Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Genetics & Development

The language of the protein universe

Andrea Scaiewicz and Michael Levitt

Proteins, the main cell machinery which play a major role in nearly every cellular process, have always been a central focus in biology. We live in the post-genomic era, and inferring information from massive data sets is a steadily growing universal challenge. The increasing availability of fully sequenced genomes can be regarded as the 'Rosetta Stone' of the protein universe, allowing the understanding of genomes and their evolution, just as the original Rosetta Stone allowed Champollion to decipher the ancient Egyptian hieroglyphics. In this review, we consider aspects of the protein domain architectures repertoire that are closely related to those of human languages and aim to provide some insights about the language of proteins.

Address
Department of Structural Biology, Stanford University, Stanford, CA 94305-5126, United States

Corresponding author: Levitt, Michael (michael.levitt@stanford.edu)

Current Opinion in Genetics & Development 2015, **35**:50–56

This review comes from a themed issue on **Genomes and evolution**

Edited by **Antonis Rokas** and **Pamela S Soltis**

For a complete overview see the Issue and the Editorial

Available online 3rd November 2015

http://dx.doi.org/10.1016/j.gde.2015.08.010

Biological linguistics

Connections between linguistics and biological processes have been described previously [1,2]. The human genome has been referred as the 'book of life' [3] and DNA as the 'cell language'. Chomsky's linguistic rules [4,5] have led to appealing terms such as 'molecular linguistics' [6] and 'RNA folding grammar' [7]. This is not surprising since biological molecules, at their most basic level, can be regarded as strings of letters derived from some alphabet.

There is a wide information gap between the exponentially growing number of protein sequences and the biological information on their structures and functions. Without experimental confirmation, the vast majority of the sequences have a 'hypothetical' role by nature. This has made the classification of the protein sequence universe a central task of computational biology.

It is well established that protein domains are essential blocks of protein evolution [8]. Denoting proteins sequences as the N-to-C terminal sequential order of their domains, or domain architecture, parallels viewing sequences as sentences (multi domain architectures) composed of words (single domains). This approach enables us to think of the repertoire of domain architectures in the protein universe as a language, and hence use linguistic approaches to understand the language of proteins. Figure 1 shows the correspondence between the human and protein languages at different levels.

The vocabulary of proteins

Protein domains correspond to the words in the proteins language. Domains can be distinguished by their sequence (sequence profiles) or their structure (classification databases). Sequence profiles are most commonly represented using statistical models such as position specific scoring matrices (PSSM) and hidden Markov models (HMM). HMM-based methods include Pfam [9], EVEREST [10], SMART [11], and PANTHER [12]. PSSM-based methods include PRINTS [13], PROSITE [14] and ProDom [15]. Here we refer to sequence profiles as domains or words. Structure-based classifications including SCOP [16,17], SCOP2 [18] and CATH [19–21], as well as predicted domain structures as in SUPERFAMILY [22,23], Gene3D [24,25], ECOD [26] and COPS [27] are not considered here.

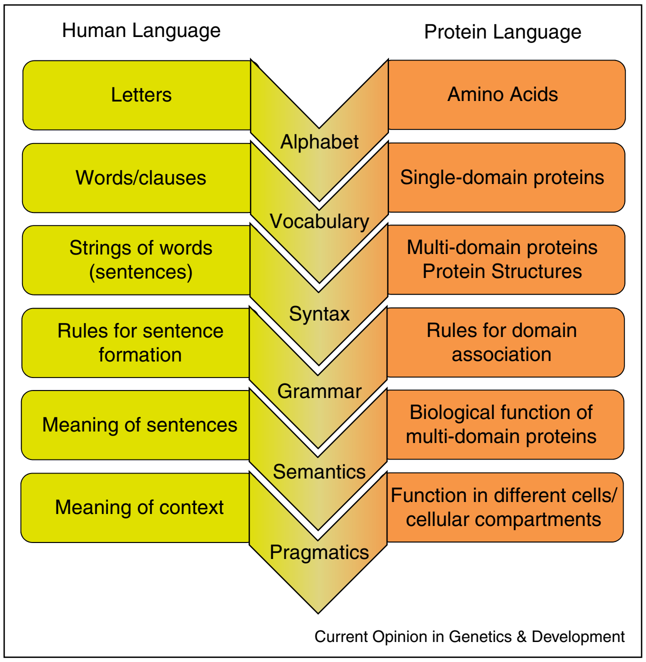
Sequence-based methods differ in coverage, level of curation and definition of families. Compilation databases like CDART [28–30] and InterPro [31] match all sets of profiles to all known sequences (Figure 2). This facilitates the classification of the protein universe into protein families by providing the locations of different domains along every sequence. Often two different sequence profiles match the same region of sequence leading to several domains, or words, for the same physical object. This can lead to confusion and such synonyms need to be recognized and possibly eliminated [32••].

