S and to compute the likelihood for the nucleotides at $S(j)$ at each node of the tree. If the length of all edges, which corresponds to a certain time interval, is v_i and if all nucleotides e_i are known for all nodes i, which implies a certain ancestral labeling, then the likelihood of the tree shown in Figure 1 is:

$$L(tree)=\pi_{e_0}p_{e_0e_6}(v_6)p_{e_6e_1}(v_1)p_{e_6e_2}(v_2)p_{e_0e_8}(v_8)\\ p_{e_8e_3}(v_3)p_{e_8e_7}(v_7)p_{e_7e_4}(v_4)p_{e_7e_5}(v_5) \quad (8)$$

However, the states (nucleotides) of the internal nodes are not known and therefore it is necessary to sum

over all possible parameter values (nucleotides at internal nodes) which results in

$$L(tree)=\sum_{e_0}\sum_{e_6}\sum_{e_7}\sum_{e_8}\pi_{e_0}p_{e_0e_6}(v_6)p_{e_6e_1}(v_1)p_{e_6e_2}(v_2)p_{e_0e_8}(v_8)\,p_{e_8e_3}(v_3)p_{e_8e_7}(v_7)p_{e_7e_4}(v_4)p_{e_7e_5}(v_5) \tag{9}$$

As introduced above, π_{e0} is the prior probability of nucleotide e_0. The value $L(tree)$ can be computed quite efficiently after a rearrangement of terms, which considers the topology of the tree; compare the pattern of brackets {[][]} {[][()()]} in equation 10 and the tree topology shown in Figure 1:

$$L(tree)=\sum_{e_0}\pi_{e_0}\left\{\sum_{e_6}p_{e_0e_6}(v_6)[p_{e_6e_1}(v_1)][p_{e_6e_2}(v_2)]\right\}\left\{\sum_{e_8}p_{e_0e_8}(v_8)[p_{e_8e_3}(v_3)]\left[\sum_{e_7}p_{e_8e_7}(v_7)(p_{e_7e_4}(v_4))(p_{e_7e_5}(v_5))\right]\right\} \tag{10}$$

This arrangement of terms suggests a bottom-up computation based on the likelihood values $L^k_{e_k}$ of all states e_k at node k. The calculation starts in the first iteration with the known likelihood values of the leaves, which are 1 for the observed nucleotide and 0 for all others. The $L^k_{e_k}$ values of internal nodes are computed in a bottom-up fashion by considering the tree topology and by summarizing likelihood values of two children r, s, which were calculated in one of the preceding iterations:

$$L^k_{e_k}=\left(\sum_{e_r}p_{e_ke_r}(v_r)L^r_{e_r}\right)\left(\sum_{e_s}p_{e_ke_s}(v_s)L^s_{e_s}\right) \tag{11}$$

Finally $L(tree)$ is

$$L(tree)=\sum_{e_0}\pi_{e_0}L^0_{e_0} \tag{12}$$

It follows that the likelihood of a tree can be computed iteratively, if we know all transition probabilities p_{ij}. The missing length of the edges can be determined by means of expectation maximization (Dempster et al., 1977); for details see Felsenstein (1981). However, in order to find the tree with the maximal likelihood, the topology – which was taken as given so far – has to be optimized as well, which requires the creation and the assessment of alternative topologies. Due to its complexity, the problem of determining the ML tree is an NP-hard problem for the computer scientist (Chor and Tuller, 2005), which means that in practice only an approximation can be found in an acceptable time interval.

For a comparison of alternative trees, ML approaches maximize the likelihood value given in equation 12; for Bayesian inference, trees have to be sampled based on a score deduced from their likelihood and prior expectations. However, the number of tree topologies grows exponentially with the number of sequences, which necessitates heuristic approaches to sample tree space. Commonly, these approaches progressively optimize the tree by examining the score of similar trees, choose the highest scoring one as the next estimate, and finally stop, if no further improvement can be found.

Popular traversal schemes of tree space propose candidate trees by making small rearrangements on the current tree, examine each internal branch of the tree in turn, and vary the way they alter the topology. New topologies are created by means of quartet puzzling (QP) (Strimmer and Von Haeseler, 1996), nearest neighbor interchange (NNI), subtree pruning and regrafting (SPR) and tree bisection and reconnection (TBR) (see Whelan, 2008). QP and NNI break an internal edge, which gives four subtrees. These can then be combined in three different ways, which give novel candidate trees. SPR is more general than NNI and QP by adding subtrees to any edge of the other subtrees. TBR removes one edge to create two subtrees and all possible combinations of the two subtrees give new candidate trees (for details see Whelan, 2008). As expected, the time complexity of the algorithms is high: For QP, algorithms of $O(n^4)$ have been reported (Ranwez and Gascuel, 2001), a Markov chain Monte Carlo (MCMC) approach (see below) of NNI is of $O(pnl)$, where p is the number of refinement steps and n is the number of taxa, i.e. sequences of length l (Guindon et al., 2005). In contrast, the phenetic neighbor-joining algorithm is of $O(n^3)$, which means that execution time increases roughly with the third power of the number of input sequences.

This outline of these heuristic algorithms makes clear that there is no guarantee for finding the optimal tree that has the ML. However, the rearrangements of the tree topology under study expand the area of searched tree space, which increases the probability of finding a nearly optimal solution. Searching a wider range is additionally supported by starting the refinement from different points in tree space; thus, these computations are often executed in parallel on a multi-processor computer with varied starting conditions.

Bayesian inference of topologies

For Bayesian inference, the topology of the tree and the parameters of the evolutionary model are estimated

simultaneously, while providing a measure of confidence in those estimates. Bayesian inference uses the same models of evolution as ML methods, however, it addresses a number of complex questions of phylogeny; see Huelsenbeck et al. (2001). For ASR, it is important that the tree with the maximum posterior probability can be deduced from a large number of sampled trees. To do so, Bayes's theorem is used favorably to combine the prior probability of a tree $p(tree)$ with the likelihood $L(tree)=p(data|tree)$ to compute a posterior probability distribution of trees $p(tree|data)$ according to

$$p(tree|data)=\frac{p(data|tree)\cdot p(tree)}{p(data)} \quad (13)$$

The posterior probability of a tree gives the probability that the tree is correct and often the tree with the maximum a posterior probability (MAP) is chosen as the best estimate. In contrast to ML, Bayesian approaches generally include a prior expectation about the problem under study. For phylogeny, uninformative priors are frequently used, which means that all trees are equally likely and the likelihood $L(tree)$ can be calculated analogously as described above.

Unfortunately, a comprehensive analysis of the posterior distribution is not feasible as it requires a summation over all possible tree topologies. However, an approximation of the posterior distribution can be determined by means of MCMC methods, which generate a series (chain) of pseudo-random samples. MCMC approaches can correctly sample from the posterior probability, because newly proposed trees are accepted based on a probability function. The probability of being accepted depends primarily on the difference in likelihood scores of the current and the new tree. Thus, the chain will contain many trees that offer an improvement over initial trees and few trees with poor scores. If the parameters are sampled correctly, the amount of time a chain spends in different regions of tree space corresponds to the posterior distribution, which allows a straightforward approximation of the MAP tree.

For a more detailed description of an MCMC approach, which is taken from Larget and Simon (1999), we define a tree $\psi=(\tau, \beta)$ by its topology τ and associated branch lengths β. Additionally we need a likelihood model $L(x|\omega)$ for observed data x that contains several parameters, where $\omega=(\psi, \varphi)$ represents a specific choice of a tree topology, branch lengths, and model parameters φ. The corresponding parameter space $\Omega=(\Psi, \Phi)$ contains the set of all possible trees Ψ and all possible values of the model parameters Φ. As MCMCs utilize the Metropolis-Hastings criterion (Metropolis et al., 1953) to accept new solutions, the chains create a dependent series of points in Ω, $\omega^{(0)}$, $\omega^{(1)}$, $\omega^{(2)}$, …, such that after some time all subsequently sampled points are distributed approximately according to their posterior distribution. As a consequence, the long-run frequencies are arbitrarily close to their posterior probabilities. In order to scan $\Omega=(\Psi, \Phi)$, a combination of two update mechanisms has been proposed (Larget and Simon, 1999), to sample tree topologies and model parameters. The algorithm starts with an initial tree (e.g. a neighbor-joining tree) and model parameters $\omega^{(0)}=(\psi^{(0)}, \varphi^{(0)})$, which are randomly chosen. Subsequently, each individual cycle i+1 of the algorithm consists of two steps that utilize the parameters of the current state $\omega^{(i)}=(\psi^{(i)}, \varphi^{(i)})$. In the first step – while keeping the current tree $\psi^{(i)}$ fixed – the algorithm can choose new model parameters φ^* from Φ, which are – according to the Metropolis-Hastings algorithm – either accepted $\varphi^{(i+1)}=\varphi^*$ or rejected $\varphi^{(i+1)}=\varphi^{(i)}$. The second step modifies the current tree $\psi^{(i)}$, while holding the parameters $\varphi^{(i+1)}$ fixed. For more details of the update algorithms see, for example, Larget and Simon (1999).

When applying MCMC methods, the adequate sampling of the posterior distribution is related to two factors named *convergence* and *mixing*. Convergence means that the chain accurately samples from the posterior distribution, which is the case after the pre-convergence phase (named also burn-in phase). Thus, trees computed in the burn-in phase are ignored. A chain mixes well, if all trees can be reached from all other ones. In contrast, if a chain is mixing poorly, the sampling of the posterior probability is compromised. In order to assess the quality of a computation, i.e. convergence and mixing, each implementation of an MCMC algorithm offers a series of diagnostic tools that compute specific indicators. One can assume convergence, if several chains that were started in parallel with different initial parameter sets concentrate their sampling in the same region of tree space. Comparing trees taken from converged chains allows for the analysis of mixing: If these trees are clearly different, the chains are mixing poorly. Additionally, model parameters or likelihood values can be plotted vs. the sample number; more details can be found in the documentation of the respective programs. For ASR, a high posteriori probability and a short length are of specific relevance for all edges, as both are prerequisites for the reliability of the reconstructed sequences.

