Biochimie xxx (2015) 1–19

Contents lists available at ScienceDirect

Biochimie

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

Review

The phylogenomics of protein structures: The backstory

Charles G. Kurland [a, *], Ajith Harish [b]

[a] Microbial Ecology, Department of Biology, Lund University, Ecology Building (Sölvegatan 37), SE-223 62 Lund, Sweden
[b] Structure and Molecular Biology, Department of Cell and Molecular Biology, Biomedical Center, Uppsala University, 751 24 Uppsala, Sweden

ARTICLE INFO

Article history:
Received 11 January 2015
Accepted 28 July 2015
Available online xxx

Keywords:
Phylogeny
Tree of life
Horizontal gene transfer
Protein superfamily
Autogenous evolution of eukaryotes

ABSTRACT

In this introductory retrospective, evolution as viewed through gene trees is inspected through a lens compounded from its founding operational assumptions. The four assumptions of the gene tree culture that are singularly important to evolutionary interpretations are: a. that protein-coding sequences are molecular fossils; b. that gene trees are equivalent to species trees; c. that the tree of life is assumed to be rooted in a simple akaryote cell implying that akaryotes are primitive, and d. that the notion that all or most incongruities between alignment-based gene trees are due to horizontal gene transfer (HGT), which includes the endosymbiotic models postulated for the origins of eukaryotes. What has been unusual about these particular assumptions is that though each was taken on board explicitly, they are defended in the face of factual challenge by a stolid disregard for the conflicting observations. The factual challenges to the mainstream gene tree-inspired evolutionary view are numerous and most convincingly summarized as: Genome trees tell a very different story. Phylogeny inferred from genomic assortments of homologous protein structural-domains does not support any one of the four principle evolutionary interpretations of gene trees: a. 3D protein domain structures are the molecular fossils of evolution, while coding sequences are transients; b. Species trees are very different from gene trees; c. The ToL is rooted in a surprisingly complex universal common ancestor (UCA) that is distinct from any specific modern descendant and d. HGT including endosymbiosis is a negligible player in genome evolution from UCA to the present.

Contents

* Corresponding author.
E-mail address: cgkurland@yahoo.se (C.G. Kurland).

http://dx.doi.org/10.1016/j.biochi.2015.07.027
0300-9084/

1. Apologia

Sequence-based gene trees were invented in the 1960s to provide a radically new window on evolution: a molecular window on protein evolution [1]. Initially, Zuckerkandl and Pauling [1] described mutational changes in amino acid sequences of proteins plotted on a hypothetical evolutionary time scale. These first distance trees for gene families were followed by more sophisticated protein and rRNA sequence-based gene trees [2]. For a while, the rRNA trees dominated discussions of phylogeny because they were advertised as more faithful reflections of species evolution than proteins. More recently, concatenated protein gene trees have replaced rRNA trees as the standard bearers of genotypic phylogeny [3,4].

Also in the 1960s, Hennig published his immensely important treatise on Phylogenetic Systematics [5], which has remained an invaluable guide to the study of phylogeny and systematics. Contemporary molecular phylogeny is cladistic to the extent that it aspires to be based on the identification of the common ancestors of homologous descendants. In that case, the phylogenies may be based on molecular genotypes or molecular phenotypes, and mixed phylogenies of genotypic together with phenotypic characters are also feasible [6].

Two bioinformatic initiatives emerged in the 1990s that refocused attention on phenotypic molecular phylogeny. First, thousands of genomes from all sorts of organisms were fully sequenced [7,8]. Second, novel developments in structural biology and informatics facilitated the annotation of genome sequences by identifying the protein domains they encode [9,10]. The key idea that sparked the development of phylogeny based on genome contents of protein domains was the realization that the 3D structures of protein domains are homologous characters [7–10]. Such characters enable the construction of robust molecular phylogeny based on the 3D protein structures. Such 3D structural homologs enable genome-scale phylogeny that is in our view an excellent approximation to a molecular species tree since it is nearly all encompassing with respect to the protein coding sequences of organisms as well as a reconstruction based on homologous phenotypic characters [7–11].

The following is the backstory of molecular phylogeny since the 1960s. The main concerns are the contrasting features of gene phylogenies and genome phylogenies, both of which are ultimately based on sequence analysis. The data suggest that while gene trees may be irreplaceable for the study of microevolution of living populations, genome content trees based on 3D structures of compact protein domains are demonstrably superior instruments for the exploration of species trees and deep phylogeny.

