Perspectives in Science (2016) 9, 17—23

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/pisc

Reconstruction of ancestral enzymes☆

Rainer Merkl, Reinhard Sterner*

University of Regensburg, Institute of Biophysics and Physical Biochemistry, 93040 Regensburg, Germany

Received 1 December 2015; accepted 2 February 2016
Available online 26 August 2016

KEYWORDS
Ancestral sequence reconstruction;
Enzyme evolution;
Phylogenetic tree;
Promiscuous enzymes;
Thermostability

Summary The amino acid sequences of primordial enzymes from extinct organisms can be determined by an *in silico* approach termed ancestral sequence reconstruction (ASR). In the first step of an ASR, a multiple sequence alignment (MSA) comprising extant homologous enzymes is being composed. On the basis of this MSA and a stochastic model of sequence evolution, a phylogenetic tree is calculated by means of a maximum likelihood approach. Finally, the sequences of the ancestral proteins at all internal nodes including the root of the tree are deduced. We present several examples of ASR and the subsequent experimental characterization of enzymes as old as four billion years. The results show that most ancestral enzymes were highly thermostable and catalytically active. Moreover, they adopted three-dimensional structures similar to those of extant enzymes. These findings suggest that sophisticated enzymes were invented at a very early stage of biological evolution.

Introduction

Modern enzymes from contemporary organisms are sophisticated biocatalysts transforming their substrates into products with high efficiency and specificity. However, as catalytic activity usually requires a certain degree of conformational flexibility, the stability of most modern enzymes is only marginal (Jaenicke, 1991). We are interested in how enzyme activity and stability developed and changed in the course of evolution. For this purpose, it would be desirable to characterize ancient enzymes from primordial organisms. The lack of macromolecular fossils, however, seems to block the access to this interesting information. Luckily, there is a circumstantial way out of this dilemma, which is the characterization of ''extinct'' proteins after their ''resurrection'' *via* ancestral sequence reconstruction (ASR).

ASR is an *in silico* approach allowing to deduce the sequences of ancient proteins from the sequences of homologous extant proteins (Liberles, 2007). The idea of ASR is more than 50 years old when Pauling and Zuckerkandl postulated that modern proteins contain enough information to derive the sequences of common ancestors (Pauling and Zuckerkandl, 1963). However, the concept of ASR could only be realized after Fitch had implemented the first phylogenetic algorithm termed PAUP (phylogenetic analysis using

* Corresponding author.
E-mail address: Reinhard.Sterner@ur.de (R. Sterner).

http://dx.doi.org/10.1016/j.pisc.2016.08.002

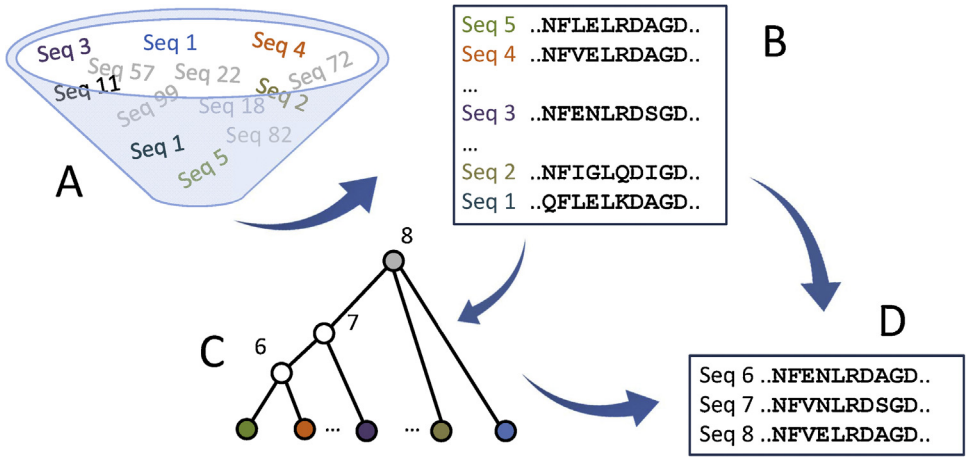

Figure 1 Protocol for ASR. Each ASR requires four steps to deduce ancestral sequences from a set of extant homologs. (A) A set of sequences is retrieved from a database. (B) These sequences are compiled to a multiple sequence alignment, which allows for the identification of mutations observed in the sequences. (C) A phylogeny is determined; the extant sequences constitute the leaves. (D) Based on this phylogenetic tree and the MSA, ancestral sequences are deduced for all internal nodes of the tree.