More complex models of evolutionary processes

The above-described methods reconstruct a phylogeny for one sequence per species – or few concatenated ones – and

aim at representing the history of the considered genes (gene products). Consequently, the resulting trees were named gene trees; however, in many cases the evolutionary history of a gene differs substantially from the history of the species from which the genes originate (Maddison, 1997; Szöllősi and Daubin, 2012). Duplication, horizontal gene transfer and gene loss cause drastic differences in the size and composition of genomes and thus produce phylogenetic discord. Moreover, horizontally transferred genes are frequently acquired from species that are extinct or do not belong to the dataset under study (Szöllősi et al., 2013). This is why a gene phylogeny does often not provide enough information to distinguish between statistically equivalent relationships, a fact that is indicated by poor posteriori probabilities of individual edges.

One reason for horizontal gene transfer is the recombination of genetic material. A statistical model that considers recombination in ancestral sequences may give rise to a graph that is no longer a tree but a network. It is feasible to estimate the number of recombination events in a given sample of sequences, however the algorithm is very computationally intensive (see Griffiths and Marjoram, 1996).

An alternative approach, which can more easily be integrated into an ASR protocol, is based on the reconciliation of a gene tree and a second one that reflects the phylogeny of the considered species. The topology of this tree is also affected by the extra processes that contribute to the evolution of species, which are speciation and lineage sorting, gene duplication and loss, and gene transfer. If these additional evolutionary events, which are ignored for gene tree computation, are considered, all models of gene family evolution can be seen as generating a tree inside a tree, that is, a gene tree inside a species tree (Szöllősi et al., 2015).

In pioneering work, an MP tree was determined that minimizes the number of nucleotide substitutions, gene duplications and gene losses (Goodman et al., 1979). Meanwhile, more complex models for species evolution were beneficially integrated in gene tree inference (Maddison, 1997; Akerborg et al., 2009; Groussin et al., 2015) and species tree can be deduced from the shared history of several gene families; see, for example, De Oliveira Martins et al. (2014); Mirarab et al. (2014); Mirarab and Warnow (2015).

Deducing ancestral sequences

For a phylogenetic analysis, two types of trees, unrooted and rooted ones, can be computed. In contrast to rooted trees, unrooted trees do not specify the location of the common ancestor. For rooting, which is required for the ASR of this ancestor, an outgroup can be used or, alternatively, the position of the root is approximated based on additional phylogenetic knowledge. For time-reversible models, which are commonly used for computation, the position of the root does not affect the likelihood score (Felsenstein, 1981), which allows for subsequent rooting. Using the set of extant sequences and the corresponding phylogenetic tree, the most plausible ancestral sequences can be deduced by a ML reconstruction following the principles introduced above. Applying the Bayesian approach, a reconstruction maximizes the probability for the set of ancient sequences given the extant ones (Pupko et al., 2000). Two variants of ancestral ML reconstructions exist (Yang et al., 1995), originating from different optimization criteria, which are the joint ML or the marginal ML, respectively. For ASR of proteins, joint reconstruction determines the most likely set of amino acids for all internal nodes at a site, which yields the maximum joint likelihood of the tree. In contrast, marginal reconstruction compares the probabilities of different amino acids at an internal node at a site and selects the one amino acid that yields the ML for the tree at that site (Cai et al., 2004). The resulting sequences may differ, and marginal reconstruction is considered to be an approximation of the joint approach (Pupko et al., 2000).

The basic idea of a ML ancestral reconstruction can be illustrated by concentrating on the internal nodes of a tree whose topology and branch lengths are assumed to be known. The tree given in Figure 1 has five operational taxonomic units (the leaves, labeled 1–5) and four hypothetical taxonomic units (HTU) labeled 0, 6, 7, 8. For each site of these four internal nodes there are 20^4 combinations of amino acids e_i. It is the aim of joint ML to identify for these four nodes a quartet ν with the largest value $p(\nu|data)$, which is in a Bayesian approach the quartet that maximizes

$$\frac{p(data|\nu)\cdot p(\nu)}{p(data)} \tag{14}$$

As $p(data)$ is identical for all candidate quartets, it is sufficient to maximize $p(data|\nu)\cdot p(\nu)$. More specifically, for this tree and the given four nodes the quartet is found by solving

$$\max_{e_0,e_6,e_7,e_8} [\pi_{e_0} p_{e_0e_6}(v_6) p_{e_6e_1}(v_1) p_{e_6e_2}(v_2) p_{e_0e_8} (v_8) p_{e_8e_3}(v_3) p_{e_8e_7}(v_7) p_{e_7e_4}(v_4) p_{e_7e_5}(v_5)] \tag{15}$$

The solution computed for equation 15 is the maximum over all possible 20^4 quartets. For larger trees with h HTUs, it is necessary to maximize overall h ancestral states,

which results in 20^h combinations. To solve this problem of joint ML reconstruction, an algorithm, which is based on dynamic programming, has been implemented that scales linearly with the number of sequences; see Pupko et al. (2000) for details. If all joint probabilities are known, the marginal distribution can by computed by marginalizing over joint probabilities; for an example see Pupko et al. (2007).

Current software protocols for ASR

In the previous section, we have introduced basic principles of phylogenetic methods, which are required for ASR. Now, we describe in more detail the software protocols that were used in those ASR experiments we will review in the following paragraphs. Generally, each protocol for ASR requires four steps (A–D) that are depicted in Figure 2.

(A) Select extant sequences: Commonly, homologous sequences were retrieved from databases like GenBank of the NCBI or UniProtKB of the EBI (see, for example, Boussau et al., 2008; Gaucher et al., 2008; Finnigan et al., 2012; Voordeckers et al., 2012; Bar-Rogovsky et al., 2013; Risso et al., 2013; Perica et al., 2014), most often with the help of BLAST (Altschul et al., 1990). If the number of hits was very large, highly similar sequences were eliminated by using CD-HIT (Li and Godzik, 2006) to create a set of sequences with 30–90% identical residues (Bar-Rogovsky et al., 2013). Alternatively, highly similar sequences, e.g. those with more that 92% pairwise sequence identity, were removed (Voordeckers et al., 2012). The number of extant sequences required for an ASR depends on protein-specific mutation rates and the time span of interest. Thus, the size of the resulting sequence sets varied between 11 (Yokoyama et al., 2008), 32 (Bar-Rogovsky et al., 2013) and up to 200 or more sequences (Perez-Jimenez et al., 2011; Harms et al., 2013). In some cases, protein sequences have been concatenated, like those of the HisF and HisH enzymes that constitute a heterodimeric complex (Reisinger et al., 2014) or 56 nearly universally distributed proteins (Boussau et al., 2008). Moreover, the protein sequences and the corresponding DNA sequences were in some cases compiled and analyzed in parallel to eliminate ambiguities (Ugalde et al., 2004; Field and Matz, 2010; Hobbs et al., 2012; Voordeckers et al., 2012).

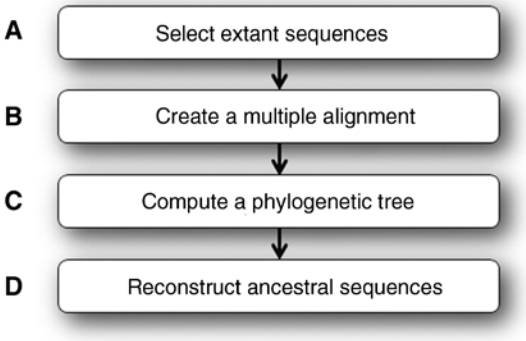

Figure 2: Protocol for ASR.
Each ASR requires four steps to deduce ancestral sequences from a set of extant homologs. (A) A set of extant sequences is retrieved from a database. (B) The selected sequences are aligned in a multiple sequence alignment, which allows for the identification of mutational events separating the sequences. (C) A phylogeny is determined; the extant sequences form the leaves. (D) Based on this phylogenetic tree and the input data, ancestral sequences are computed for each internal node of the tree.

(B) Create a multiple alignment: Due to the complexity of the algorithmic problem, heuristic approaches are the only way of computing a multiple sequence alignment (MSA), which is required to map residues to protein positions. During recent years, several algorithms have been introduced that show comparable alignment quality. Therefore, it is not surprising that different methods were used for MSA creation. Among them were Clustal X and Clustal W (Larkin et al., 2007) used in Bridgham et al. (2006) and Hobbs et al. (2012) as well as MUSCLE (Edgar, 2004), which was used frequently (Boussau et al., 2008; Richter et al., 2010; Perez-Jimenez et al., 2011; Eick et al., 2012; Perica et al., 2014). The algorithm PRANK (Löytynoja and Goldman, 2008) considers insertions and deletions as distinct evolutionary events and has been shown to prevent systematic errors related to the gap placement of more traditional MSA methods. PRANK was utilized in more recent ASR experiments (Bar-Rogovsky et al., 2013; Reisinger et al., 2014). In some cases, regions of ambiguous alignment were removed from the MSA by applying GBLOCKS (Castresana, 2000) prior to the subsequent computation of a phylogeny (see Reisinger et al., 2014).

(C) Compute a phylogenetic tree: It is state of the art to deduce a phylogeny by means of an ML or a Bayesian approach. Among ML approaches, PAUP (Swofford, 1984) has been chosen (Gaucher et al., 2003; Perez-Jimenez et al., 2011) as well as GARLI (Bazinet et al., 2014), which was used in Hobbs et al. (2012), and PAML (Yang et al., 1995) was used in Akanuma et al. (2013) and Finnigan et al. (2012). Frequently used implementations of Bayesian approaches are MrBayes (Ronquist and Huelsenbeck, 2003; see Ugalde et al., 2004; Bridgham et al., 2006; Gaucher et al., 2008; Field and Matz, 2010; Voordeckers et al., 2012; Risso et al., 2013; Perica et al., 2014), PhyML (Guindon and Gascuel, 2003; see Eick et al., 2012;

Bar-Rogovsky et al., 2013; Harms et al., 2013), and PhyloBayes (Lartillot et al., 2009; see Reisinger et al., 2014). nhPhyloBayes (Blanquart and Lartillot, 2008) is a non-homogeneous Bayesian approach that was also utilized, see Akanuma et al. (2013).

Alternatively, phylogenetic relationship between the species was deduced from the literature (Yokoyama and Radlwimmer, 2001) or a user-defined time-calibrated mammalian phylogeny was assembled via Mesquite (Maddison and Maddison, 2015; see Mirceta et al., 2013). Moreover, node age estimates were made (Hobbs et al., 2012) using the ML branch lengths and two calibration points taken from the literature (Battistuzzi et al., 2004).

