Available online at www.sciencedirect.com

ScienceDirect

Biochemical and Biophysical Research Communications 366 (2008) 607–611

BBRC

www.elsevier.com/locate/ybbrc

Mini Review

Characters of very ancient proteins

Bin-Guang Ma [a,1], Lei Chen [a,b,1], Hong-Fang Ji [a], Zhong-Hua Chen [a], Fu-Rong Yang [a], Ling Wang [a], Ge Qu [a], Ying-Ying Jiang [a], Cong Ji [a], Hong-Yu Zhang [a,*]

[a] *Shandong Provincial Research Center for Bioinformatic Engineering and Technique, Center for Advanced Study, Shandong University of Technology, Zibo 255049, PR China*

[b] *College of Life Sciences, Shandong Normal University, Jinan 250014, PR China*

Received 15 November 2007

Abstract

Tracing the characters of very ancient proteins represents one of the biggest challenges in the study of origin of life. Although there are no primitive protein fossils remaining, the characters of very ancient proteins can be traced by molecular fossils embedded in modern proteins. In this paper, first the prior findings in this area are outlined and then a new strategy is proposed to address the intriguing issue. It is interesting to find that various molecular fossils and different protein datasets lead to similar conclusions on the features of very ancient proteins, which can be summarized as follows: (i) the architectures of very ancient proteins belong to the following folds: P-loop containing nucleoside triphosphate hydrolases (c.37), TIM beta/alpha-barrel (c.1), NAD(P)-binding Rossmann-fold domains (c.2), Ferredoxin-like (d.58), Flavodoxin-like (c.23) and Ribonuclease H-like motif (c.55); (ii) the functions of very ancient proteins are related to the metabolisms of purine, pyrimidine, porphyrin, chlorophyll and carbohydrates; (iii) a certain part of very ancient proteins need cofactors (such as ATP, NADH or NADPH) to work normally.

Keywords: Ancient proteins; Molecular fossils; Protein architecture; Protein function; Power law; Catalytic site

Since proteins are not just the stage but also the players of life, there is continuing interest in the exploration of origin and evolution of proteins, in which tracing the characters of very ancient proteins is one of the biggest challenges, because no fossil remains of the primitive proteins can be accessed. However, considering the fact that some protein elements, such as architectures (folds) and short sequences, are rather conserved during evolution [1], they can serve as molecular fossils to help infer the features of very ancient proteins. The past few years have witnessed the preliminary success of this strategy.

By analyzing the fold occurrence patterns in 157 completely sequenced genomes (17 for archaea, 130 for bacteria, and 10 for eukaryotes), Deane and co-workers revealed that α/β is the most ancient protein class [2,3], which is supported by the inference based on architectures of amino acid synthases [4]. By a large-scale phylogenomic analysis on protein folds, Caetano-Anollés and co-workers indicated that: (i) the architectures used by the very ancient proteins belong to the following folds (from early to late): P-loop containing nucleotide triphosphate hydrolases (c.37), DNA/RNA-binding 3-helical bundle (a.4), TIM β/α-barrel (c.1), NAD(P)-binding Rossmann-fold domains (c.2), Ferredoxin-like (d.58), Flavodoxin-like (c.23) and Ribonuclease H-like motif (c.55) [5–7]; (ii) the functions of the primitive proteins are related to the metabolisms of purine, pyrimidine, porphyrin and chlorophyll, etc. (Table 1) [8]. Through searching the common amino acid sequences in 131 prokaryotic proteomes, Trifonov and co-workers found that some sequences are conserved across the proteomes and there exists a correlation between the conservation level and the age of the short sequences [9,10]. These findings strongly suggest that the most widely shared sequences descended from the common ancestor of

[*] Corresponding author. Fax: +86 533 2780271. *E-mail address:* zhanghy@sdut.edu.cn (H.-Y. Zhang).

[1] These authors contributed equally to this work.

