Biochimie xxx (2015) 1–9

Contents lists available at ScienceDirect

Biochimie

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

Research paper

Function-selective domain architecture plasticity potentials in eukaryotic genome evolution

Viktorija Linkeviciute [a, b], Owen J.L. Rackham [a, c], Julian Gough [a], Matt E. Oates [a], Hai Fang [a, *]

[a] Computational Genomics Group, Department of Computer Science, University of Bristol, The Merchant Venturers Building, Bristol BS8 1UB, UK
[b] School of Biological Sciences, University of Edinburgh, Darwin Building, The King's Buildings, Edinburgh EH9 3BF, UK
[c] Centre for Computational Biology, Duke-NUS Graduate Medical School, Singapore 169857, Singapore

ARTICLE INFO

Article history:
Received 31 October 2014
Accepted 6 May 2015
Available online xxx

Keywords:
Domain architectures
Eukaryotic genomes
Evolution
Function

ABSTRACT

To help evaluate how protein function impacts on genome evolution, we introduce a new concept of 'architecture plasticity potential' – the capacity to form distinct domain architectures – both for an individual domain, or more generally for a set of domains grouped by shared function. We devise a scoring metric to measure the plasticity potential for these sets, and evaluate how function has changed over time for different species. Applying this metric to a phylogenetic tree of eukaryotic genomes, we find that the involvement of each function is not random but highly selective. For certain lineages there is strong bias for evolution to involve domains related to certain functions. In general eukaryotic genomes, particularly animals, expand complex functional activities such as signalling and regulation, but at the cost of reducing metabolic processes. We also observe differential evolution of transcriptional regulation and a unique evolutionary role of channel regulators; crucially this is only observable in terms of the architecture plasticity potential. Our findings provide a new layer of information to understand the significance of function in eukaryotic genome evolution. A web search tool, available at http://supfam.org/Pevo, offers a wide spectrum of options for exploring functional importance in eukaryotic genome evolution.

1. Introduction

1.1. The importance of protein-domain architectures in understanding genome evolution

Elucidating the importance of function in directing eukaryotic evolution is vital to explain the phenotypic diversity of observed living forms. We present a first attempt towards a systematic description and comparison of gene function over the evolution of eukaryotic proteomes. A proteome is an entire protein repertoire (encoded by a genome), and is composed of proteins comprised of structural units or domains [1]. For simplicity, hereinafter the words 'genome' and 'proteome' are used interchangeably (e.g. protein domain assignments for a genome actually means assignments for the proteome encoded by the genome). Also, the referred to in this work are those with well-defined 3D structure, although other types of domains and their functional importance have been described elsewhere [2–6]. As building blocks, domains are either found alone or combined to create multi-domain proteins. It is generally accepted that domains often act as functional units [7,8] creating a basis for the complete functional repertoire for proteins. This modularity of proteins is likely favoured by evolution because it allows for combining pre-existing domains to acquire new functions [9]. The sequential order of the domains that make up a protein is referred to as its domain architecture (or 'architecture' in brief). Our previous analysis has shown that most extant architectures evolve from ancient architectures, and convergent/polyphyletic evolution of architectures resulting in the same architectures in eukaryotic species of different lineages is rare [10]. Furthermore, studies by others show that the evolutionary changes to architectures are more common by domain insertions than deletions, and the insertion of domains is preferred at the terminus

Abbreviations: FDR, false discovery rate; GO, Gene Ontology; HMM, hidden Markov models; PP, plasticity potential; SCOP, Structural Classification of Proteins; sTOL, species tree of life.

[*] Corresponding author. Tel.: +44 (0) 7582289050.
E-mail address: hfang@cs.bris.ac.uk (H. Fang).

http://dx.doi.org/10.1016/j.biochi.2015.05.003
0300-9084/© 2015 Published by Elsevier B.V.

Please cite this article in press as: V. Linkeviciute, et al., Function-selective domain architecture plasticity potentials in eukaryotic genome evolution, Biochimie (2015), http://dx.doi.org/10.1016/j.biochi.2015.05.003

over internally [11]. Proteins with the same or similar architectures tend to be homologous and functionally similar [12,13]. The emergence of new architectures is thought to be a major mechanism of new functionality [14,15]. Our recent study on the evolution of human cells suggests that the exaptation (opposed to adaptation) of existing architectures is probably a major source of cell types [16]. So far, studying domain architectures at the genome scale (both in extant and ancestral genomes) is the most realistic approach to comprehensively understand the evolutionary forces shaping eukaryotic genomes.