A large set of supplementary programs support phylogenetic studies and ASR. To select the best fitting model for the data set at hand, ProtTest (Abascal et al., 2005) was used (Bar-Rogovsky et al., 2013), which aims at identifying the best generating evolutionary model. In one case, the resulting parameters were fed into GARLI (Bazinet et al., 2014) to find the best ML tree (Hobbs et al., 2012). The validity of the phylogeny was confirmed with different approaches. The quality of the PhyML tree (Guindon and Gascuel, 2003) was assessed by means of a bootstrap resampling test (Bar-Rogovsky et al., 2013). These bootstrap values were calculated with RAxML (Stamatakis, 2006) and the topology was evaluated by means of a Shimodaira-Hasegawa test (Shimodaira and Hasegawa, 2001; Anisimova et al., 2011), which was used similarly in Finnigan et al. (2012). Often, MCMC convergence was checked, e.g. by using the AWTY program (Nylander et al., 2008) as in (Voordeckers et al., 2012) or by determining other parameters indicating convergence and well mixing of the chains. For example, the maximum difference of posterior probabilities of tree bipartitions and the posterior number of biochemical profile categories was estimated (Richter et al., 2010). Additional tests were performed to exclude long-branch attraction artifacts (see, for example, Voordeckers et al., 2012).

(D) Reconstruct ancestral sequences: The extant sequences chosen in step (A) and the phylogenetic tree determined in step (C) combined with a substitution model form the basis for the computation of the ancestral sequences. Most often, the one sequence with the highest likelihood has been considered for each internal node, see, for example, Perica et al. (2014). These ancestral sequences were deduced by means of MrBayes (Ronquist and Huelsenbeck, 2003) and a simple F81-like model (Felsenstein, 1981). Frequently, the functions CODEML and ML of PAML (Yang, 1997) were utilized (Yokoyama and Radlwimmer, 2001; Bridgham et al., 2006; Gaucher et al., 2008; Yokoyama et al., 2008; Perez-Jimenez et al., 2011; Hobbs et al., 2012; Akanuma et al., 2013; Ingles-Prieto et al., 2013) in combination with gamma distributions modeling variable replacement rates across sites. However, different substitution matrices were chosen: in some cases the JTT model (Jones et al., 1992) was the basis for a marginal reconstruction and the synthesis of ancestral enzymes (Voordeckers et al., 2012), the same model was used in (Eick et al., 2012) in combination with the Lazarus software (Hanson-Smith et al., 2010). Alternatively, the WAG (Whelan and Goldman, 2001) substitution model was utilized (Risso et al., 2013). In Field and Matz (2010) and Ugalde et al. (2004) PAML was combined with three alternative ML models. Those were the amino acid based JTT (Jones et al., 1992), the codon-based M5 (Yang et al., 2000), and the nucleotide based GTR+G3 (Tavaré, 1986) model. In this case, the posterior probability of the marginal reconstruction at each site served as a measure of accuracy. Alternatively, the ML approach of FastML (Pupko et al., 2000) was applied (Bar-Rogovsky et al., 2013). In Mirceta et al. (2013), only the most probable amino acid sequence was considered, which was determined utilizing the ML approach implemented in MEGA5 (Tamura et al., 2011) in combination with the Dayhoff+G (Dayhoff et al., 1978) model. Moreover, the non-homogeneous models nhPhyML (Boussau and Gouy, 2006) and nhPhyloBayes (Blanquart and Lartillot, 2008) were also integrated into sequence reconstruction (see Boussau et al., 2008; Richter et al., 2010).

Open issues and problems to be solved in ASR

Is ASR applicable to any protein of interest? Presumably not. A phylogeny can be computed if a sufficiently large set of sequences is at hand. However, it is the quality of the resulting tree that decides on the meaningfulness of a reconstruction: if the length of any edge indicates that a rate of more than one mutation per site separates the adjacent nodes, the corresponding sequences cannot be reconstructed reliably. Additionally, the topology of the tree has to be unambiguous – which demands for high a posteriori probabilities or bootstrap values – although this property is not extremely crucial for ASR (Hanson-Smith et al., 2010).

Moreover, one has to keep in mind that phylogenetic models make implicit assumptions about the data. For example, it is assumed that the proteins under study share a common ancestor and that all sequences have evolved independently. The first assumption is violated, for example if multi-domain proteins are examined that possess only one common domain. The second one is violated, if sequences were exchanged via horizontal gene transfer. For the reconstruction of sequences related to the LUCA, i.e. the most

ancestral node, midpoint rooting is frequently applied as no outgroup exists. However, this method is not generally accepted (Perez-Jimenez et al., 2011).

The major concern with respect to the reliability of ASR is whether the resurrected proteins display the same characteristics as the authentic ancestral proteins (Gaucher et al., 2008). A reconstructed sequence is some kind of consensus sequence and it has been argued, for example, that the higher thermostability observed in many ASR projects is due to selecting the most probable residue at each site (Williams et al., 2006). Higher equilibrium frequencies of hydrophobic residues in the amino acid substitution matrices may strengthen this effect, especially if the underlying tree contains long branches. Moreover, these matrices have been deduced from extant proteins and their use for ASR has been questioned (Brooks and Gaucher, 2007). Thus, if thermostability is a major issue, special care has been taken to exclude these effects in some applications of ASR (see, for example, Gaucher et al., 2008).

For the practitioner, two further problems complicate the application of ASR: These are (i) the selection of a representative sample, if a large number of sequences is at hand and (ii) the correct modeling of larger insertions and deletions.

To tackle the first problem, tools that choose sequences leading to an unambiguous phylogeny would be helpful. Alternatively, reconciliation methods like the recently introduced TERA approach (Scornavacca et al., 2015) incorporate species trees into gene tree reconstruction and promise drastic improvements in accuracy. Thus, it seems reasonable to integrate these concepts into ASR protocols in order to simplify sequence selection. A first application of a species tree-aware ASR, namely the resurrection of the LeuB enzyme for the ancestor of Firmicutes was successful (Groussin et al., 2015).

The correct modeling of loops, which underlies the second problem, is still in its infancies. It has to be shown whether algorithms like PRANK (Löytynoja and Goldman, 2008) in combination with an adapted reconstruction protocol (Perica et al., 2014) contribute to the correct phenotype.

Applications of ASR

Deducing environmental conditions of the Precambrian era

An important application of ASR is to "replay the molecular tape of life" (Gaucher, 2007). Along these lines, billions-of-years old Precambrian proteins have been resurrected (Gaucher et al., 2008; Perez-Jimenez et al., 2011; Akanuma et al., 2013; Ingles-Prieto et al., 2013; Risso et al., 2013; Reisinger et al., 2014; Risso et al., 2014). As a side-effect, these studies allow one to obtain information about environmental conditions surrounding Precambrian life.

For example, G.E. Gaucher and colleagues determined the thermal melting temperature (T_m) of resurrected translation elongation factors from organisms living from 3.5 to 0.5 Gyr ago (Gaucher et al., 2008). The results suggest that ancient life cooled progressively by 30°C during this period. In accordance with this finding, an almost identical cooling trend for the ancient ocean was inferred from the deposition of silicon isotopes (Robert and Chaussidon, 2006).

An analogous cooling trend has been deduced from the analysis of ancestral thioredoxins (Perez-Jimenez et al., 2011). Seven Precambrian thioredoxin enzymes dating back between about 4 Gyr and 1.4 Gyr were resurrected, which are related to the last bacterial common ancestor (LBCA), the last archaeal common ancestor (LACA), and the archaeal-eucaryotic common ancestor (AECA). These organisms are thought to have inhabited Earth 4.2–3.5 Gyr ago diverging from the LUCA, which could not be reconstructed due to technical difficulties. DSC measurements showed that these enzymes are up to 32°C more stable than modern enzymes, and a plot of the T_m vs. geological time revealed a linear decrease with a slope of about 6°C/Gyr (Figure 3). An activity assay based on single molecule spectroscopy and an artificial substrate showed that ancient thioredoxins used a similar reaction mechanism as modern thioredoxins.

The crystal structures of the resurrected thioredoxins possess the canonical thioredoxin fold (Ingles-Prieto et al., 2013), meaning that the chemistry and three-dimensional structure of thioredoxin were already established around 4 Gyr ago. This observation suggests that the step from simple reducing compounds to well-structured and functional enzymes occurred early in molecular evolution (Nisbet and Sleep, 2001). Remarkably, ancient thioredoxins are significantly more active at pH 5 than extant ones, which fits to the proposed acidity of the ancient oceans (Walker, 1983). Taken together the natural habitat in which LBCA, LACA, and AECA lived was likely acidic as well as hot in accordance with the plausible hypothesis that early life thrived in seawater.

The assumption of a hot environment for early life is further supported by an analysis of nucleotide kinases (NDKs) (Akanuma et al., 2013). Ancestral NDK sequences were computed based on two different resurrection

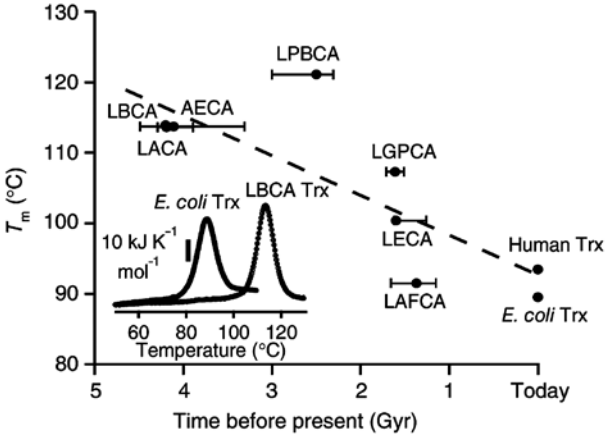

Figure 3: Denaturation temperatures (T_m) versus geological time for ancestral thioredoxin (Trx) enzymes.
Modern *Escherichia coli* and human Trx enzymes are also indicated. The dashed line represents a linear fit to the data. Inset, experimental DSC thermograms for *E. coli* Trx and LBCA Trx. For ancestral thioredoxins, the following abbreviations were used: LBCA, last bacterial common ancestor; LACA, last archaeal common ancestor; AECA, archaeal-eukaryotic common ancestor; LECA, last eukaryotic common ancestor; LPBCA, last common ancestor of the cyanobacterial and deinococcus and thermus groups; LGPCA, last common ancestor of γ-proteobacteria; LAFCA, last common ancestor of animals and fungi. Figure taken from Perez-Jimenez et al. (2011).

strategies. At pH 6, the reconstructed archaeal NDKs have T_m values of around 110°C, and the bacterial ones have T_m values of 109°C and 102°C. These values are higher than those for the thermophilic archaeon *Archaeoglobus fulgidus* and the thermophilic bacterium *Thermus thermophilus*, respectively.