Grammar of proteins

Words are the basic components of languages, but it is the combination of the words that defines the language. Proteins domain architectures (sentences) define structure and biological function of proteins; the repertoire of domain architectures is the language of the protein universe. Studies revealing its insights involve three types of representations: classifications, maps and networks (see review [33•]).

Classifications use protein domains as the units, have high information content and are often used when studying a specific protein family. Levitt [32••] provided a detailed

Figure 1


Human Language
Protein Language
Letters
Amino Acids
Alphabet
Words/clauses
Single-domain proteins
Vocabulary
Strings of words (sentences)
Multi-domain proteins Protein Structures
Syntax
Rules for sentence formation
Rules for domain association
Grammar
Meaning of sentences
Biological function of multi-domain proteins
Semantics
Meaning of context
Function in different cells/ cellular compartments
Pragmatics
Current Opinion in Genetics & Development

Analogy between human and proteins languages. In this comparison, the vocabulary (domains) of proteins is built from an alphabet of amino acids. The syntax principles enable domain association to form multi-domain architectures, a process governed by hierarchical rules (grammar), that determine the structure and hence the biological function (semantics) of proteins. In several languages, for example in English, a number of different classes of words exist (nouns, adjectives, verbs, adverbs, pronouns, conjunctions). Each class has its task in the language, that is, nouns name words, adjectives describe nouns, verbs are action words, conjunction connect words. Analogously, one can also distinguish different classes of domains with different tasks (motors, binding proteins, enzymes, signaling proteins, structural proteins, targeting proteins).

classification of the domains in the entire protein universe, revealing that despite the exponential increase in the number of sequences since 1980, the number of domains is saturating while the number multi domain architectures is increasing as fast as sequences. In 2009, about a quarter of the sequences in the protein universe could not be assigned a domain and comprised the dark matter. Our latter study confirms these findings and also shows a decrease in the fraction of dark matter in 2014 to 22% (Scaiewicz & Levitt, unpublished results). InterPro reports 16.5% dark matter (83.5% coverage) for the sequences in UniProtKB release 2014 [31]. Redefining the dark matter including inter-domain residues [34], suggested a much higher percentage of dark matter (∼45%) [34].

Networks, which can detect novel relationships that could be missed by pairwise-based methods, suggest evolutionary paths between domains. Recently, computational tools that use protein similarity networks to illustrate functional relationships between huge groups of proteins have emerged. Nepomnyachiy *et al.* [35••] constructed networks to present similarities among a representative set of all known SCOP domains showing that protein space has a continuous region formed by a large connected component and a much smaller region comprised by isolated islands. This allows two remotely related proteins to be connected through a transitive path.

More recent approaches identify homologous sequences by global alignment of domain architectures [36]. In RADS/RAMPAGE [37], proteins are globally compared as strings of domains providing a fast, sensitive homology search. MDAT [38] uses a domain similarity matrix to score domain pairs and aligns the domain arrangements by a progressive alignment method. ADASS [39] compares and classifies protein domain architectures by recognizing similarity between the domain architectures even if the proteins share very poor sequence similarity; it includes neighbor information in its score. The Enrichment of Network Topological Similarity (ENTS), is a framework to improve the performance of large-scale similarity searches; it considers a continuous protein space and performs well on the fold recognition problem [40]. ArchSchema [41] uses a domain-graph of related domain arrangements. DoMosaics [42] unifies protein domain annotation (via InterProScan and HMMER), domain arrangement analysis and visualization of domain architecture evolution in a single tool. CDvist [43] (a comprehensive domain visualization tool) searches a series of domain databases using the best algorithms in a user-friendly framework.