parsimony) (Fitch, 1971). Although phylogenetic models and algorithms have considerably improved since then, the workflow of ASR has remained essentially the same (Fig. 1). A set of homologous sequences is put together in a multiple sequence alignment (MSA), which is the basis for the subsequent calculations: first, a phylogenetic tree is constructed whose outermost nodes (the leaves) are represented by the extant sequences. After the calculation of this tree, the precursor sequences corresponding to the internal nodes are being calculated. These calculations are commonly based on a maximum likelihood approach and a phylogenetic model which allows for the sampling of mutational frequencies in a position-specific residue manner (Merkl and Sterner, 2016).

In this short review, we will first describe state-of-the art *in silico* methods that have been developed for ASR. We will then provide examples on how ASR has been used to resurrect ancient enzymes from the Precambrian era, among them translation elongation factors (Gaucher et al., 2008), thioredoxins (Ingles-Prieto et al., 2013; Perez-Jimenez et al., 2011), 3-ispropylmalate dehydrogenases (Hobbs et al., 2015; Hobbs et al., 2012), nucleotide kinases (Akanuma et al., 2013), β-lactamases (Risso et al., 2013; Zou et al., 2015), imidazole glycerol phosphate synthase (Reisinger et al., 2014), and ribonuclease H1 (Hart et al., 2014). We will conclude by summarizing the most important insights that have been gained from these studies with respect to our ability to ''replay the molecular tape of life'' (Gaucher, 2007).

In silico methods for ASR

A detailed introduction into stochastic concepts and phylogenetic models that is needed to understand modern ASR methods is beyond the scope of this review and can be found elsewhere; see (Merkl and Sterner, 2016) and references therein. Here, we present a short summary of the algorithms required to deduce phylogenetic trees and ancestral sequences.

Computing a phylogenetic tree by means of maximum likelihood

The prerequisite for the computation of a phylogenetic tree is a stochastic model that describes the probability for DNA or protein sequences to acquire mutations within a certain time interval t_i. For this purpose, a probability to *acquire any mutation* within t_i is combined with a *substitution model*. The latter explains in detail with which probability a nucleotide or amino acid residue is replaced by another one. Instead of using fixed mutation rates, it is meanwhile state of the art to sample these probabilities from a continuous distribution, which provides every site with a specific rate (Susko et al., 2003).

Based on such a model, the likelihood of a tree can be computed. Likelihood is the probability for observing the data (*i.e.*, sequences) *given* (i) the parameters of the chosen evolutionary model and (ii) the topology of the tree under study. Commonly, mutations at different sites are considered as independent events. Thus, this likelihood of a complete sequence is the product of all site-specific values. To explain the principle, it is therefore 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 all time intervals t_i and all nucleotides e_i are known for all nodes $i = 1, \ldots, 8$, the likelihood of the tree shown in Fig. 2 is:

$$L(tree) = p_{e_1e_8}(t_1)p_{e_2e_8}(t_2)p_{e_3e_7}(t_3)p_{e_4e_6}(t_4)p_{e_5e_6}(t_5) \\ p_{e_6e_7}(t_6)p_{e_7e_8}(t_7). \quad (1)$$

However, the states (nucleotides) of the internal nodes are not known and therefore it is necessary to sum over all possible states (nucleotides at internal nodes) which results in

$$L(tree) = \sum_{e_8}\sum_{e_7}\sum_{e_6} p_{e_1e_8}(t_1)p_{e_2e_8}(t_2)p_{e_3e_7}(t_3)p_{e_4e_6}(t_4) \\ p_{e_5e_6}(t_5)p_{e_6e_7}(t_6)p_{e_7e_8}(t_7)). \quad (2)$$