However, not all enzymes followed this general cooling trend in their evolution. An interesting case of a more recent thermal adaptation to the local habitat is the metabolic enzyme 3-isopropylmalate dehydrogenase (LeuB) reconstructed for *Bacilli* (Hobbs et al., 2012). Four ancestral sequences (ANC1–ANC4) of LeuB were determined. Each was positioned progressively deeper in the phylogeny and further back in time. ANC1, 2, 3, and 4 are approximately 679, 820, 850, and 950 million years old, respectively. All four ANC enzymes exhibited kinetic parameters comparable to their homologs from contemporary *Bacillus* species. The thermoactivity profiles and the thermal melting temperatures of ANC1–ANC4 were compared and showed a sharp decline in thermophily from ANC1 to ANC2, followed by a gradual increase in thermophily from ANC2, through ANC3 to ANC4. A more detailed sequence analysis demonstrated that the mechanisms of thermal stability differ between ANC1 and ANC4, i.e. thermophily within *Bacillus* LeuB evolved twice. As there is a good correlation between the growth temperature of an organism and the thermostability of its proteins (Gromiha et al., 1999), the authors hypothesize that the observed fluctuations in thermophily reflect changes in the microenvironment encountered by the evolving *Bacillus* species.

In the previous examples, environmental properties were estimated by characterizing resurrected proteins with the help of biophysical methods. A pure *in silico* analysis was presented by the group of M. Gouy, which reconstructed rRNA and protein sequences for the LUCA and the ancestors of the three domains of life (Boussau et al., 2008). The rRNA sequences consisted of 1043 sites from the double-stranded regions of the small and the large ribosomal subunit, and the protein sequences comprised 3336 sites from 56 nearly universally distributed proteins. These ancestral sequences were used to deduce the ambient temperature from the G+C content of the rRNA and from the content of the amino acids I, V, Y, W, R, E, and L in the protein sequences. For both parameters, a strong correlation with the optimal growth temperature is known (Galtier and Lobry, 1997; Zeldovich et al., 2007). In contrast to other findings, these parameters characterize the LUCA as a non-hyperthermophilic species. Interestingly, the bacterial ancestor and the archeal ancestor as well as the common ancestor of Archaea and Eukaryotes were estimated as being thermophilic or hyperthermophilic species. In summary, these findings argue for a temperature increase during the era of the LUCA descendants.

Promiscuity of ancestral enzymes

It has been postulated that ancestral enzymes were promiscuous, i.e. processed several different substrates during the very first steps of their evolution (Khersonsky and Tawfik, 2010). This hypothesis has been tested for several protein families using ASR.

J. Sanchez-Ruiz and colleagues have studied four Precambrian ancestors of β-lactamase (Risso et al., 2013). These ancient β-lactamases were indeed promiscuous: activity assays showed that they are able to hydrolyze various β-lactam antibiotics with catalytic efficiencies similar to those of an average modern enzyme. In contrast, the modern β-lactamase TEM-1 is a specialist that can only hydrolyze penicillin. The comparison of crystal structures of extant and ancestral β-lactamases revealed the same topologies and no significant differences of the active sites. It was concluded that substrate promiscuity of the ancestral β-lactamases is probably due to altered dynamics, a hypothesis that was substantiated by extensive

molecular dynamics simulations. The modern TEM-1 β-lactamases showed a relatively rigid active site region, likely reflecting adaptation for efficient degradation of a specific substrate (penicillin), whereas the observed enhanced flexibility in the ancestral β-lactamases might allow for the binding of antibiotics with different sizes and geometries (Zou et al., 2015). In spite of their high flexibility, the ancestral proteins were about 35°C more thermostable than extant β-lactamases including the one from the thermophile *Bacillus licheniformis*.

This ancestral promiscuity can be explained if ancient bacteria benefitted from producing a variety of β-lactam antibiotics. Such antibiotics could have served as a device to achieve nutrients by killing competitors, and β-lactamases might have arisen as a mechanism of defense (Risso et al., 2014).

However, promiscuity is not a general feature of ancient enzymes: A chymase is a serine protease that converts angiotensin I to angiotensin II. An ancestral chymase, which was deduced from mammalian sequences, had an efficient and specific angiotensin II-forming activity. Thus, it was postulated that the less specific serine proteases evolved later and broadened their substrate spectra, thereby losing specificity (Chandrasekharan et al., 1996). Moreover, the reconstructed cyclase subunit of imidazole glycerol phosphate synthase from LUCA (LUCA-HisF) can convert only the native HisF substrate into product but not the related HisA or TrpF substrates (Reisinger et al., 2014).

The LUCA is a construct that may exemplify a single organism (Woese, 1998) or may represent populations of organisms capable of sharing large amounts of genetic information through horizontal gene transfer (Doolittle, 2000). Either way, it is clear now that organisms at the time of the LUCA possessed many of the fundamental features present in modern organisms and likely exhibited a level of sophistication comparable with modern Bacteria or Archaea (Becerra et al., 2007). The findings of Reisinger et al. (2014) are in full agreement with this notion: LUCA-HisF forms a high-affinity imidazole glycerol phosphate synthase complex with its associated glutaminase subunit LUCA-HisH.

Long-term persistence of beneficial residues in proteins

It is an interesting question whether the preferences for different amino acids at a given site in a protein change during evolution or remain essentially constant. V. Risso and colleagues have used resurrected thioredoxins to address this problem (Risso et al., 2015). To this end, the authors have experimentally determined the effects of 21 mutations on the stability of both *Escherichia coli* thioredoxin and the thioredoxin of the LBCA. Fourteen mutations were identical for the extant and ancestral proteins in terms of the introduced amino acid exchange. For example, there is valine at position 16 both in the *E. coli* and the LBCA thioredoxin. Thus, the Val16Ile exchange was studied in both cases. In contrast, seven mutations had to be analyzed in opposite directions for the extant and ancestral backgrounds. For example, there is an isoleucine at position 23 in *E. coli* thioredoxin but a valine at the same position in LBCA thioredoxin. Thus, the Ile23Val mutation was studied in the extant background whereas the Val23Ile mutation was studied in the ancestral background. Importantly, the extant and ancestral proteins substantially differ in the residues present in the molecular neighborhood of the targeted positions. Nevertheless, the effect of the mutation (stabilizing or destabilizing) – when considered in the same direction – was qualitatively identical for the extant and the ancestral proteins. Taken together, this study suggests that site-specific amino acid preferences in a protein have essentially remained unchanged over long geological timescales even when the amino acids themselves changed during evolution. The evolutionary persistence of a destabilizing mutation might be explained by the fact that it leads to an enhanced fitness of the organism caused by functional advantages.

Following the evolution of receptor-ligand interactions

Can few mutations induce major shifts in protein function and if so, what are the underlying mechanisms? To study this question, the group of J. Thornton analyzed the structural basis of the different hormone sensitivities of the estrogen receptors (ERs) and the non-aromatized steroid receptors (naSRs) (Harms et al., 2013). Previous investigations had revealed that the ancestor of the entire steroid receptor family (AncSR1) had ER characteristics with respect to hormone binding, whereas its immediate phylogenetic descendant (AncSR2) was sensitive to androgens, progestogens, mineralocorticoids, and glucocorticoids, and thus had naSR characteristics (Figure 4A) (Eick et al., 2012). The AncSR1 sequence is most similar to those of the extant ERs, whereas that of AncSR2 is most similar to the naSRs, and this pattern is most pronounced at sites in the ligand-contacting pockets (Figure 4B). Altogether, AncSR1 and AncSR2 differ by 171 residues corresponding to a sequence divergence of 70%. Among them

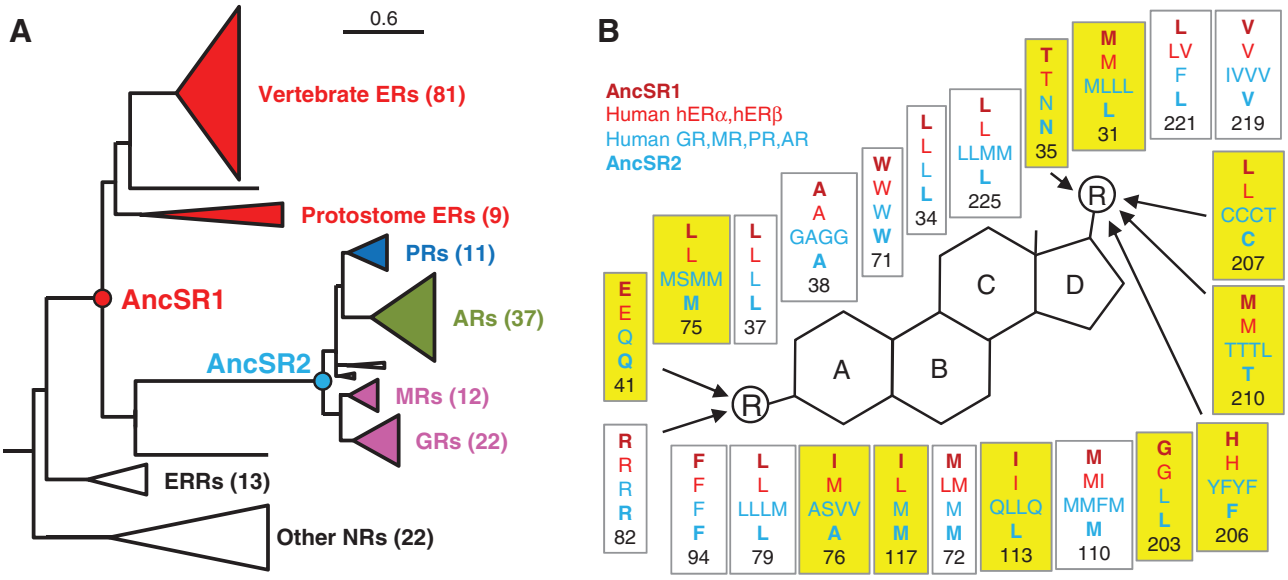

Figure 4: Evolution of steroid receptors and their ligands.
(A) Phylogeny of the SR gene family. Receptors are color-coded by the classes of ligands to which they are most sensitive. These are estrogens (red), progestagens (blue), androgens (green), and corticosteroids (purple). ERRs are estrogen related receptors and "Other NRs" are other nuclear receptors. The ancestral steroid receptors (AncSR1 and AncSR2) that were resurrected are marked as circles. The number of sequences is shown for each clade in parentheses. (B) Maximum likelihood reconstruction of ligand-contacting residues in AncSR1 and AncSR2 and the residues at homologous sites in extant human SRs. A circled R indicates a polar functional group, at which the major steroid classes differ from each other and arrows indicate residues within hydrogen bonding distance. Residues that differ between AncSR1 and AncSR2 are highlighted in yellow. Modified image adapted from Eick et al. (2012).

