Exon–domain correlation and its corollaries

Mingyi Liu, Shaoping Wu, Heiko Walch and Andrei Grigoriev*
GPC Biotech AG, Fraunhoferstrasse 20, 82152 Martinsried, Germany

Received on March 11, 2005; revised on May 16, 2005; accepted on May 18, 2005
Advance Access publication May 19, 2005

Contact: andrei.grigoriev@gpc-biotech.com

Functional proteins are known to contain stretches of amino acid sequences highly conserved in different protein families and across species. These conserved sequences constitute protein domains that are generally integral structural units, conferring specific functionalities and often self-folding. The observation of highly conserved protein domains dispersed across different families and organisms provoked the question whether the domains were reused in some way during evolution.

A compelling mechanism for reusing the domains was put forth by the exon shuffling theory (Gilbert, 1978). In many eukaryotic species, the coding sequences of genes, or exons, are frequently interrupted by non-coding sequences, the introns. When the exon–intron split structure of genes correlates with the organization of protein domains, i.e. the exons match the domains, then duplication, permutation and rearrangement of such exons would create novel genes with reused functional properties. Shuffling of exons could be accomplished through various biological processes such as illegitimate recombination or retrotransposition (van Rijk and Blobel, 2003). Studies on individual proteins, groups of ancient proteins and genome-wide surveys of alpha helices and beta strands indicated gene-structure and protein-structure correlation (Barik, 2004; de Souza et al., 1996; Holmes and Parham, 1985). Intron phase match studies suggested that domains tend to be bounded by symmetrically phased exons, another trait of exon shuffling process (de Souza et al., 1998; Holmes and Parham, 1985; Kaessmann et al., 2002; Pathy, 1996).

A ‘one domain ↔ one exon’ match can be often obscured due to either the loss of bordering introns or the insertion of other introns into the domain-encoding sequence. To compensate for these effects, one can analyze the match only between the borders of domains and exons, thus taking into account cases such as ‘one domain ↔ many exons’ and ‘one exon ↔ many domains’. High-quality domain border annotations can be obtained by analyzing protein sequences with the domain definitions from the Pfam database, which is built from conserved protein sequences in a wide spectrum of species. Using this approach, we have recently demonstrated that correlation between the borders of protein domains and their encoding exons is a genome-wide phenomenon in multiple eukaryotic organisms (Liu and Grigoriev, 2004). Further, we have shown that exon-bordering domains probably contributed more to the expansion and diversification of proteomes than other domains as a result of duplications and exon shuffling, as they preferentially expanded into more genes than other domains during evolution (Liu et al., 2005). To highlight the fact of their remarkable genomic mobility, we use the terms ‘mobile domains’ and ‘exon-bordering domains’ interchangeably in the text below. In this study, we consider two main corollaries of this exon–domain correlation: (1) the impact of mobile domains on the domain network and (2) possible refinement of definitions of individual protein domains.

EXON–DOMAIN CORRELATION AND DOMAIN NETWORKS

Global properties of various networks, such as World Wide Web or biological networks (of interacting and co-occurring protein domains, as well as metabolic networks and those of protein–protein interactions, transcriptional regulation, etc.), have received significant attention in recent years (Apic et al., 2001; Jeong et al., 2000; Luscombe et al., 2002; Ye and Godzik, 2004). Graph representation of such networks abstracts them as nodes connected by edges (e.g. proteins connected by their interactions); the number of edges of a given node is called the node’s degree; nodes with many edges are often called ‘hubs’. Network properties are mainly analyzed from the prospective of their node connectivity (or degree distribution). Many of the networks have been shown to possess scale free property, which means that their degree distribution follows a power law

\[ N(d) \sim d^{-c}, \tag{1} \]

where \( N(d) \) is the number of nodes of degree \( d \) and \( c \) is a constant (generally, 2 < c < 3 for biological networks). Thus there are many poorly connected nodes and very few hubs in such networks.

Earlier, we have shown (Liu et al., 2005) that as a result of preferential amplification exon-bordering domains became on average more abundant and present in more genes than other protein domains. Exon-bordering domains also co-occur with a larger number of different domains to form mosaic proteins with diverse domain architectures. This property suggests that exon-bordering domains should be found among the highly connected hubs and that the evolution of domain networks (at least in terms of degree distribution) is likely to be largely driven by the evolution of exon-bordering domains and their propagation into genes via exon shuffling and duplication mechanisms.

