Biochimie xxx (2014) 1–6

Contents lists available at ScienceDirect

Biochimie

journal homepage: www.elsevier.com/locate/biochi

Review

Two fundamental questions about protein evolution

David Penny*, Bojian Zhong [1]

Institute of Fundamental Sciences, Massey University, Palmerston North, New Zealand

ARTICLE INFO

Article history:
Received 23 July 2014
Accepted 15 October 2014
Available online xxx

Keywords:
Markov models
Phylogenetics
Origin of proteins
Eigen limit

ABSTRACT

Two basic questions are considered that approach protein evolution from different directions; the problems arising from using Markov models for the deeper divergences, and then the origin of proteins themselves. The real problem for the first question (going backwards in time) is that at deeper phylogenies the Markov models of sequence evolution must lose information exponentially at deeper divergences, and several testable methods are suggested that should help resolve these deeper divergences. For the second question (coming forwards in time) a problem is that most models for the origin of protein synthesis do not give a role for the very earliest stages of the process. From our knowledge of the importance of replication accuracy in limiting the length of a coding molecule, a testable hypothesis is proposed. The length of the code, the code itself, and tRNAs would all have prior roles in increasing the accuracy of RNA replication; thus proteins would have been formed only after the tRNAs and the length of the triplet code are already formed. Both questions lead to testable predictions.

1. Introduction

Two questions about proteins are discussed here, and they approach evolutionary aspects of proteins from different directions. The first question is more biological in its nature and goes backwards in time (a 'top down' approach) and discusses the accuracy of trees from protein/genes for deeper phylogeny. The second is a more chemical approach, and comes forwards in time (the 'bottom up' approach) and considers the origin of proteins themselves. This 'top down' and 'bottom up' approach idea has been used previously [1].

For the first question, it has been shown mathematically by Mossel and Steel [2] that the Markov models used with sequence data lose information at deeper divergences; and that the falloff is exponential at deeper times. They do show that the situation is improved linearly with increasing number of sequences. Thus having more sequences will help, but with an exponential decay at deeper times, versus a linear increase with the number of sequences, there is no doubt that the exponential decay with time will eventually be the dominant factor. There has been considerable interest in this issue for many years on down-weighting some characters/sites (see e.g. Refs. [3,4]), but in general these do not deal directly with the issues raised by Mossel and Steel [2]. The issues of deeper phylogeny are crucial to understand many questions, including the deeper divergences among eukaryotes. With this question, people are generally convinced about 5 or 6 main eukaryote groups [5], but the order of branching among these deeper groups is not well established. This is a problem that needs to be resolved. Despite this concern about the deep phylogeny of eukaryotes, many people appear to accept even older divergences for Bacteria and Archaea. However, there are several approaches for increasing the accuracy for deeper divergences that are suggested here.

For the second question, the origins of proteins themselves, there appears to be an impasse of good testable models. Both Crick [6] and Orgel [7] raised the question early about the origin of the code. The RNA-world scenario was produced several decades ago, and there has been good progress on our understanding of the continuum of processes it entails. However, we need a genuine evolutionary theory of the origin of both proteins and the genetic code, and one that does not 'think ahead'. It is estimated that there are about 10^{84} possible codes (see Ref. [8]) – but we are not interested here in 'which code'. We suggest that for a later stage that proto-tRNAs could be used to increase the accuracy of ribonucleotide addition – this gives testable predictions that there was a longer time (still maybe only milliseconds) for the potential recognition and checking for the addition of a ribonucleotide, and therefore (potentially) a higher accuracy. Next we suggest that triplets of ribonucleotides were added, and a recent review of the origin of proteins [9] refers to this as a 'replicase' model. This would

* Corresponding author.
E-mail address: d.penny@massey.ac.nz (D. Penny).
[1] Present address: College of Life Sciences, Nanjing Normal University, Nanjing, China.

http://dx.doi.org/10.1016/j.biochi.2014.10.020
0300-9084/

increase the length of the RNA that could be produced before going into Eigen's 'error catastrophe' (see later). As a later step, some copies of the proto-ribosome could then start synthesising short peptides, and so on. These are the two questions analysed.

2. Increasing the resolution of deep divergences

An important point is that there is no claim [2] that all information about deeper divergences is lost from the proteins. It is just that the Markov models we use for reconstructing the relationships of protein sequences must eventually lose power, and are also susceptible to deviations from the basic assumptions - such as equality of nucleotide composition (see Ref. [10]). There is however information left, for example, in the three-dimensional structure of proteins – proteins can still retain information about deeper divergences [11]. The question is to be able to use such information with reasonable confidence.

Several suggestions are considered.