are 22 residues that are in AncSR1 identical to the residue observed in the extant ERs and in AncSR2 identical to the residue observed in the extant naSRs. Two residues out of the 22 differences appeared to be involved in differential hormone binding according to a comparison of the structures of AncSR1 and AncSR2. The two AncSR2-specific residues (Gln41 and Met75) were replaced by the AncSR1-specific residues (Glu41 and Leu75), and vice versa. Characterization of the AncSR2 variant showed that it had AncSR1-like hormone binding characteristics, and characterization of the AncSR1 variant showed that it had AncSR2-like hormone binding characteristics. Thus, just two relatively minor amino acid differences are responsible for the distinct ligand specificities of these two major clades of vertebrate hormone receptors.

Understanding the evolution of biological systems consisting of tightly integrated parts is difficult, due to the mutual dependency of the interacting partners. J. Thornton and colleagues used a vertical approach to elucidate the stepwise adaptation in the functional interaction between the steroid hormone aldosterone and its binding partner, the mineralocorticoid receptor (Bridgham et al., 2006; Ortlund et al., 2007). The authors were interested in identifying the key residue differences between different steroid receptors in two related systems, namely the mineralocortocoid receptor (MR), which is activated by aldosterone and to a lesser extent by cortisol, and the glucocorticoid receptor (GR), which is activated by cortisol only (Bridgham et al., 2006; Ortlund et al., 2007). It was found that the common ancestor of all MRs and GRs (AncCR) was MR-like. By resurrecting successive ancestors in the GR lineage, it was shown that cortisol-specificity evolved between AncGR1 (MR-like phenotype) and AncGR2 (GR-like phenotype). Within this branch, 37 residue differences occurred but only five have been conserved in one state in the MRs and in another state in the GRs. These residues were introduced into AncGR1, singly and in pairs. None of the single mutations increased cortisol-specificity, but the combination of Ser106Pro and Leu111Gln did. A strong epistatic effect with respect to these two mutations was observed: Leu111Gln alone had little effect on sensitivity to any hormone, but Ser106Pro dramatically reduced activation by all ligands. Only the combination switched receptor preference from aldosterone to cortisol. Introducing these substitutions into the human MR yielded a completely non-functional receptor, as did reversing them in the human GR. These results emphasize that the functional impacts of historical substitutions can only be evaluated with the ancestral sequence at hand.

Elucidating mutational routes leading to highly specific chromophores

Two other studies, focusing on the absorbance and fluorescence properties of chromophores, also nicely illustrate how ASR can help to identify key residues distinguishing highly specific functions of homologs.

In the first study, the sequence determinants that are relevant for the absorption of opsins in the green or the red wavelength region were identified (Yokoyama et al., 2008). Earlier work had shown that the replacement of three putative key residues in red opsin with the corresponding residues in green opsin yielded a pigment absorbing in the green wavelength region. However, the reverse procedure did not yield an opsin absorbing in the red (Asenjo et al., 1994). ASR showed that the common ancestor of all red and green opsins absorbed maximally in the red. A sequence comparison identified five positions that were specifically conserved in red and green opsins but differed between the two forms. When the five green-specific residues were introduced in the ancestral opsin, it displayed a shift from the red to the green. Then, each mutation was introduced singly and in sets of two or three. The results showed that a large fraction of the total green shift was the result of epistatic mutations rather than the direct effects of the individual mutations.

In the second study, the sequence determinants responsible for the different fluorescence properties of GFP variants from the coral suborder Faviina were characterized (Ugalde et al., 2004; Field and Matz, 2010). ASR revealed that the ancestral GFP-like protein fluoresced in the green; subsequent diversification resulted in the emission of a variety of colors (Figure 5). The authors were interested to identify residues that lead to a shift from the green ancestor to the red GFP that is present in a certain star coral. They first found that these two proteins differed by 37 residues. As it was impossible to test all possible combinations of exchanges, they generated a library of variants that comprised about half of the residues in the ancestral green state and half in the derived red state. Then, the fluorescence of a large number of clones from this library was correlated with the amino acids found

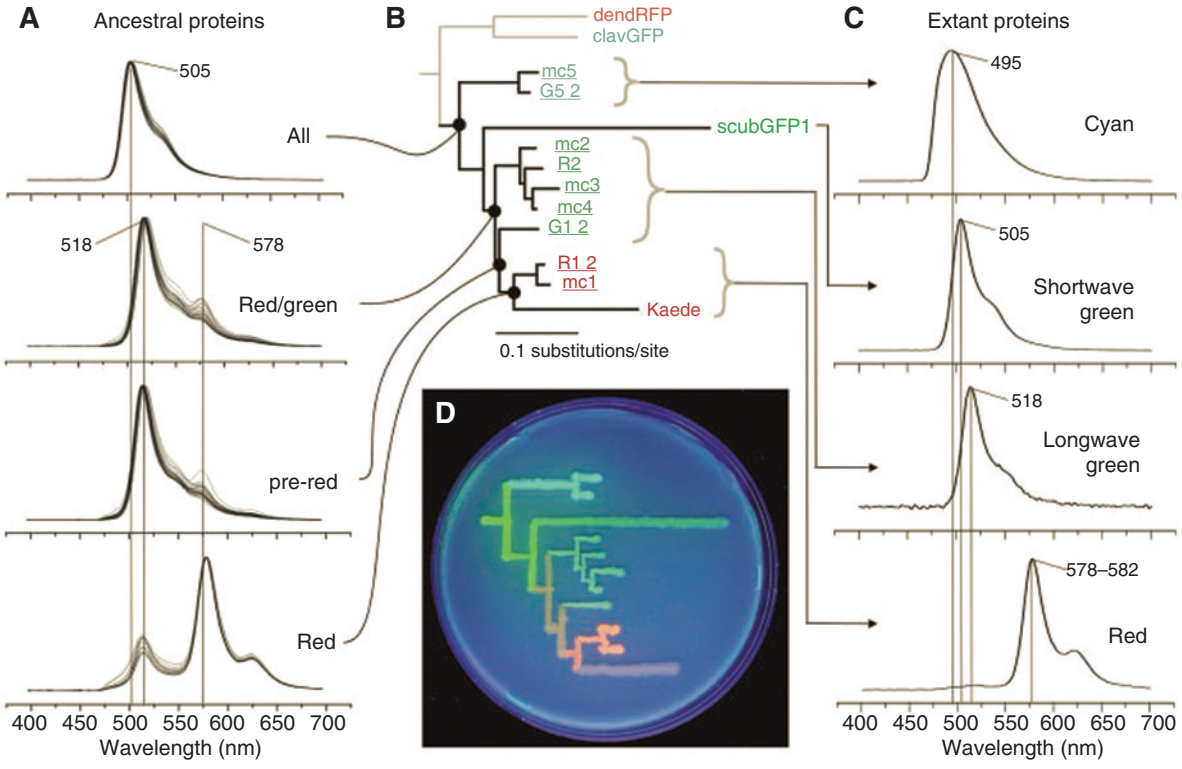

Figure 5: ASR of GFP variants.
(A) Fluorescence spectra of the reconstructed ancestral proteins. Multiple curves correspond to clones bearing variations at degenerate sites. (B) Phylogeny of GFP-like proteins from the great star coral *Montastraea cavernosa* and closely related coral species. The red and cyan proteins from soft corals (dendRFP and clavGFP) represent an outgroup. (C) Fluorescence spectra of extant proteins. (D) Phylogenetic tree of colors from the great star coral, drawn on a petri dish with bacteria expressing extant and ancestral proteins, under ultraviolet light. Figure taken from Ugalde et al. (2004).

at different positions. The statistical analysis of the data indicated that 12 of the 37 residues were crucial for red fluorescence. The introduction of these 12 residues into the green ancestral protein yielded a protein that emitted in the red. This work thus illustrates that crucial amino acids responsible for different properties of proteins can be identified by a combination of ASR with library selection.

Gene duplications and their contribution to the evolution of modern enzymes

ASR was also used to study the effect of gene duplication on evolutionary innovation (Voordeckers et al., 2012). Following gene duplication, three evolutionary scenarios are feasible that explain the subsequent function of the gene products: (i) one copy can retain the old function and the other copy can adopt a new one (neofunctionalization); (ii) it is also possible that the ancestral gene product has two different functions, which might be split between the two copies (subfunctionalization); (iii) finally, the two copies may preserve the same activity; in such a case, gene duplication would increase the activity by increasing the concentration of the encoded protein (gene dosage effect).

In this study, a family of fungal enzymes (MALS) was analyzed that hydrolyze disaccharides. These enzymes all originated from the same ancestral gene and underwent several duplication events. Activity data were obtained for the very first preduplication enzyme ancMALS, for the subsequent ancestral enzymes ancMAL-IMA, ancMAL and ancIMA1-5, and for the seven extant MALS enzymes from *Saccharomyces cerevisiae* (Figure 6).