0006-291X/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.bbrc.2007.12.014

Table 1
Characters of very ancient proteins identified by a phylogenomic analysis on protein architectures

KEGG pathway[a]	Fold[a]	Enzyme[a]/EC number[b]	Amino acid composition of catalytic site[c]	Cofactor[d]
Purine metabolism	c.37	Adenylate kinase/2.7.4.3	Lys21, Arg132, Arg138, Asp140, Asp141, Arg149	ATP, AMP
	c.37	Adenylyl-sulfate kinase/2.7.1.25	Arg421, Arg522, His425, His428	ATP
	c.37	Deoxynucleoside monophosphate kinase/2.7.1.76	Arg68	ATP
	c.37	Deoxyguanosine kinase/2.7.1.113	Glu70, Arg142	ATP, UTP
	c.37	Guanylate kinase/2.7.4.8	—	ATP
	c.37	Sulfate adenylyltransferase/2.7.7.4	Arg199, Arg292, His203, His206	ATP
	c.37	3′,5′-cyclic-nucleotide phosphodiesterase/3.1.4.17	Gln70	—
	c.37	dGTPase/3.1.5.1	Asp21, His85	dGTP
	c.37	Adenosinetriphosphatase/3.6.1.3	Lys71	ATP
	c.37	Adenylate cyclase/4.6.1.1	Arg1150	ADP, AMP
	c.37	Adenylosuccinate synthase/6.3.4.4	Asp13, His41, Gln224	GTP, IMP
Purine metabolism Pyrimidine metabolism	c.37	Deoxycytidine kinase/2.7.1.74	Glu53, Arg128	ATP, UTP
Purine metabolism Pyrimidine metabolism	c.37	DNA-directed DNA polymerase/ 2.7.7.7	Lys51, Lys141, Arg215, Arg215, Thr157	—
Pyrimidine metabolism	c.37	Cytidylate kinase/2.7.4.14	Lys19, Arg131, Arg137, Arg148, Asp139, Asp140	ATP, dCMP
	c.37	Thymidine kinase/2.7.1.21	Gly59, Glu83, Arg163, Arg222	ATP, CTP, GTP, TTP
	c.37	Uridine kinase/2.7.1.48	—	ATP
	c.37	dTMP kinase/2.7.4.9	Gly12, Glu37, Arg92	ATP, dTMP
	c.37	Nucleoside-triphosphate-adenylate kinase/2.7.4.10	Lys18, Arg126, Arg159, Arg170, Asp161, Asp162	AMP
	c.37	Cytidylate kinase/2.7.4.14	Lys16, Arg134, Arg140, Arg151, Asp142, Asp143	ATP, dCMP
Porphyrin and chlorophyll metabolism	c.37	Cob(I)yrinic acid a,c-diamide adenosyltransferase/2.5.1.17	Lys41, Thr43	ATP

[a] Data from ref. [8]. Only enzymes that participate in the three most ancient functions (*i.e.*, purine metabolism, pyrimidine metabolism and porphyrin and chlorophyll metabolism) and belong to c.37 fold are listed.
[b] Derived from MODBASE [20].
[c] Derived from Catalytic Site Atlas [13].
[d] Derived from BRENDA [22].

associated species and the more conserved the sequence is, the more ancient the motif is likely to be. It is interesting to note that in the 50 most conserved sequence motifs (octapeptides) [9,10], 22 are located in domains of c.37 fold. However, since all of these inferences eluded experimental verification, further theoretical examinations by alternative strategies appear warranted.

As above described, one can find that the prior strategies to trace the characters of very ancient proteins were based on two points: (i) using vast genomes as starting datasets; (ii) using protein architectures or conserved short sequences as molecular fossils, which inspired us to propose a new strategy by making two innovations. First, a small set of relatively early proteins [11] is used as the starting dataset. Second, the amino acid composition of catalytic sites of enzymes is used as a molecular fossil, because the catalytic sites are rather conserved during evolution [12,13]. According to the coevolutionary theory of genetic code [14], the primitive protein life mainly consisted of naturally occurring small amino acids, from which the large amino acids were derived later through biosynthesis. This theory provides a frame of reference to identify the late-occurring proteins by molecular fossil of amino acid composition, that is, if an enzyme uses late originated amino acids (i.e., His, Phe, Cys, Met, Tyr, and Trp [15,16]) to do catalysis, it is very likely to appear late. This tactic is preliminarily supported by the finding that the catalytic sites of very ancient enzymes revealed by Caetano-Anollés et al. are mainly composed of early amino acids (∼90%) (Table 1) [8]. The new data set and the alternative molecular fossil constitute the core of the new strategy to infer the characters of very ancient proteins. It is intriguing to investigate whether the new strategy reaches conclusions consistent with the previous ones.

The set of relatively early proteins is derived from yeast proteome. Recently, yeast proteins have been classified into five age groups, according to the occurring patterns of their orthologs in other species [11]. The oldest age group, consisting of 1806 members, includes yeast proteins that can be traced back to eubacterial genomes. The amino acid sequences corresponding to these proteins were downloaded from SGD database [17]. Then, they were assigned