Firstly, we expect a Gamma distribution of rates to retain information longer.
Secondly, if there is bimodal distribution of rates, then eliminate the faster sites.
Thirdly, inferring ancestral sequences appears robust and should help estimate deeper divergences.
Fourthly, inferred three-dimensional structures (3-D) should retain information longer.
Fifthly, only taking the sequence crossing the central 3D region might help.
Sixthly, weighting the partitions that are consistent could also help.
Seventhly, gene order information might be useful.

This list is by no means exhaustive – there are other possibilities, but these examples illustrate the principles.

2.1 & 2.2 The first two suggestions (a gamma distribution of rates across sites, and a bimodal distribution of rates) will be considered together. The approach of Goremykin et al. [12] of identifying the faster evolving sites, and sequentially eliminating them, is predicted to be more robust for somewhat deeper divergences. In this approach, the faster evolving sites are identified by comparing each site with every other site, and this does not involve a preliminary tree being inferred. This process allows all aligned sites to be ranked in order of their apparent rate of evolution, and the faster sites can be sequentially eliminated until a stopping point is found based on correlations between distances. The analysis includes all the slower evolving sites, and it does appear to be more robust tin resolving deeper divergences [12]. The method relies on being able to successfully detect the faster evolving sites, and it has been shown that these faster evolving sites are not fitting the model very well [13]. However, the current approach has primarily been tested on the divergence of flowering plants [12], and although it has also been evaluated for the origin of land plants [14] it does need to be evaluated on say, the relationships between the 5 or 6 main groups of eukaryotes [5]. In this case simulations are a good test of the approach.

2.3 For the third suggestion, using Ancestral Sequence Reconstruction (ASR), there appears yet to be limited evidence, but that limited evidence is encouraging. There are two types of evidence discussed here. The first is that of Collins [15] who found that by using inferred ancestral sequences from plant and animal proteins associated with the spliceosome, they could identify related proteins in the more distant eukaryote *Giardia*. Without ASR, the BLAST search found no convincing hits. Even more striking, Finnigan et al. [16] have shown that they can reconstruct synthetic proteins having the inferred ancestral sequence, and that these proteins do have the expected metabolic properties – they call it Ancestral Sequence Resurrection, also ASR! As yet we do not appear to have a good method that uses this approach for estimating deep relationships, However, Daly et al. [17] have used traditional ASR to demonstrate that the major vault protein appears to be in all major lineages of eukaryotes – and therefore almost certainly was in LECA – the Last Eukaryote Common Ancestor. This method is promising, but its utility is not yet demonstrated, and the approach might yet be amenable to more formal proofs.

2.4 Fourthly, inferred three-dimensional structures should retain information longer. Illegard et al. [11] suggest that 3D structures retain information about 3–10 times longer than primary sequences, thus there should be useful information in tertiary structures. The main problem here is that we currently have little idea on how to use this structural information for resolving deeper phylogeny. Additionally, gene content will be important [18] but we will have to take into account that different groups may have quite different strategies in relation to their requirements; for example, many eukaryotes appear currently to rely on prokaryotes to synthesis some amino acids and some cofactors (vitamins). Although we think that there must be good information in tertiary structures, we do need additional information on how best to use it. One suggestion is the next alternative.

2.5 Fifthly, only taking the sequence crossing the central 3-D region might help; this is thought to be the most conserved part of the protein. For example, Wang et al. [19] report that an interesting possibility arises because it appears that eukaryotes have longer linker regions and Light et al. [20] show that the expansions/contractions occur in 'intrinsically disordered' regions, that is, not in the core part of the structure. The net result is that in attempting to align proteins, when there are differences in length of the different regions of the proteins, these differences could lead to errors in alignment – and consequently potential errors with phylogenetic analysis. The interest here is whether only aligning and using the central part of the protein (possibly more conserved) would give more accurate alignment and help to resolve the deeper phylogeny. In a similar manner, Daly et al. ([17], Fig. 9) have also used structural information to identify sites where a new amino acid affects the central core structure, versus amino acids that jut out into the outside (or inside) portion of the protein. There is the possibility that understanding how amino acid changes affect the central core structure might be especially helpful and there is also the use of structural (3-D) information to help with alignments.

2.6 Sixthly, weighting the partitions that are consistent (or inconsistent) with other sites might help. This is illustrated in Fig. 1A with 9 taxa (the rows) and twelve 2-state characters (the columns). The principles are the same for multi-state characters (4 for nucleotides, 20 for amino acids), and pairs of characters are 'compatible' if the character states can be joined with the minimum number of changes for a pair of characters; otherwise they are

Please cite this article in press as: D. Penny, B. Zhong, Two fundamental questions about protein evolution, Biochimie (2014), http://dx.doi.org/10.1016/j.biochi.2014.10.020