The results show that ancMALS and ancMAL-IMA were promiscuous, preferring maltose-like substrates

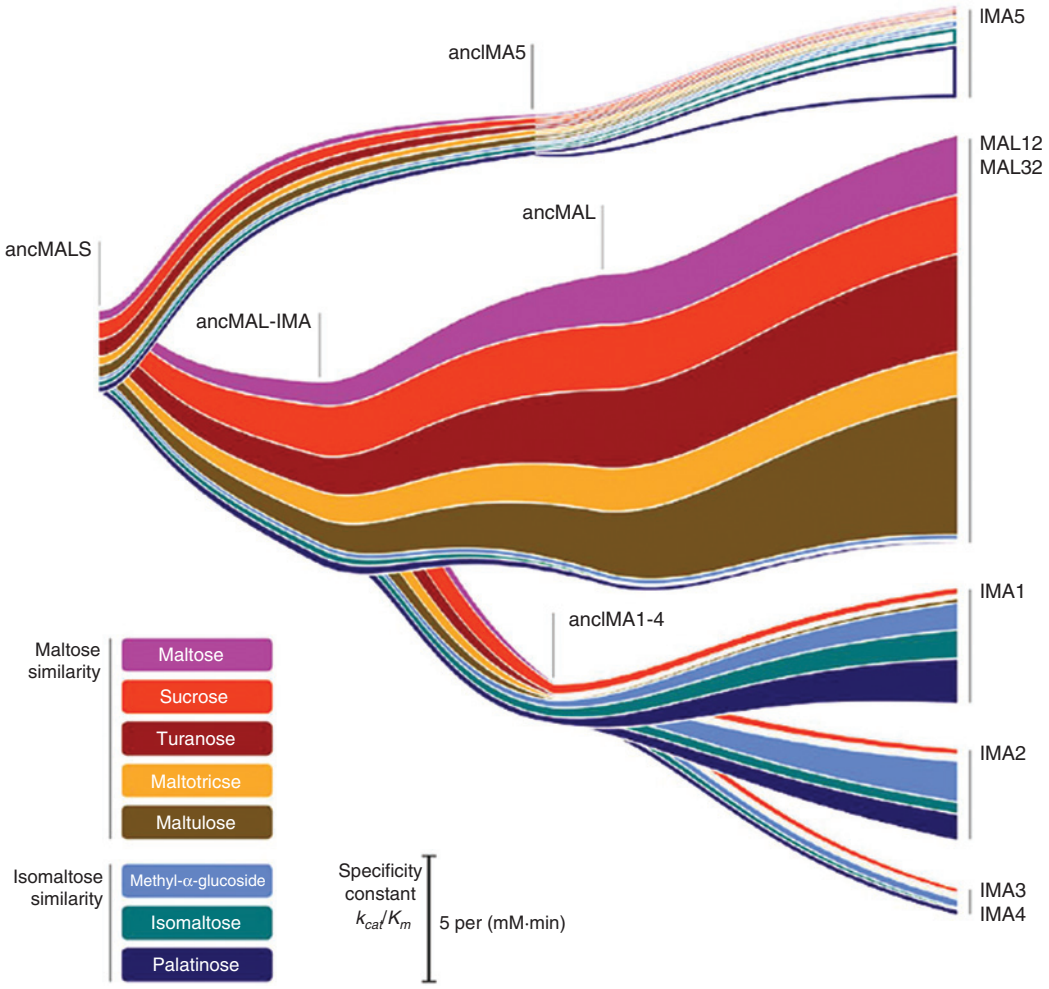

Figure 6: Duplication events and changes in specificity and activity during the evolution of *Saccharomyces cerevisiae* MALS enzymes. The hydrolytic activity of seven modern MAL and IMA enzymes and of key ancestral enzymes (prefix anc) is given. The width of the colored bands corresponds to the k_{cat}/K_m-value of the enzyme for a specific substrate. For details see Voordeckers et al. (2012).

such as maltose, maltotriose, maltulose, sucrose, and turanose, but also displaying trace activity towards isomaltose-like sugars such as palatinose and isomaltose. A clear divergence of both subfunctions only occurred after duplication of ancMAL-IMA, resulting in the specialized ancMAL and ancIMA1-4 proteins. This subdivision is also present in the seven extant enzymes: two enzymes (MAL12 and MAL32) show high activity towards maltose-like substrates, whereas the enzymes of the second class (IMA1-4) show high activity for isomaltose-like substrates. These findings illustrate how, after duplication, the different copies diverged and specialized in one of the functions present in the preduplication enzyme. Interestingly, it was found that evolution has taken two different molecular routes to optimize isomaltose-like activity (the evolution of ancMAL-IMA to ancIMA1–4 and ancIMA5 to IMA5; Figure 6). Molecular modeling and site-directed mutagenesis studies revealed that the observed different substrate specificities are caused by different evolutionary routes: when going from ancMAL-IMA to ancIMA1–4 position 279 is crucial, whereas when going from ancMALS to IMA5 the same shift in substrate specificity is caused by residue 219.

Taken together, the data suggest that the evolutionary history of the MALS family exhibits aspects of all three classical models of gene evolution after duplication: the preduplication enzyme was multifunctional and already contained the different activities found in the postduplication enzymes, which is in agreement with the idea of subfunctionalization. However, the isomaltose-like activity was very weak in the preduplication ancestor and only fully developed through mutations after duplication, which resembles neofunctionalization. Moreover, considerable fitness costs that were observed when one of two almost identical copies of extant MALS proteins was deleted suggest that gene dosage effects may also play an important role in the evolution of this enzyme group.

Gene duplication and subsequent specialization are also the basis for the evolution of increased complexity in a molecular machine (Finnigan et al., 2012). The V_0 ring of extant V-ATPases from fungi contains three different subunits, Vma3, Vma11, and Vma16, which are arranged in a specific orientation (Figure 7A). Phylogenetic analysis showed that Vma3 and Vma11 are sister proteins that are derived from an ancestral protein (Anc3–11) via a gene duplication event. Anc3–11 as well as the last common ancestors of Vma3 (Anc3), Vma11 (Anc11), and Vma16 (Anc16) were reconstructed (Figure 7B). It was found that Anc16 can complement a ΔVma16 strain. Likewise, Anc3–11 (but not Anc3 or Anc11) could complement a yeast ΔVma11ΔVma3 double deletion strain. These findings show that an ancestral two-subunit ring can functionally replace the extant three-subunit ring of yeast. Subsequent subunit fusion experiments demonstrated that

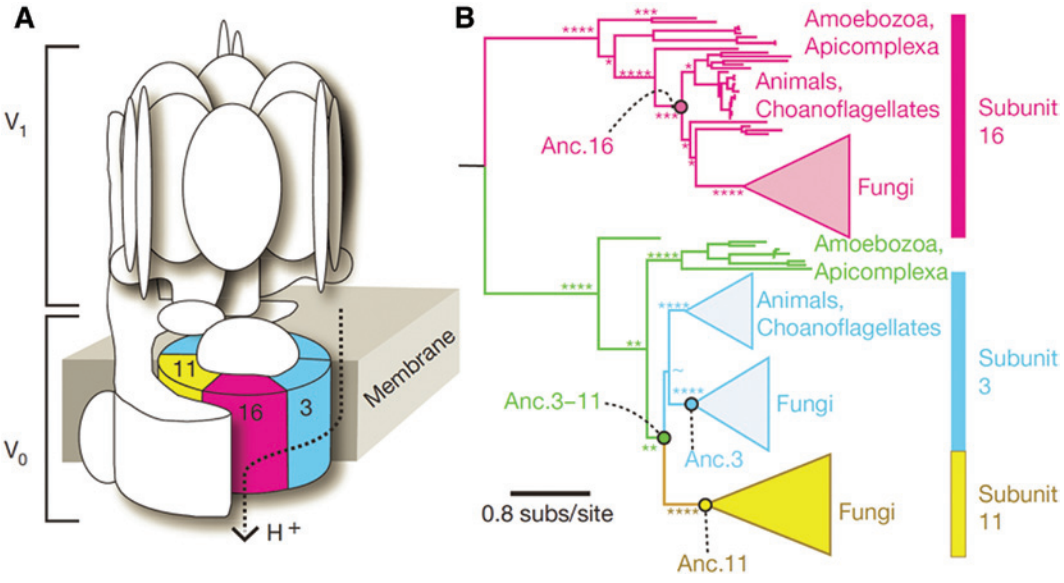

Figure 7: Structure and evolution of the V-ATPase complex.
(A) In *Saccharomyces cerevisiae*, the V-ATPase contains two subcomplexes: the octameric V1 domain on the cytosolic side of the organelle membrane, and the membrane bound hexameric V0 ring. Subunits Vma3, Vma11, and Vma16 are color-coded. (B) Maximum likelihood phylogeny of V-ATPase subunits Vma3, Vma11, and Vma16. The genomes of all eukaryotes contain subunits 3 and 16, but Fungi also contain subunit 11. Circles show reconstructed ancestral proteins, colors correspond to those of subunits in panel (A); unduplicated orthologs of Vma3 and Vma11 are green. Asterisks show approximate likelihood ratios for major nodes. Figure taken from Finnigan et al. (2012).

Vma3 and Vma11 evolved their specialized roles because they lost specific asymmetric interactions present in Anc3–11 that are required for ring assembly. These losses were complementary, so both copies, Vma3 and Vma11, became obligate components with restricted spatial roles in the complex. Site-directed mutagenesis with Anc3–11 was used to recapitulate this asymmetric degeneration: a single amino acid replacement that occurred on the branch leading to Anc11 abolished the capacity of Anc3–11 to function as subunit 3. Conversely, a single amino acid replacement that occurred on the branch leading to Anc3 radically reduced the capacity of Anc3-11 to function as subunit 11.

Following the evolution of quaternary complexes

Commonly, the quaternary configuration of homologous protein complexes is highly conserved. However, in some protein families, different oligomeric states are observed. An example is PyrR, which is a pyrimidine operon attenuator in *Bacillaceae*. Here, the thermophilic ortholog (BcPyrR) forms a tetramer whereas the mesophilic ortholog (BsPyrR) is a dimer. In order to dissect the role of the 49 substitutions that distinguish BsPyrR from BcPyrR, S. Teichmann and colleagues combined ASR with biophysical methods and structural analysis (Perica et al., 2014). Comparing the 3D structures and residue contact networks of variants, 11 allosteric key mutations were identified that control the oligomeric state. The results made clear that evolution utilized the intrinsic dynamics of this protein to toggle a conformational switch in the same manner as the binding of a small molecule does, which is related to the function of this attenuator.