Indeed, many properties of the network of co-occurring protein domains, where each domains in human is a node and an edge represents co-occurrence of two domains (not necessarily adjacent) in one protein, are similar to other biological networks described. As in many of these networks, there is one large component, containing >42% of all nodes and >93% of all edges (Fig. 1A), 172 much smaller components with 311 total edges and

*To whom correspondence should be addressed.
Fig. 1. (A) Snapshot of network representation using PINS software (Grigoriev, 2004). (Left) Statistics of the complete network, including elements hidden from view. (Right) Mobile (red parallelograms) and other domains (green rectangles) are shown as nodes, while edges correspond to one or more human proteins, harboring connected pairs (thick edges indicate 10 or more proteins). Shaded nodes (such as Ig or VWD) are connected to additional nodes, which are hidden from view. Red edges connect pairs of mobile domains. (B) Number of domain pairs is plotted versus the number of proteins containing these pairs, together with the power law trendline. (C) Network parameters after removal of nodes representing mobile domains or the corresponding number of random nodes. Means, SDs and Z-scores (difference with the mean in SD units) calculated from 1000 trials, exon-bordering domain selection was done using two E-value thresholds as described previously (Liu et al., 2005), resulting in groups of 112 and 235 domains; results are similar for both groups and data are shown in the case of 235 domains. Removal of mobile domains results in substantial fragmentation of the network and a drop in the average degree (all values are significantly different between mobile and random nodes, as shown by the Z-scores of at least 13 SD units).

We found that this undirected network is also scale-free (data not shown, this result is analogous to already published reports (Apic et al., 2001; Ye and Godzik, 2004)). In addition to this property we also observed that a distribution of the number of different pairs of domains contained in human proteins also follows power law (Equation (1)), in this case $N(d)$ being the number of domain pairs connected by $d$ proteins and $c = 2.08$ (Fig. 1B). Thus, most of the domain pairs can be found only in one protein per pair. For example, out of 4677 detected domain pairs, only $\sim 200$ pairs occur in more than 10 human proteins. Such proteins, however, are often domain-rich.

We also calculated the expected distribution of co-occurring pairs by modeling domain co-occurrence as a Bernoulli process (where a pair frequency would be proportional to the product of frequencies...
Exon–domain correlation and its corollaries

Fig. 2. Exon–domain correlation and domain definitions. (A) The exon–domain correlation graph for Ig domain shows the amino acid positions near the domain border, where the numbers of exon boundaries are compared with the random expectation. Y-axis shows logarithm of $\chi^2$ P-value calculated from expected and observed numbers of exon boundaries for each amino acid position as described previously (Liu et al., 2005), with negative log used when the observed value is higher than the expected. The Pfam Ig domain shows a ‘negative correlation’ as there were less exon borders observed than expected at every amino acid position inside the $[-10, +10]$ domain border boxes (blue diamonds). However, when we switched to using the SMART definitions of Ig and IgC1, both of which are longer than Pfam Ig definition, we observed a much higher than expected number of exon borders inside domain border boxes at positions $-6$ and $-1$ (red squares and green triangles), respectively. (B) The gene structure (top) and protein domain organization (bottom) for gene DDR2 are shown. In this instance, the FA58C domain correlates with exons 2–4 and is annotated by HmmPfam as from amino acid positions 33–182 on DDR2 protein (ENSP00000294781). However, when we examined the amino acid sequences close to the domain borders, the two highly conserved cystein residues at positions 30 and 185 (marked by arrowheads) were conspicuously outside the HmmPfam annotation. If exon borders between exons 1 and 2 and exons 4 and 5 were chosen to represent the start and end of this domain, respectively, all the conserved residues (highlighted by red font) would be included in the domain.

EXON–DOMAIN CORRELATION COULD HELP RESOLVE CONFLICTS BETWEEN DOMAIN DEFINITIONS FROM DIFFERENT DATABASES

In our method, exon–domain correlation was analyzed as follows. For computational prediction of protein domains in human proteins retrieved from the Ensembl (Birney et al., 2004) genome database, we used HmmPfam (Eddy, 1998) with domain definitions from Pfam (Bateman et al., 2002). For comparison, we also used CD-search (Marchler-Bauer and Bryant, 2004) with domain definitions from SMART (Letunic et al., 2004), a database that uses a mixture of analyses to create protein alignments and domain definitions for a relatively small number of signaling and extracellular domains. We collected statistics for only one multi-exon transcript per gene whose protein translation had at least one domain. Domain prediction may not be exact, so for each domain border, frequencies of exon boundaries were calculated for a window of 10 amino acids (Fig. 1C). Thus, mobile domains appear to be the major determinants of the network topology and evolution.