The evolutionary story told by genome content phylogeny based on 3D structural characters, the compact domains of proteins, is decidedly different from mainstream scenarios that have been elaborated with the aid of gene trees [2,4,11,12].

2. Introduction

The Stockholm meeting on “Protein structure, protein evolution” was held June 2–6, 2014. It featured in-depth presentations about protein folding, bioinformatics, genome content-based phylogeny, proteolytic editing as well as the evolution of novel proteins. The barriers to retrieving reliable information from deep time were a common subtext for many of the talks.

Cosmology shares with molecular evolution a dependence on mathematical reconstructions to recover and describe phenomena buried in deep time. Both sorts of searches are vulnerable to uncertainties in the boundary conditions that steer their respective extrapolations. These similarities make it difficult to understand why, unlike those interested in molecular evolution, the cosmologists respond so well to challenges arising from the discovery of potential artifacts that may trouble their searches.

For example, in March of 2014 a US team, BICEP reported that a particle pattern in the sky over Antarctica may have signaled the presence of remnants of the rapid expansion of space (cosmic inflation) that occurred just fractions of a second after the celebrated “Big Bang” [13]. By September of the same year a relevant new background analysis from the European Space Agency's Planck satellite was released. It supported the suspicions that the BICEP group may have underestimated the extent of contamination by dust in our own galaxy, which might account for BICEP's apparent sightings of “cosmic inflation” [14,15]. The upshot is that within six months of the initial report from BICEP, collaboration was initiated with the Planck group to assess the impact of the dust “background” on the calculations that tentatively identified cosmic inflation [14,15]. A joint assessment of the impact of that newly reported dust background on the recent cosmic inflation calculations has concluded that the original BICEP study was indeed in error [14,15].

That enviably straightforward example of constructive confrontation between adversarial groups is a model for optimal research efficiency. In contrast, the decades long attempt by mainstream molecular evolutionists to reconstruct a tree of life (ToL) has been virtually impervious to corrective measures [12,16]. It would not be an exaggeration to characterize much of the molecular evolutionary enterprise since the 1990s as a study in missed opportunities.

There are understandable reasons for this predicament. One is the natural preoccupation of molecular biologists with the collinear sequences that make up the DNA, RNA, and protein trichotomy of genetic information. This preoccupation had the unfortunate consequence that it obscured a century old, biological tradition to exploit homologous structural characters to study evolution. Closer to home, molecular evolutionists seemed little interested in the recurrent 3D protein structures that present themselves as homologous evolutionary characters at the molecular level [16]. Ironically, it was precisely the availability of such homologous characters at the molecular level that was emphasized by publications in the early 1990s that offered user-friendly public databases of homologous 3D structures identified at atomic resolution [7,8,17].

In a better world, such protein structure databases might have supported the re-evaluation of global phylogeny that was on Mayr's [12] agenda in his 1998 exchange with Woese [16]. However, Woese deemed that re-evaluation impossible because, as he insisted, there were no homologous structural (phenotypic) characters on the molecular level that were suitable for phylogenetic studies of microorganisms [16]. The issues at stake were straightforward: Mayr and earlier others [12,18–20] reckoned that any natural taxonomy would have to identify Archaea and Bacteria as sister clades in a monophyletic taxon because of their very similar cellular phenotypes as well as their shared divergence from the phenotypes of Eukaryote cells.

Woese [16] on the other hand preferred the contrary results of gene tree reconstructions with rRNA and a few proteins: His normative, chimeric tree consisted of the unrooted sequence-based gene tree of rRNA to which had been added the Dayhoff rooting of several pairs of paralogous proteins [2,21–27]. In this chimeric tree the Archaea are sister clades of Eukaryotes, and rooted in Bacteria [27]. In other words, two phenotypically similar taxa, Archaea and Bacteria are separated in a paraphyletic arrangement with Archaea and Eukaryotes as sister clades diverging from a Bacterial root. The latter, arrangement made no sense to a “biologist”, as Mayr insisted [12].

Please cite this article in press as: C.G. Kurland, A. Harish, The phylogenomics of protein structures: The backstory, Biochimie (2015), http://dx.doi.org/10.1016/j.biochi.2015.07.027