Conclusion

The above examples illustrate that evolutionary analysis can help to solve biochemical and biological problems, which are not accessible with other methods. However, to address these problems, it is not sufficient to simply reconstruct the protein sequences. Instead, the functionally important mutations have to be identified, and the physical effects mediating them have to be uncovered by means of a biochemical and biophysical characterization of the resurrected proteins.

A number of ASR experiments have confirmed that Precambrian life was thermophilic, which is in accordance with several scenarios, including that ancestral oceans were hot, that ancient life thrived in hot spots such as hydrothermal systems, or that only robust thermophilic organism survived bombardment events in the young Earth (Risso et al., 2014). Moreover, several publications suggest that essential enzymes had already reached a high level of functional sophistication in the LUCA era. Furthermore, crystal structure analysis of Precambrian thioredoxins (Ingles-Prieto et al., 2013), β-lactamases (Risso et al., 2013), nucleoside kinases (Akanuma et al., 2013) and the imidazole glycerol phosphate synthase HisF (Reisinger et al., 2014) made clear that the three-dimensional structures of these proteins are similar to those of the corresponding extant proteins, supporting a relatively slow evolution of protein structure and function as compared to amino acid sequences.

Although ASR is unavoidably uncertain to some extent, the presented studies show that ASR is validated to a significant degree at the phenotypic level by the fact that the properties of the proteins resurrected in the laboratory are typically robust. Moreover, their capacities are consistent with the ancestral properties expected from physical science and paleogeology. Additionally, state-of-the-art applications of ASR acknowledge that reconstructed ancestors are approximations of historical reality. For example, several studies carefully explored the robustness of their functional inferences to uncertainty about the reconstructed ancestors by experimentally characterizing alternate plausible reconstructions (see, for example, Thornton, 2004; Ugalde et al., 2004).

We have shown that ASR counts on state-of-the-art phylogeny. These methods will further improve due the permanent increase of computing power, which allows for the implementation of more sophisticated models. Additionally, a much larger number of extant sequences can be exploited, which makes plausible that the uncertainty related to ASR will further decrease.

It has been argued that ASR has a tendency to overpredict highly stable predecessors (Perica et al., 2014), a suspicion that cannot be ruled out completely. Conversely, hyperstability in combination with promiscuity is a winning combination from the protein-engineering point of view, because both features contribute to high evolvability (Risso et al., 2013).

What are the limitations of ASR? This technology is entirely dependent on phylogenetic trees, which consequently hampers the analysis of protein evolution for the pre-LUCA era. Thus, the experimental simulation of this very early phase of evolution requires alternative approaches, which again combine *in silico* analyses and proteins characterization (see, for example, Farias-Rico et al., 2014).

What comes next? The next steps are the integration of ancestral protein complexes into modern organisms – for first examples see Finnigan et al. (2012) and Reisinger et al. (2014) – and in the long-run the reconstruction of a full ancestral microorganism. To do so, the ancestral genomic content has to be determined at first. This task is feasible with methods resembling the approaches introduced here (see Tuller et al., 2010; Jones et al., 2012; Yang et al., 2012), but a more detailed survey is out of the scope of the current review.

Acknowledgments: Work by the authors was supported by the Deutsche Forschungsgemeinschaft (ME2259/2-1, STE891/9-1).

References

Abascal, F., Zardoya, R., and Posada, D. (2005). ProtTest: selection of best-fit models of protein evolution. Bioinformatics *21*, 2104–2105.

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. USA *110*, 11067–11072.

Akerborg, O., Sennblad, B., Arvestad, L., and Lagergren, J. (2009). Simultaneous Bayesian gene tree reconstruction and reconciliation analysis. Proc. Natl. Acad. Sci. USA *106*, 5714–5719.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403–410.

Anisimova, M., Gil, M., Dufayard, J.F., Dessimoz, C., and Gascuel, O. (2011). Survey of branch support methods demonstrates accuracy, power, and robustness of fast likelihood-based approximation schemes. Syst. Biol. *60*, 685–699.

Asenjo, A.B., Rim, J., and Oprian, D.D. (1994). Molecular determinants of human red/green color discrimination. Neuron *12*, 1131–1138.

Bar-Rogovsky, H., Hugenmatter, A., and Tawfik, D.S. (2013). The evolutionary origins of detoxifying enzymes: the mammalian serum paraoxonases (PONs) relate to bacterial homoserine lactonases. J. Biol. Chem. *288*, 23914–23927.

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

Bazinet, A.L., Zwickl, D.J., and Cummings, M.P. (2014). A gateway for phylogenetic analysis powered by grid computing featuring GARLI 2.0. Syst. Biol. *63*, 812–818.

Becerra, A., Delaye, L., Islas, S., and Lazcano, A. (2007). The very early stages of biological evolution and the nature of the last common ancestor of the three major cell domains. Annu. Rev. Ecol. Evol. Syst. *38*, 361–379.

Blanquart, S. and Lartillot, N. (2008). A site- and time-heterogeneous model of amino acid replacement. Mol. Biol. Evol. *25*, 842–858.

Boussau, B. and Gouy, M. (2006). Efficient likelihood computations with nonreversible models of evolution. Syst. Biol. *55*, 756–768.

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.

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

Brooks, D.J. and Gaucher, E.A. (2007). A thermophilic last universal ancestor inferred from its estimated amino acid composition. In: Ancestral Sequence Reconstruction, D.A. Liberles, ed. (Oxford, UK: Oxford University Press), pp. 200–207.

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

Cai, W., Pei, J., and Grishin, N.V. (2004). Reconstruction of ancestral protein sequences and its applications. BMC Evol. Biol. *4*, 33.

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

Chandrasekharan, U.M., Sanker, S., Glynias, M.J., Karnik, S.S., and Husain, A. (1996). Angiotensin II-forming activity in a reconstructed ancestral chymase. Science *271*, 502–505.

Chor, B. and Tuller, T. (2005). Maximum likelihood of evolutionary trees: hardness and approximation. Bioinformatics *21*(Suppl 1), i97–106.

Cunningham, C.W. (1999). Some limitations of ancestral character-state reconstruction when testing evolutionary hypotheses. Syst. Biol. *48*, 665–674.

Dayhoff, M.O., Schwartz, R.M., and Orcutt, B.C. (1978). A model of evolutionary change in proteins. In: Atlas of Protein Sequence and Structure, M. Dayhoff, ed. (Washington, DC, USA: National Biomedical Research Foundation), pp. 345–352.

De Oliveira Martins, L., Mallo, D., and Posada, D. (2014). A Bayesian supertree model for genome-wide species tree reconstruction. Syst. Biol. pii:syu082.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Ser. B. (Stat. Method.) *39*, 1–38.

Doolittle, W.F. (2000). The nature of the universal ancestor and the evolution of the proteome. Curr. Opin. Struct. Biol. *10*, 355–358.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. *32*, 1792–1797.

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

Eyre-Walker, A. (1998). Problems with parsimony in sequences of biased base composition. J. Mol. Evol. *47*, 686–690.

Farias-Rico, J.A., Schmidt, S., and Höcker, B. (2014). Evolutionary relationship of two ancient protein superfolds. Nat. Chem. Biol. *10*, 710–715.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. *17*, 368–376.

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.

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

Fitch, W.M. (1971). Toward defining the course of evolution: minimum change for a specific tree topology. Syst. Biol. *20*, 406–416.

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

Galtier, N. and Lobry, J.R. (1997). Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes. J. Mol. Evol. *44*, 632–636.

Gaucher, E.A. (2007). Ancestral sequence reconstruction as a tool to understand natural history and guide synthetic biology: realizing and extending the vision of Zuckerkandl and Pauling. In: Ancestral sequence reconstruction, D.A. Liberles, ed. (Oxford, UK: Oxford University Press), pp. 20–33.

Gaucher, E.A., Thomson, J.M., Burgan, M.F., and Benner, S.A. (2003). Inferring the palaeoenvironment of ancient bacteria on the basis of resurrected proteins. Nature *425*, 285–288.

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

Gerlt, J.A. and Babbitt, P.C. (2009). Enzyme (re)design: lessons from natural evolution and computation. Curr. Opin. Chem. Biol. *13*, 10–18.

Goodman, M., Czelusniak, J., Moore, G.W., Romero-Herrera, A.E., and Matsuda, G. (1979). Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. Syst. Zool. *28*, 132–163.

Griffiths, R.C. and Marjoram, P. (1996). Ancestral inference from samples of DNA sequences with recombination. J. Comput. Biol. *3*, 479–502.

Gromiha, M.M., Oobatake, M., and Sarai, A. (1999). Important amino acid properties for enhanced thermostability from mesophilic to thermophilic proteins. Biophys. Chem. *82*, 51–67.

Groussin, M., Hobbs, J.K., Szöllősi, G.J., Gribaldo, S., Arcus, V.L., and Gouy, M. (2015). Toward more accurate ancestral protein genotype–phenotype reconstructions with the use of species tree-aware gene trees. Mol. Biol. Evol. *32*, 13–22.

Guindon, S. and Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. *52*, 696–704.

Guindon, S., Lethiec, F., Duroux, P., and Gascuel, O. (2005). PHYML Online–a web server for fast maximum likelihood-based phylogenetic inference. Nucleic Acids Res. *33*, W557–W559.

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.

Harms, M.J., Eick, G.N., Goswami, D., Colucci, J.K., Griffin, P.R., Ortlund, E.A., and Thornton, J.W. (2013). Biophysical mechanisms for large-effect mutations in the evolution of steroid hormone receptors. Proc. Natl. Acad. Sci. USA *110*, 11475–11480.

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.

Huelsenbeck, J.P. and Bollback, J.P. (2001). Empirical and hierarchical Bayesian estimation of ancestral states. Syst. Biol. *50*, 351–366.

Huelsenbeck, J.P., Ronquist, F., Nielsen, R., and Bollback, J.P. (2001). Bayesian inference of phylogeny and its impact on evolutionary biology. Science *294*, 2310–2314.

Ingles-Prieto, A., Ibarra-Molero, B., Delgado-Delgado, A., Perez-Jimenez, R., Fernandez, J.M., Gaucher, E.A., Sanchez-Ruiz, J.M., and Gavira, J.A. (2013). Conservation of protein structure over four billion years. Structure *21*, 1690–1697.

