Molecular evolution

The architectural design of networks of protein domain architectures

Chia-Hsin Hsu\(^{1,4,2}\), Chien-Kuo Chen\(^2\) and Ming-Jing Hwang\(^{1,4,3}\)

\(^1\)Bioinformatics Program, Taiwan International Graduate Program, \(^2\)Institute of Molecular Biology, and 
\(^3\)Institute of Biomedical Sciences, Academia Sinica, Taipei, Taiwan, Republic of China 
\(^4\)Institute of Biomedical Informatics, National Yang-Ming University, Taipei, Taiwan, Republic of China

Protein domain architectures (PDAs), in which single domains are linked to form multiple-domain proteins, are a major molecular form used by evolution for the diversification of protein functions. However, the design principles of PDAs remain largely uninvestigated. In this study, we constructed networks to connect domain architectures that had grown out from the same single domain for every single domain in the Pfam-A database and found that there are three main distinctive types of these networks, which suggests that evolution can exploit PDAs in three different ways. Further analysis showed that these three different types of PDA networks are each adopted by different types of protein domains, although many networks exhibit the characteristics of more than one of the three types. Our results shed light on nature’s blueprint for protein architecture and provide a framework for understanding architectural design from a network perspective.

1. Introduction

Most proteins are made up of more than one module of folding, function and evolution, known as domains. The sequential order in which domains are linked together in a protein is referred to as protein domain architecture (PDA; [1]). The deluge of genomic sequence data in recent years has allowed genome-scale comparative analyses to uncover a number of interesting observations about PDAs. For example, we now know that (i) PDAs, even of only two domains, are highly conserved, meaning that they originated from an ancestral architecture which was created once, then diversified through mechanisms such as domain duplication, fusion and insertion/deletion [2–4], although cases involving multiple origins seem to occur more often than previously thought [5]; (ii) gains and losses of domain in non-repeat PDAs tend to occur at a terminal rather than in internal part of the architecture [6–7]; (iii) despite the exponential growth of sequence data, few new single domain architectures are being discovered, whereas novel and species-specific multiple-domain architectures are continuing to be identified, leading to the notion that PDA is a major means by which protein functional and species diversity evolves [8] and (iv) some domains, such as many of those involved in signal transduction, have a penchant to promiscuously combine with other domain families to form many different PDAs [9]. These findings underscore the wealth of knowledge that we are just beginning to glean from PDA analysis, especially as it relates to protein evolution. Owing to the very large number of PDAs (more than 70 000 and growing rapidly), within which various numbers of more than 10 000 (growing slowly) single domain families (SDFs) are sequentially linked, PDAs are especially suitable for network-based analysis. In this work, we constructed a type of PDA network (PDAN) for every SDF catalogued in the Pfam-A database [10] to investigate the mechanisms of PDA growth from a network perspective.
Figure 1. The three main types of PDAN: star, tail and tetragon. Examples of the three main types of PDAN identified in this study: (a) star, (b) tail and (c) tetragon. The parent SDF (indicated by an arrow) from which the network was constructed is indicated at the bottom of each figure, with its Pfam ID in parenthesis. Nodes (i.e. PDAs) found in species of the three superkingdoms were given a proportionally thicker width, with those connecting the parent SDF and its repeats coloured black and the rest grey. The size of the circle gives a rough idea of the node’s degree, or the number of its connections.

2. Results and discussion

We retrieved the PDAs, along with their species information, that are annotated and assigned to each of the 11 912 SDFs in the Pfam-A database (v. 24) for the proteins collected in UniprotKB (v. 57.6). All of the protein domains, including split and partial domains, were handled according to these PDA annotations. For each SDF, we constructed its PDAN by connecting any two of its PDAs that differed by only one domain, which usually resulted in a large connected network, a few much smaller networks and a number of singletons (more details in the electronic supplementary material, figure S1). We focused our analysis on 656 large PDANs, one for each SDF, in order to observe discernible network architectures.

(a) Three types of network architecture

Inspection of these 656 PDANs revealed three distinct types of network architecture, which, from their overall appearance, will be referred to as ‘star’, ‘tail’ and ‘tetragon’ (figure 1). A typical star network (figure 1a) originates from the parent SDF located at the network centre, from which the network proliferates as a result of the parent SDF being linked to many different domains to form two-domain architectures, only a portion of which are then expanded into three-domain architectures, with increasingly longer architectures becoming increasingly less frequent. By contrast, a tail PDAN is one in which significant expansion from the parent SDF occurs in only one direction—by repeated addition of the same parent SDF; furthermore, along the path of the repeat, or the tail’s backbone, other domains are added, such that each node (i.e. PDA) of the repeat may resemble the centre of a burgeoning star, the size of which usually decreases as one moves down the backbone (figure 1b). For tetragon PDANs, named because they contain many tetragons (figure 1c), every node in the network other than the terminal ones is fairly equal, in that they are connected to a similar number of adjacent nodes, and this number is usually small and thus, unlike star and tail PDANs, the parent SDF of a tetragon PDAN does not seem to play a leading role in the evolution of the network. The distribution of the three main PDAN characteristics and the criteria used to quantify them are shown in figure 2. Note that while exclusively tail PDANs were few in number (14 versus 229 for star and 166 for tetragon, figure 2g), ‘tail’ was given a type to itself because repeats are a distinctive type of protein domain [11] and they play an important role in protein evolution and exhibit distinct duplication patterns [12,13].

