Kelley and Sternberg *Genome Biology* (2015) 16:100
DOI 10.1186/s13059-015-0663-8

Genome Biology

RESEARCH HIGHLIGHT Open Access

Partial protein domains: evolutionary insights and bioinformatics challenges

Lawrence A Kelley and Michael JE Sternberg*

Abstract

Protein domains are generally thought to correspond to units of evolution. New research raises questions about how such domains are defined with bioinformatics tools and sheds light on how evolution has enabled partial domains to be viable.

With the rapid expansion in the number of determined protein sequences - over 92 million in UniProt in March 2015 - an ever-increasing number of biologists are using bioinformatics tools for annotation of these sequences. One widely used strategy is to identify occurrences of Pfam families within the sequence of interest [1]. A Pfam family is a multiple sequence alignment of the occurrences of a particular domain both in different species and in different regions of the same protein. The concept underpinning Pfam is that proteins typically comprise one or more domains (regions), each of which is an evolutionary unit that generally has a well-defined biological function. A significant sequence similarity between a query protein and a Pfam family provides the basis for annotations. Two recent articles [2,3] in *Genome Biology* evaluate the implications of having the query sequence only matching part of a Pfam family, which is an intriguing finding, given that a Pfam family is considered to be an evolutionary unit.

What is a protein domain?

In discussing these articles, it must be emphasized that the definition of a protein domain is complex. One approach to define a protein domain is from a structural perspective. This was highlighted in Wetlaufer's seminal article [4], where he noted that several protein structures can readily be divided into distinct structural regions (called domains), with few interactions between the domains and few chain crossings between them. The suggestion was that each domain was an independent folding unit. Moreover, certain domains were associated with a particular function, such as NAD binding, and thus these domains were considered an evolutionary unit.

Another approach, currently used in Pfam [1], defines domains from a sequence perspective. This approach aims to identify a section of the protein sequence that shares significant sequence similarity elsewhere in the protein or in other proteins. The use of multiple alignments is central to assessing whether remote sequences are actually similar. In Pfam, the multiple sequence alignment is stored as a hidden Markov model (HMM), which is a statistical formulation that quantifies the archetypal sequence signature that is characteristic of that domain. Underpinning the sequence-based approach is the concept that a domain is an evolutionary unit.

In many proteins, the structural and the sequence-based approaches yield a very similar definition of the domains forming a protein. However, this is not always the case. In Pfam (A Bateman, personal communication), the families are primarily derived based on sequence alignment, but additionally take into account the concept that a Pfam family is likely to be a structural domain. However, as we do not have the structures for all proteins, a Pfam family subsequently could be split once a structure is determined - but of course this takes time. Moreover, the structural definition of a domain involves substantial subjective input, and algorithms are able only to provide guidance. The extent of the problem of defining domains is highlighted in a recent study to map CATH (for 'class-architecture-topology-homologous superfamily') [5] to SCOP ('structural classification of proteins') [6] domains undertaken in the Genome3D consortium [7]. The principle behind SCOP is that a region has to be seen independently to merit being defined as a domain, and so there is a class in SCOP called multi-domain proteins that has multi-lobal structures. CATH, by contrast,

* Correspondence: m.sternberg@imperial.ac.uk
Structural Bioinformatics Group, Centre for Integrative Systems Biology and Bioinformatics, Department of Life Sciences, Imperial College London, London SW7 2AZ, UK

would split the multi-lobal structure into component domains. Only 60% of CATH domains are similar to a SCOP counterpart (the silver standard in Genome3D) (N Nadzirin and C Orengo, personal communication).

Explanations for a partial match to an archetypal Pfam domain

Two recent *Genome Biology* articles investigated partial matches to Pfam domains. Triant and Pearson [2] show that almost 4% of Pfam domains from a representative subset (RefProtDom2 (RPD2) with 136 families) are shorter than 50% of the length of the HMM of that family. Detailed examination identifies several reasons for these partial matches. The largest fraction of 50% partial domains are split-partials - partials where a complete domain has been split into smaller regions. After inspection, one could reassemble the full-length domain, but a naive user of Pfam might not be aware of the problem. In many other cases, there is strong evidence that this partial alignment is the result of errors in the identification of the protein sequence in the truncated domain, and this is particularly prevalent in eukaryotic sequences from poorly annotated species. Other occurrences of partial domain matches can be explained by the domain being split by an inserted region or an alignment that should extend beyond the matched region, but has failed to do so owing to a weak sequence similarity signal. However, there are a few observed instances (18 out of the 136 Pfam families) where the Pfam HMM corresponds to multiple CATH [5] domains or multiple ‘vector alignment search tool’ VAST+ [8] domains, or both. Thus, the apparent partial match will typically be the result of matching to one of these component domains. This study highlights the challenges in protein annotation and in using sequence families to identify independent structural domains.