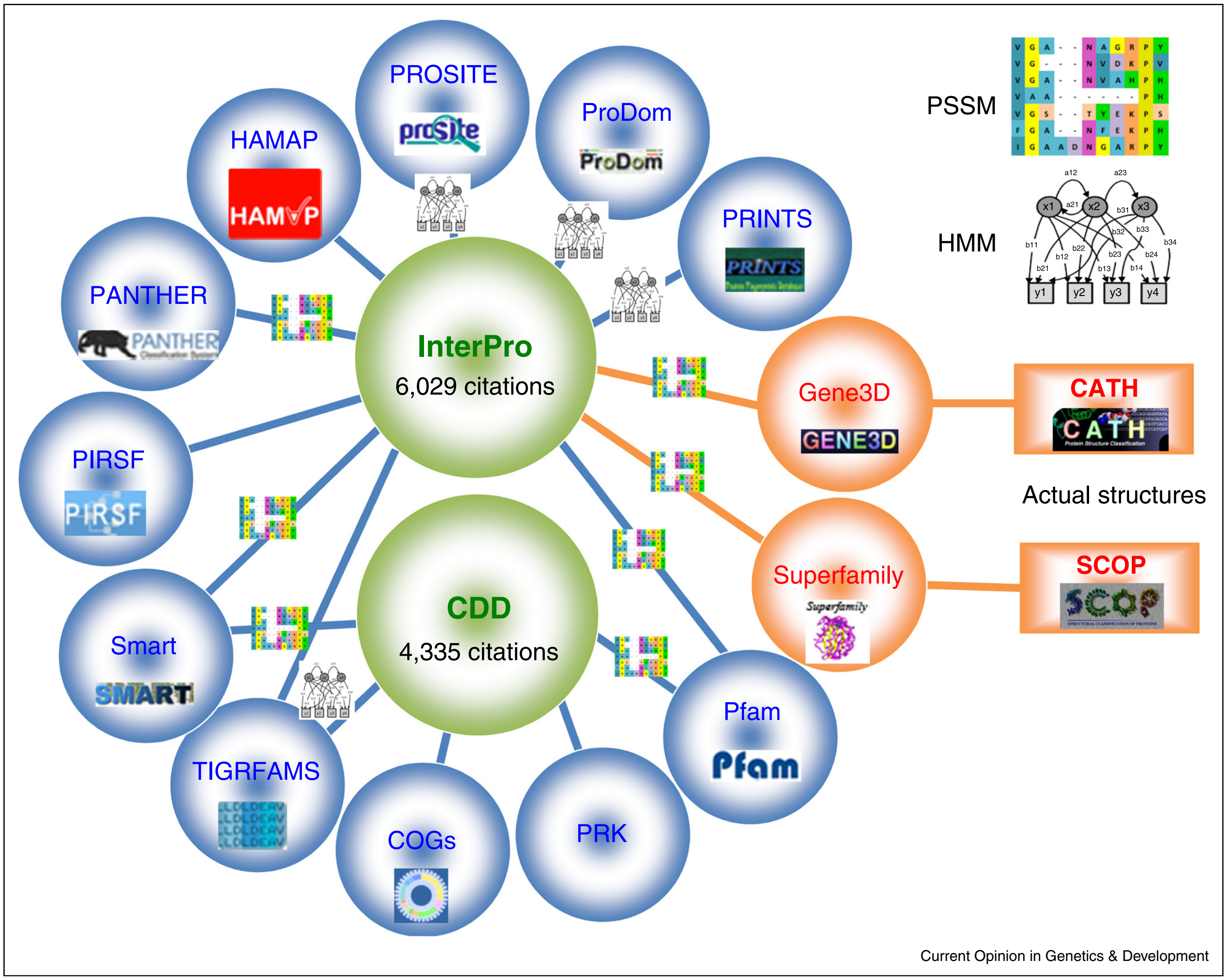
Maps simplify protein space as a set of points (proteins or domains) and the distances (usually structural or sequence similarity) between them. Biological properties can be shown by coloring the points enabling a better visualization of protein space [33•].

The Supra-domain is a more sophisticated syntax element of protein language; it is a conserved combination of domains in different proteins [44,45]. Supra-domains may involve of contacts between its constituent domains, a step beyond consecutive modular folding units [46].