(b) Superkingdom occurrence and network expansion

As apparent from figure 1 and the electronic supplementary material, figures S2–S4, incorporating species information showed that most PDAs (87%, 63 210/72 629) were present in just one superkingdom (Archaea, Bacteria or Eukarya) and the two end nodes (PDAs) of most network edges (93%, 108 167/115 824) were present in no more than one common superkingdom. A similar result, though to a lesser extent (74%, 85 982/115 824), was obtained using species instead of superkingdom presence data (see the electronic supplementary material, table S1). However, for most (86%, 566/656) of the PDANs analysed, their parent SDF was present in the most ancient superkingdom of all of the network’s nodes, and nodes immediately connecting the parent SDF or forming the backbone of the network were usually more ancient than peripheral nodes (see figure 1...
Figure 2. The distribution and a score for the three main PDANs and their PDAs’ length and number of unique domains. The distribution of network size versus the three parameters used in this work to characterize the three main types of PDAN: largest degree for (a) star, (b) repeat length for tail and (c) ratio of tetragon nodes to network size for tetragon. (d–f) are the cumulative percentage of the 656 PDANs analysed that exhibited the three characteristics, respectively, at varying thresholds. Using the cumulative percentage as score, every one of the 656 PDANs can be assigned an S score (for star), an R score (for tail) and a T score (for tetragon), as illustrated in the three examples shown in the electronic supplementary material, figure S4 and also in the data shown in table S4. The score, a value between 0 and 1, indicates the extent of its corresponding characteristic exhibited by a network as compared with other networks within the set of 656 PDANs. The thresholds used to generate the Venn diagram for Pfam-A data and for InterPro data (see the electronic supplementary material, figure S5) are indicated. (g) The threshold of 30 nodes used to select the 656 PDANs is also indicated. The choices of these thresholds are admittedly arbitrary, but were made to find PDANs with a strong presence of star, tail or tetragon characteristics. As indicated by these distribution data, and also by the data shown in the electronic supplementary material, figure S4 and table S4, most PDANs had a substantial score (i.e. non-negligible presence) from all the three characteristics. (h) The distribution of PDA length (number of domains in a PDA) for the three types of PDANs, showing that tetragon networks usually have longer PDAs than star networks. (i) The distribution of the number of unique domains used by the three types of PDAN, showing that, for the same network size, tetragon networks usually use a smaller number of unique domains than star networks. In the analysis for (b) and (i), only networks with exclusively star, tail or tetragon characteristics as defined by the thresholds of (d–f) were used. The number of exclusively tail networks (14) was not sufficient to observe a clear distribution tendency. (Online version in colour.)

and electronic supplementary material, figures S2–S4 for examples and table S2 for data), the exceptions (14%, 90/656) being those with increased tetragon characteristics (see the electronic supplementary material, table S3). These results corroborate the above suggestions of how PDANs might evolve that were made solely from observing network topologies, although more species data and a deeper phylogenetic analysis are required to further elucidate the evolutionary path of PDAN expansion.

(c) Results using InterPro
Other databases may define protein domain differently from Pfam-A, resulting in different PDAs and hence PDANs from those presented above. Nevertheless, we observed the same three main architectures of PDAN using InterPro (v. 39) [11], which integrates a number of sequence- and structure-derived protein domain databases, although there were many more tetragon PDANs and PDANs with characteristics of more than one type (see the electronic supplementary material, figure S5). Therefore, the observation of the three types of PDAN was not a result of the analysis done on a particular database or using a particular set of protein domains.

(d) Architectural design principles
Many studies have shed considerable light on mechanisms of protein evolution [14–17], but the architectural design of
‘domain accretion’ remains unclear. Our construction of PDAN necessarily ignored more realistic evolutionary events such as positional preferences and insertion/deletion of more than one single domain [12,13], because otherwise the network would become too convoluted to fathom easily. Nevertheless, our observation of the existence of the three PDAN architectures suggests that there are three distinctive architectural designs, each of which may be favoured by specific types of protein domains. Several lines of evidence support this suggestion. First, star seems to be the blueprint used by ‘promiscuous’ domains [9] to build domain architectures whereby these domains can be linked to a very large number of different domains (e.g. 151 for protein tyrosine kinase domain (PF07714)). Indeed, all but four of the 25 most ‘promiscuous’ domains surveyed by Basu et al. [9] exhibit the characteristics of a dominant star architecture, and even in these four exceptions, a small star centred on the parent SDF is also evident (see the electronic supplementary material, figure S2). Second, tail seems to be the main PDAN architecture of well-known repeat domains, such as ankyrin repeat, WD40 and EGF-like domains, which, in order to serve as a repeat unit, usually have a relatively small domain size [12]. Third, compared with star, the PDAs of a tetragon network tend to link a larger number of domains (i.e. the PDAs are longer), but, for the same network size, use a smaller number of unique domains (figure 2bi,i). In many tetragon PDANs, there appears to be a core set of functionally related domains from which domain architectures are made up of. An example is the network of RNA polymerase Rpb2 domain 3 (PF04565; electronic supplementary material, figure S3), which has 91 nodes (PDAs), which are made up from 16 unique domains, all but two of which are RNA polymerase subunit domains. When the PDAN of the other subunit domains of the RNA polymerase complex is a substantial size (greater than 30 nodes), it forms a network of the tetragon type (data not shown), suggesting that core domains of functional modules in protein complexes may have a tendency to form tetragon PDANs.

The three design forces of PDANs may interact, or even compete, with each other, as it is quite common to observe a mixture of star, tail and tetragon characteristics in these networks (figure 2g for statistics and electronic supplementary material, S4 for examples). An interesting question that arises as a result of this study is what are the functional, and probably structural, constraints that determine if a protein domain evolves a star, tail or tetragon PDAN. Although further studies are required (see for example, electronic supplementary material, figure S9 and table S5 for preliminary results of a domain-centric gene ontology analysis [18]), the identification of these networks and their constituent protein domains and domain architectures should help greatly in investigating this question.

References