As a group, exon-bordering domains show a much higher connectivity (Fig. 1) and, as expected, they comprise most of the hubs in the network. We analyzed the level of network fragmentation after the removal of mobile domains by calculating the number of components, or remaining connected subgraphs, and average degree. We also estimated the distributions of these parameters for 1000 networks obtained from the network we studied by removal of the corresponding number of random nodes. Removal of mobile domains results in substantial fragmentation of the network and a drop in the average degree, significantly different from random node removals of individual domains, derived in this case from the number of proteins containing a domain, rather than domain numbers). Best-fit power law trendline for the expected distribution generated a much less steeper curve ($c = 0.88$, $R^2 = 0.74$), indicating that a large proportion of pairwise domain combinations are under-represented in human proteins. Similar findings obtained by other methods have been very recently published for the domain families in SCOP (Vogel et al., 2005), which annotates domains based on structural data, in contrast to the sequence-based Pfam that we had used.

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/21/15/3213/195541 by guest on 20 March 2019
immediately outside the domain and 10 amino acids inside it (marked as [−10;10] window). If a window contained one or more amino acid positions with frequencies significantly different from random expectation (threshold of $P < 10^{-7}$), the domain was deemed exon bordering.

Remarkably, when we detected correlation of the borders of protein domains with encoding exons, it was nearly always positive, i.e. we observed significantly higher numbers of exon borders than that expected in the domain border boxes. However, there was one notable exception: the immunoglobulin domain displayed a negative correlation with exons, with the number of exon borders contained in its border boxes being much smaller than expected. This was rather surprising since the immunoglobulin domain was considered to be mobile and its bounding introns to have phase 1–1, which is the characteristic of mobile domains (Kolkman and Stemmer, 2001). Upon further investigation, we noticed that the Pfam definition of Ig domain was actually 8–20 amino acids shorter than its counterpart domain definition from the SMART database. Owing to this reason, the amino acid positions immediately outside the Ig domain border boxes as defined by Pfam were actually right inside the domain border boxes as defined by SMART. This indicates a preference for the SMART domain definition because we consistently observed lower numbers of exon borders inside Pfam-specific Ig domain border boxes (Fig. 2A) than expected.

When we switched to using SMART domain definition for Ig domains, we discovered that the two most prevalent Ig-related domains in SMART, IgC1 and Ig, were ranked #2 and #7, respectively, out of all human mobile domains, with both having positive correlation with exons in contrast to the results obtained from Pfam’s Ig domain definition. This contrast is even more obvious on the correlation graph that gives us information on where the exon borders preferentially fall at the start and end of domains (data not shown). SMART IG domain has a few peaks from positions −10 to −6 amino acids outside starting position of IG domain, and has a major peak at the first amino acid outside the ending position of the domain. This suggests that the exons correlate with the actual IG domain quite well and that additional analysis of residue conservation between positions −10 and −1 preceding the IG might further improve the domain annotation.

Another interesting example is the FA58C (F5/8 type C) domain present in blood coagulation factors that is thought to be involved in cell adhesion. Our study identified FA58C as a mobile domain that correlates with multiple exons and displays a very strong preference for phase 1–1 introns. The majority of the exon borders inside the two domain border boxes for FA58C fall onto positions −3 to −9, somewhat distant to the domain borders. We investigated the properties of this domain and found that its Pfam prediction of FA58C could be improved by taking into account the exon border positions. In the illustrated example (Fig. 2B), the DDR2 protein contains Pfam-annotated FA58C domain at the N-terminus (amino acids 33–182). However, the highly conserved cysteine residues at both ends of the domain that form a disulfide bond were actually at positions 30 and 185, both excluded by the Hmnpfam prediction yet right inside the exon borders at both ends of FA58C domain. We also found that the domain annotation by pfscan program using Prosise profile is closer to the exon borders and it included the two cysteine residues.

From these examples, it is apparent that at least for identified mobile domains, exon borders could in some cases serve as indicators of domain coordinates to improve (or choose between) predictions of computational tools. In fact, presence of exon borders in the vicinity of domain borders may potentially be used prediction tools themselves.

ACKNOWLEDGEMENTS

We would like to thank Jonathon Blake for helping with the data collection and early discussions and two anonymous reviewers for helpful suggestions.

Conflict of Interest: none declared.

REFERENCES