A related question is considered by Prakash and Bateman [3]. They identify a few proteins where the protein sequence lacks one-third or more of the residues of the Pfam model. Manual inspection and filtering identifies proteins where the partial match could be explained by gene-prediction error or by other well-recognized effects such as multi-domain proteins. However, after filtering, they identify sequences where there is a partial match to an independently folded domain, and the authors introduce the term 'domain atrophy'. They note that domain atrophy is very rare (0.06% of all Pfam domains), but when it does occur it raises the question of how such a partial domain is able to fold into a functional stable unit. For 75 domains where such atrophy occurs, there is an available structure - either experimental or inferred from the coordinates of a homolog. Detailed inspection shows that often domain-domain or subunit-subunit interactions lead to stabilization of the atrophied domain.

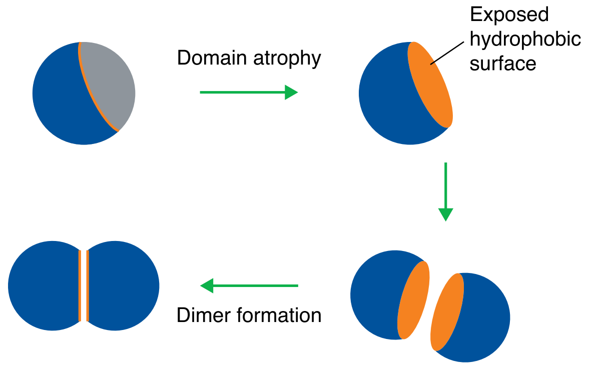

Figure 1 Schematic illustrating how domain atrophy (loss of gray region from blue sphere) can lead to exposure of hydrophobic residues (orange). This exposed hydrophobic surface can subsequently be stabilized by dimerization. See Prakash and Bateman [3] in this issue for detailed examples.

A particularly striking case (illustrated schematically in Figure 1) is the bacterial luciferase domain from *Photobacterium phosphoreum* LuxF [9]. This domain lacks one β-strand and three α-helices from the standard structure of bacterial luciferase that comprises eight buried β-strands forming a β-barrel surrounded by α-helices. One might expect that the fold would be highly unstable. Indeed, the crystal structure of this protein reveals that the monomer has a large hydrophobic cleft that is not buried. However, a homo-dimeric interaction buries this cleft and thus stabilizes the protein. This and the other examples that the authors list highlight the mechanisms by which evolution has managed to ensure that atrophied domains remain viable.

General implication for protein bioinformatics

The difficulties that can occur in Pfam-based protein annotations can also arise when other strategies are applied, such as in prediction of protein structure. For example, a web-based resource for template-based modelling is available for the community - Phyre2 [10]. These two articles will provide for the users valuable suggestions as to the possible causes for Phyre2 predicting only a partial protein domain. Thus, when interpreting results of bioinformatics resources, biologists need to consider both the possibility of erroneous data and the fascinating diversity of mechanisms that can occur during evolution to deliver biological function.

Abbreviations

CATH: Class-architecture-topology-homologous superfamily; HMM: Hidden Markov model; NAD: Nicotinamide adenine dinucleotide; RFPD2: RefProtDom2; SCOP: Structural classification of proteins.

Competing interests

MJES is a director and shareholder in Equinox Pharma Ltd, which applies bioinformatics and chemoinformatics to drug discovery. Phyre2 might shortly be available for commercial use.

Authors' contributions

LAK and MJES jointly wrote this article and have approved the final manuscript.

Acknowledgements

This work was supported by the Biotechnology and Biological Sciences Research Council (BBSRC) (BB/J019240/1) and The Wellcome Trust (WT/104955/Z/14/Z).

Published online: 19 May 2015

References

1. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, et al. Pfam: the protein families database. Nucleic Acids Res. 2014;42:D222–30.
2. Triant DA, Pearson WR. Most partial domains in proteins are alignment and annotation artifacts. Genome Biol 2015;16:99
3. Prakash A, Bateman A. Domain atrophy creates rare cases of functional partial protein domains. Genome Biol 2015;16:88
4. Wetlaufer DB. Nucleation, rapid folding, and globular intrachain regions in proteins. Proc Natl Acad Sci U S A. 1973;70:697–701.
5. Cuff AL, Sillitoe I, Lewis T, Redfern OC, Garratt R, Thornton J, et al. The CATH classification revisited—architectures reviewed and new ways to characterize structural divergence in superfamilies. Nucleic Acids Res. 2009;37:D310–14.
6. Andreeva A, Howorth D, Brenner SE, Hubbard TJ, Chothia C, Murzin AG. SCOP database in 2004: refinements integrate structure and sequence family data. Nucleic Acids Res. 2004;32:D226–9.
7. Lewis TE, Sillitoe I, Andreeva A, Blundell TL, Buchan DW, Chothia C, et al. Genome3D: exploiting structure to help users understand their sequences. Nucleic Acids Res. 2014;D382-6.
8. Madej T, Lanczycki CJ, Zhang D, Thiessen PA, Geer RC, Marchler-Bauer A, et al. MMDB and VAST+: tracking structural similarities between macromolecular complexes. Nucleic Acids Res. 2014;D42:297–303.
9. Kita A, Kasai S, Miyata M, Miki K. Structure of flavoprotein FP390 from a luminescent bacterium *Photobacterium phosphoreum* refined at 2.7Å resolution. Acta Crystallogr D Biol Crystallogr. 1996;52:77–86.
10. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE. The Phyre2 web portal for protein modelling, prediction and analysis. Nat Protoc. 2015 (in press)