1.2. Genomic content of protein domain architectures in eukaryotic genomes

The Structural Classification of Proteins (SCOP) database is a gold standard for classifying protein domains of known structure [17]. According to SCOP, a domain superfamily is defined to group together domains for which there is structural, sequence and functional evidence for a common ancestor. Hereinafter, 'domain superfamilies' and 'superfamilies' are used interchangeably. Using this definition of domains, the SUPERFAMILY database builds hidden Markov models (HMMs) for assigning domain compositions for genome sequences [18]. It provides the most comprehensive assignment of SCOP domain architectures to publicly available genome sequences [19], including those in eukaryotic genomes and their ancestral architectures reconstructed from the eukaryotic species tree of life (sTOL) [20]. Fig. 1 illustrates the status quo for eukaryotic genome information in the SUPERFAMILY database. Across genomes there is a remarkably similar number of superfamilies but a much higher variation for the number of architectures. On average, the number of proteins (with domains) is higher in plant genomes than in animal genomes, but the reverse is true for architectures (Fig. 1A). When plotting the architecture number against superfamily number for each of eukaryotic genomes (Fig. 1B), it becomes clear that it is the repertories of domain architectures that more closely correlate with organism complexity than protein domains. As superfamilies increase in number, architectures undergo an exponential increase, indicating that the emergence of architectures (rather than superfamilies) contributes to the organism complexity. Still, there exists the 'G-value paradox' (the gene/protein number is not expectedly related to the complexity [21]), even in terms of architecture number.

1.3. Concept of protein domain architecture plasticity potential

To better describe the relationship between genomes, superfamilies and architectures, we introduce the concept of *'architecture plasticity potential'*, the capacity of a domain superfamily to occur in different architectural contexts (i.e., the number of different architectures) within a genome. From this concept, architecture plasticity potential differs from one superfamily to another. The upper panel of Fig. 1C illustrates architecture plasticity potential for superfamilies across extant eukaryotic genomes. For an extant genome, most superfamilies occur only in a small number of architectures, but with a few superfamilies present in many architectures. This power-law-like pattern is similar to the previous report for domain combinations [22] and for domain architectures [23], suggesting that architecture plasticity potential is likely an intrinsic property of superfamilies (i.e. superfamily-specific). This superfamily-specific potential also differs between genomes. For a given superfamily, in general animal genomes have a higher degree of architecture diversity than plant and fungi genomes, and this potential is evolvable in a highly lineage-specific manner (the lower panel of Fig. 1C). Notably, our concept of 'architecture plasticity' looks similar to but is different from the previous concepts such as 'domain versatility' [24] and 'domain promiscuity' [25,26]. The architecture plasticity is closely related to the (unique) architectural design of the proteins, while the domain versatility/promiscuity is much related to the combinatory nature of domains (observed within domain architectures).

1.4. Opportunity for studying functional significance in eukaryotic genome evolution

Based on preliminary data present in Fig. 1 and the concept of architecture plasticity potential introduced in Section 1.3, we intend to examine dynamic changes of architecture diversity during eukaryotic evolution, not only for an individual superfamily, but also for a collection of superfamilies, for instance those sharing a certain biological property (especially function). A somewhat overlooked area of research is the need for functional annotations of protein domains (even though their importance as functional units has been widely recognised). Recently, we released the dcGO database [8], together with open-source software 'dcGOR' [27], providing a systematic annotation of domains using a panel of ontologies including Gene Ontology (GO) and expanding our sparse manual functional annotations [28]. This resource has been assessed in the CAFA competition [29,30], and has been effectively utilised for cross-knowledge and cross-species studies [31]. As well as defining architecture plasticity potentials for individual superfamilies, we also generalize the definition to describe a collection of functionally related superfamilies (e.g., annotated by a GO term in the dcGO database). As such we are able to address the question of how functional information carried by protein domains influences the architectural diversity over the course of eukaryotic genome evolution.

2. Materials and methods

2.1. Genomic domain assignments and architectures

Domain assignments for sequenced genomes were obtained from the SUPERFAMILY database [32], a routinely updated resource that was initially developed for structural genomics analysis [18] but now has been extended to phylogenomics analysis [20]. At the time of writing (September 2014) SUPERFAMILY contains 437 eukaryotic proteomes and 1674 superfamilies (defined by SCOP [17] at the superfamily/evolutionary level with an evidence for a common ancestor). Each proteome is annotated using HMMs based on these superfamilies and subsequently each protein sequence is converted into a sequence of SCOP superfamily domains or gaps, i.e. the protein's domain architecture. Here we are interested in, given a genome, the potential of a superfamily to be present in different architectures. Thus we prepared a matrix of 1674 superfamilies × 437 genomes, wherein each element corresponds to the number of different architectures associated with a superfamily (in a row) that is present in a genome (in a column). This matrix is comprised of domain architectures for extant eukaryotic genomes.

2.2. Ancestral genomic architectures in eukaryotic evolution

Recently, we published the sTOL [20], a tree of (sequenced) life that provides an evolutionary context for genome-wide studies. The sTOL is a fully resolved binary tree, with each internal node either being mapped onto a known ancestral species or left unlabelled as a hypothetical unknown ancestor. Since the convergent evolution of domain architectures is rare, particularly in eukaryotes [10], we have applied Dollo parsimony [33] to reconstruct ancestral states of domain architectures for ancestral genomes in the

Please cite this article in press as: V. Linkeviciute, et al., Function-selective domain architecture plasticity potentials in eukaryotic genome evolution, Biochimie (2015), http://dx.doi.org/10.1016/j.biochi.2015.05.003