Types of words: domain versatility

Although some domains always form in the same combinations, others are seen in many different arrangements. These are called versatile (or promiscuous) domains, and may be pictured as frequently used words. Each domain has an intrinsic combinatorial propensity, and its effects have been studied using measures of domain versatility such as the number of different direct neighbors of a given domain, number of domains with which a given domain occurs in the same arrangement and number of triplets, to name a few. Weiner *et al.* [47] proposed a domain versatility index (DVI), which eliminates dependence of

Figure 2



Sequence profile databases can be sequence-based (blue circles) or structure-based (orange circles). Sequence-based profiles are derived by mainly two methods: HMMs (Hidden Markov Models) or PSSMs. (Position Sensitive Sequence Matrices). Structure-based profiles in Gene3D and superfamily are generated from HMMs built from actual structures coming from CATH and SCOP, respectively. Two main integrative resources, CDART and InterPro, are shown (green circles) with the databases they include.

versatility on domain frequency and is 'age' independent, where 'age' is the first appearance of a domain in the species tree. Higher DVI values were seen for domains also occurring in single domain proteins as well as domains seen frequently at protein termini. This agrees with findings that domain evolution is driven by fusion of pre-existing arrangements including single domains and loss of domains at protein termini [48,49•,50]. A recent study [51] modeled domain combinations as bigrams and introduced the 'networking versatility' measure to suggest that more versatile domains are much more highly adaptive.

Domain versatility has been shown to depend on function: versatile domains are required for specific functions such as protein–protein interactions and play roles in signaling pathways. It was suggested that versatility appeared because a domain had the potential to be useful in various contexts [52]. A study on fungi associated cell survival and interaction with the environment with versatile domains [53].

Eukaryote domains have higher versatility than prokaryote and such domains evolve more slowly than non versatile domains [52]. Domain versatility follows a power law distribution: a small number of large families with many types of neighbors, and a large number of families with few types of neighbors [54]. Versatility is a volatile characteristic in evolution; only a few domains retain their versatility status throughout evolution, usually acquired independently in different lineages [54].

Hsu *et al.* [55] used Pfam-A domains to build networks of connected domain architectures that had been derived from the same single domain. They found three distinctive types of networks: 'star' networks are generally adopted by versatile domains; 'tail' networks are adopted by well known repeat domains with a relatively small domain size; and 'tetragon' networks link a larger number of domains (longer domain architectures), but contain a smaller number of unique domains (involve a core set of functionally related domains). These networks are each adopted by different types of domains, although some networks exhibited the characteristics of more than one type.

Protein semantics

Semantics refers to the meaning of a linguistic expression. Understanding the semantics of the protein language is a challenging task, mainly because the 'words' in protein sentences may have multiple meanings in different contexts, as do polysemous words in complex languages. For example, the word 'table' can mean a piece of furniture, or a graphic representation of facts. In the same way, domains with high sequence similarity but different structure or function exist. Examples of the same sequence adopting different structures are reviewed in [56]. Some homologous genes may adopt novel functions during evolution, sometimes quite different from their original ones; see [57–59] for examples of databases that collect such genes. Multi-tasking (or moonlighting) proteins are proteins that perform different functions in different environments [60–62].

Since most available sequences lack biological annotation, there is a strong need for computational methods for automatic annotation of protein function. Domains are less studied than proteins or genes in terms of ontology annotation. The dcGO database [63] addresses this need and provides a systematic annotation of domains using a panel of ontologies. The large-scale community-based critical assessment of protein function annotation (CAFA) experiment [64••] assessed fifty-four state of the art methods for protein function prediction. Best performing methods were those recently developed and based on machine learning, but there is still a considerable need for improvement of currently available tools. All evaluated methods showed higher accuracy on single domain proteins.

Evolution of languages and protein domains

Biology and linguistics are interweaved in many aspects as both share methods and evolutionary theories. Linguistics uses homology to compare different vocabularies and elucidate the history of languages. Swadesh [65] compiled a list of less than 200 words (core vocabulary) which reflect the main concepts expressed in all known human languages and possibly resilient to change. By comparing similarities between corresponding words in different languages, he identified words that presumably originated from the same 'ancestor language' (cognates) and proposed that languages sharing a higher percentage of cognates were more closely related and may have separated more recently. The suggestion that an average rate of word substitution over time is an intrinsic constant of languages, is a concept parallel to the 'molecular clock' in sequence-based evolutionary biology.