Jermann, T.M., Opitz, J.G., Stackhouse, J., and Benner, S.A. (1995). Reconstructing the evolutionary history of the artiodactyl ribonuclease superfamily. Nature *374*, 57–59.

Jones, D.T., Taylor, W.R., and Thornton, J.M. (1992). The rapid generation of mutation data matrices from protein sequences. Comput. Appl. Biosci. *8*, 275–282.

Jones, B.R., Rajaraman, A., Tannier, E., and Chauve, C. (2012). ANGES: reconstructing ANcestral GEnomeS maps. Bioinformatics *28*, 2388–2390.

Jukes, T.H. and Cantor, C.R. (1969). Evolution of protein molecules. In: Mammalian protein metabolism, H.N. Munro, ed. (New York, USA: Academic Press), pp. 21–132.

Khersonsky, O. and Tawfik, D.S. (2010). Enzyme promiscuity: a mechanistic and evolutionary perspective. Annu. Rev. Biochem. *79*, 471–505.

Kimura, M. (1981). Estimation of evolutionary distances between homologous nucleotide sequences. Proc. Natl. Acad. Sci. USA *78*, 454–458.

Koshi, J.M. and Goldstein, R.A. (1996). Probabilistic reconstruction of ancestral protein sequences. J. Mol. Evol. *42*, 313–320.

Larget, B. and Simon, D.L. (1999). Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. Mol. Biol. Evol. *16*, 750–759.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., et al. (2007). Clustal W and Clustal X version 2.0. Bioinformatics *23*, 2947–2948.

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

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, UK: Oxford University Press).

Liò, P. and Bishop, M. (2008). Modeling sequence evolution. In: Bioinformatics, J.M. Keith, ed. (Totowa, NJ, USA: Springer), pp. 255–285.

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

Maddison, W.P. (1997). Gene trees in species trees. Syst. Biol. *46*, 523–536.

Maddison, W.P. and Maddison, D.R. (2015). Mesquite: a modular system for evolutionary analysis. http://mesquiteproject.org.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., and Teller, E. (1953). Equation of state calculations by fast computing machines. J. Chem. Phys. *21*, 1087–1092.

Mirarab, S. and Warnow, T. (2015). ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. Bioinformatics *31*, i44–52.

Mirarab, S., Reaz, R., Bayzid, M.S., Zimmermann, T., Swenson, M.S., and Warnow, T. (2014). ASTRAL: genome-scale coalescent-based species tree estimation. Bioinformatics *30*, i541–548.

Mirceta, S., Signore, A.V., Burns, J.M., Cossins, A.R., Campbell, K.L., and Berenbrink, M. (2013). Evolution of mammalian diving capacity traced by myoglobin net surface charge. Science *340*, 1234192.

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

Nylander, J.A., Wilgenbusch, J.C., Warren, D.L., and Swofford, D.L. (2008). AWTY (are we there yet?): a system for graphical exploration of MCMC convergence in Bayesian phylogenetics. Bioinformatics *24*, 581–583.

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.

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

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

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

Pupko, T., Pe'er, I., Shamir, R., and Graur, D. (2000). A fast algorithm for joint reconstruction of ancestral amino acid sequences. Mol. Biol. Evol. *17*, 890–896.

Pupko, T., Doron-Faigenboim, A., Liberles, D.A., and Cannarozzi, G.M. (2007). Probabilistic models and their impact on the accuracy of reconstructed ancestral protein sequences. In: Ancestral sequence reconstruction, D.A. Liberles, ed. (Oxford, UK: Oxford University Press).

Ranwez, V. and Gascuel, O. (2001). Quartet-based phylogenetic inference: improvements and limits. Mol. Biol. Evol. *18*, 1103–1116.

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.

Risso, V.A., Gavira, J.A., Mejia-Carmona, D.F., Gaucher, E.A., and Sanchez-Ruiz, J.M. (2013). Hyperstability and substrate promiscuity in laboratory resurrections of Precambrian β-lactamases. J. Am. Chem. Soc. *135*, 2899–2902.

Risso, V.A., Gavira, J.A., and Sanchez-Ruiz, J.M. (2014). Thermostable and promiscuous Precambrian proteins. Environ. Microbiol. *16*, 1485–1489.

Risso, V.A., Manssour-Triedo, F., Delgado-Delgado, A., Arco, R., Barroso-delJesus, A., Ingles-Prieto, A., Godoy-Ruiz, R., Gavira, J.A., Gaucher, E.A., Ibarra-Molero, B., et al. (2015). Mutational studies on resurrected ancestral proteins reveal conservation of site-specific amino acid preferences throughout evolutionary history. Mol. Biol. Evol. *32*, 440–455.

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

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

Saitou, N. and Nei, M. (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. Mol. Biol. Evol. *4*, 406–425.

Sankoff, D. (1975). Minimal mutation trees of sequences. SIAM J. Appl. Math. *28*, 35–42.

Schultz, T.R., Cocroft, R.B., and Churchill, G.A. (1996). The reconstruction of ancestral character states. Evolution *50*, 504–511.

Scornavacca, C., Jacox, E., and Szöllősi, G.J. (2015). Joint amalgamation of most parsimonious reconciled gene trees. Bioinformatics *31*, 841–848.

Shimodaira, H. and Hasegawa, M. (2001). CONSEL: For assessing the confidence of phylogenetic tree selection. Bioinformatics *17*, 1246–1247.

Stackhouse, J., Presnell, S.R., McGeehan, G.M., Nambiar, K.P., and Benner, S.A. (1990). The ribonuclease from an extinct bovid ruminant. FEBS Lett. *262*, 104–106.

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

Strimmer, K. and Von Haeseler, A. (1996). Quartet puzzling: a quartet maximum-likelihood method for reconstructing tree topologies. Mol. Biol. Evol. *13*, 964–969.

Susko, E., Field, C., Blouin, C., and Roger, A.J. (2003). Estimation of rates-across-sites distributions in phylogenetic substitution models. Syst. Biol. *52*, 594–603.

Swofford, D.L. (1984). PAUP: Phylogenetic analysis using parsimony (Champaign: Illinois Natural Historical Survey).

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

Szöllősi, G.J. and Daubin, V. (2012). Modeling gene family evolution and reconciling phylogenetic discord. In: Evolutionary Genomics, A. Anismimova, ed. (New York, USA: Springer), pp. 29–51.

Szöllősi, G.J., Tannier, E., Lartillot, N., and Daubin, V. (2013). Lateral gene transfer from the dead. Syst. Biol. *62*, 386–397.

Szöllősi, G.J., Tannier, E., Daubin, V., and Boussau, B. (2015). The inference of gene trees with species trees. Syst. Biol. *64*, e42–62.

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.

Tavaré, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. Lectures Math. Life Sci. *17*, 57–86.

Thornton, J.W. (2004). Resurrecting ancient genes: experimental analysis of extinct molecules. Nat. Rev. Genet. *5*, 366–375.

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

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

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

Walker, J.C.G. (1983). Possible limits on the composition of the Archaean ocean. Nature *302*, 518–520.

Whelan, S. (2008). Inferring trees. In: Bioinformatics, J.M. Keith, ed. (Totowa, NJ, USA: Springer), pp. 287–309.
Whelan, S. and Goldman, N. (2001). A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol. Biol. Evol. *18*, 691–699.
Williams, P.D., Pollock, D.D., Blackburne, B.P., and Goldstein, R.A. (2006). Assessing the accuracy of ancestral protein reconstruction methods. PLoS Comp. Biol. *2*, e69.
Woese, C. (1998). The universal ancestor. Proc. Natl. Acad. Sci. USA *95*, 6854–6859.
Yang, Z. (1994). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. *39*, 306–314.
Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. *13*, 555–556.
Yang, Z. and Nielsen, R. (2000). Estimating synonymous and non-synonymous substitution rates under realistic evolutionary models. Mol. Biol. Evol. *17*, 32–43.
Yang, Z., Kumar, S., and Nei, M. (1995). A new method of inference of ancestral nucleotide and amino acid sequences. Genetics *141*, 1641–1650.
Yang, Z., Nielsen, R., Goldman, N., and Pedersen, A.M. (2000). Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics *155*, 431–449.
Yang, K., Heath, L.S., and Setubal, J.C. (2012). REGEN: Ancestral Genome Reconstruction for Bacteria. Genes (Basel) *3*, 423–443.
Yokoyama, S. and Radlwimmer, F.B. (2001). The molecular genetics and evolution of red and green color vision in vertebrates. Genetics *158*, 1697–1710.
Yokoyama, S., Yang, H., and Starmer, W.T. (2008). Molecular basis of spectral tuning in the red- and green-sensitive (M/LWS) pigments in vertebrates. Genetics *179*, 2037–2043.
Zeldovich, K.B., Berezovsky, I.N., and Shakhnovich, E.I. (2007). Protein and DNA sequence determinants of thermophilic adaptation. PLoS Comp. Biol. *3*, e5.
Zou, T., Risso, V.A., Gavira, J.A., Sanchez-Ruiz, J.M., and Ozkan, S.B. (2015). Evolution of conformational dynamics determines the conversion of a promiscuous generalist into a specialist enzyme. Mol. Biol. Evol. *32*, 132–143.

Bionotes

Rainer Merkl
Institute of Biophysics and Physical Biochemistry, University of Regensburg, Universitätsstrasse 31, D-93053 Regensburg, Germany,
Rainer.Merkl@ur.de

Rainer Merkl studied biomedical engineering and computer science and obtained his PhD from the University of Göttingen. He has been with the University of Regensburg since 2005, where he is now an adjunct professor of computational biology. His main research interests are protein evolution and enzyme design.

Reinhard Sterner
Institute of Biophysics and Physical Biochemistry, University of Regensburg, Universitätsstrasse 31, D-93053 Regensburg, Germany,
Reinhard.Sterner@ur.de

Reinhard Sterner studied biology and obtained his PhD from the University of Munich. He then worked as a postdoctoral fellow at the University of Basel, as a junior group leader at the University of Göttingen, and as a professor at the University of Cologne. Since 2004, he has held a chair of Biochemistry at the University of Regensburg. His main research interests are enzyme evolution and design.