An analysis of the distribution of SCOP domains across 40 genomes from archaea, bacteria, and eukaryotes found that a majority of domain families are common to all three kingdoms of life, and thus likely to be ancient [66]. A later study of 172 complete eukaryotic genomes compared the Pfam-based domain architectures using a maximum parsimony. They found prevalent independent evolution of domain combinations: about a quarter of the domain combinations have evolved multiple times and this fraction is even higher in individual species. Moreover, 70% of all domain combinations in the human genome independently appeared in at least one other eukaryotic genome. This difference in specificity suggests the existence of a core set of domain combinations constantly recurring in different species, and a lesser number of unique domain combinations that do not appear anywhere else [67••]. This is in agreement with the findings that most multidomain architectures are species-specific while single domains are common to all species [32••,68]. However, early studies indicate that domain architecture reinvention is a common phenomenon and convergent domain architecture evolution is rare [69,70]. Recursion, a fundamental characteristic of languages, is also seen in the proteins language where a limited set of domains are re-used to create new more complex multi-domain architectures [32••,46].

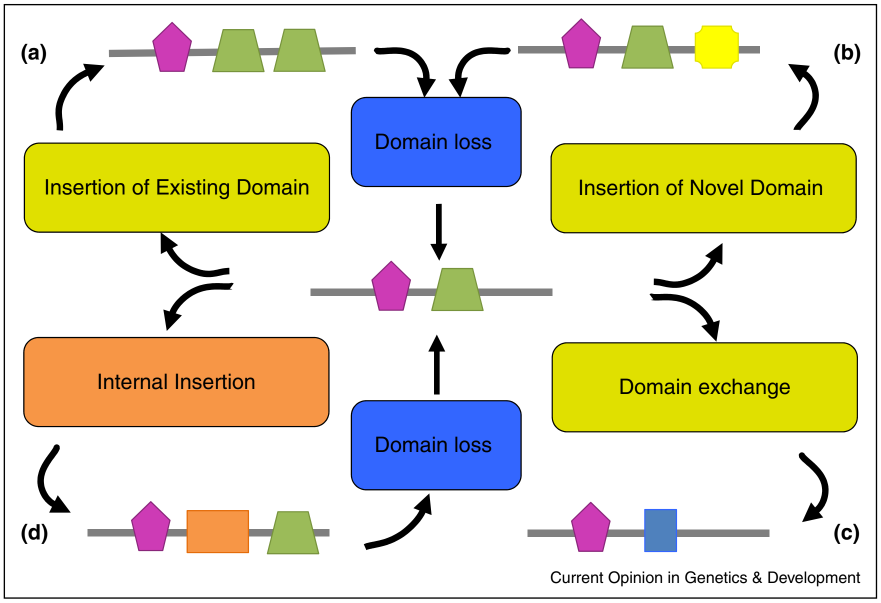
Many domains in multidomain structures could have once existed as independent proteins. A 'domain-protein-reaction network' built from 13 species suggested the majority of novel domains arise to catalyze novel reactions in the metabolic system [71].

Mechanisms for domain evolution

The study of protein domain architectures and the migration of protein domains within and across different genomes provide clues about the evolution of protein function. The modularity of proteins is likely favored by evolution as it permits the combination of pre-existing domains to acquire new functions. Most domain architectures evolve from ancient ones, and convergent evolution resulting in the same architectures in eukaryotic species of different lineages is rare [72]. The majority of genomic proteins, two-thirds in unicellular organisms and more than 80% in metazoa, are multidomain proteins created as a result of gene duplication events [66,73].

The average length of a protein domain is about 120 amino acids, hence changes in domain architecture are

Figure 3


(a)
(b)
Domain loss
Insertion of Existing Domain
Insertion of Novel Domain
Internal Insertion
Domain exchange
Domain loss
(d)
(c)
Current Opinion in Genetics & Development

Mechanisms of domain architecture change. New domain architectures can be created by **(a)** insertion or loss of existing domains (duplication), **(b)** insertion or loss of novel domains or by **(c)** substitution of one domain for another (exchange), which is almost always a two-step process comprising loss/fission and fusion. Domain insertion and loss are a consequence of gene fusion and fission respectively. Domain insertions or losses can be **(d)** internal (between two domains, also called domain shuffling) or may occur either at the N or C terminus (a) & (b).

underlined by large alterations at the gene level, such as gene fusion and fission, exon shuffling through intronic recombination, alternative gene splicing and retropositioning [48,74••,75••]. The elementary mechanisms for new domain architectures creation may be classified into three classes: (1) domain insertion or loss, (2) domain exchange and (3) domain repetition. Figure 3 shows an inserted domain can be fused at the C-termini or N-termini and can be a novel domain (B) or an already existing one (A), which is a domain duplication. Internal insertions of domains (D) are a consequence of exon shuffling and domain losses are a consequence of gene fission (C). Although gene fusion can add domains only at terminal positions, domain-shuffling can insert domains both at internal and terminal positions. Studies of the differences between domain architecture changes in internal positions versus terminal positions enable the assessment of the relative contribution of gene fusion and domain-shuffling to evolution. Changes in domain architecture preferentially occur at protein termini in metazoan proteins through fusion [48,50]. New domains are seen in higher copy numbers than older domains and it is their highly adaptive potential that allows them to dynamically spread across genomes mainly through gene duplication, fusion and terminal domain loss. Transposition, exons shuffling and recombination play only a minor role [74••].

Although fusion mostly effects binding-related functionalities, fission shows stronger effects on catalytic activities. Both fusion and fission effect terms across all levels of signaling pathways, from signal triggering components to transcription factor activity [68]. Prokaryotes and eukaryotes share dominant evolutionary mechanisms in the early stage but diverge substantially along each clade [71].

Conclusion

This short review cannot do justice to a very complicated and burgeoning field that deals with millions of protein sequences in thousands of species whose genomes have been determined. This data is growing exponentially by doubling every year. The scope of the project is massively ambitious, aimed as it is, at understanding the language of living organisms. Although there has been great improvement on computational technologies and availability of methods, we are likely to be involved with such a difficult task for many years to come.

Acknowledgement

This work was supported by the NIH grant GM063817.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Searls DB: **The language of genes**. *Nature* 2002, **420**:211-217.

2. Gimona M: **Protein linguistics — a grammar for modular protein assembly?** *Nat Rev Mol Cell Biol* 2006, **7**:68-73.

3. Eisenhaber F: **A decade after the first full human genome sequencing: when will we understand our own genome?** *J Bioinf Comput Biol* 2012, **10**.

4. Chomsky N: **Logical-structures in language**. *Am Doc* 1957, **8**:284-291.

5. Chomsky N: **Fundamentals of language — jakobson,r, halle,m**. *Int J Am Linguist* 1957, **23**:234-242.

6. Botstein D, Cherry JM: **Molecular linguistics: extracting information from gene and protein sequences**. *Proc Natl Acad Sci U S A* 1997, **94**:5506-5507.

7. Group NP: **Folding as grammar**. *Nat Struct Biol* 2002, **9** 713-713.

8. Chothia C, Gough J: **Genomic and structural aspects of protein evolution**. *Biochem J* 2009, **419**:15-28.

9. Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K *et al.*: **The pfam protein families database**. *Nucleic Acids Res* 2010, **38(Database Issue)**:D211-D222.

10. Portugaly E, Linial N, Linial M: **Everest: a collection of evolutionary conserved protein domains**. *Nucleic Acids Res* 2007, **35(Database Issue)**:D241-D246.

11. Letunic I, Doerks T, Bork P: **Smart 6: recent updates and new developments**. *Nucleic Acids Res* 2009, **37(Database Issue)**:D229-D232.

12. Mi HY, Dong Q, Muruganujan A, Gaudet P, Lewis S, Thomas PD: **Panther version 7: improved phylogenetic trees, orthologs and collaboration with the gene ontology consortium**. *Nucleic Acids Res* 2010, **38(Database Issue)**:D204-D210.

13. Attwood TK, Bradley P, Flower DR, Gaulton A, Maudling N, Mitchell AL, Moulton G, Nordle A, Paine K, Taylor P *et al.*: **Prints and its automatic supplement, preprints**. *Nucleic Acids Res* 2003, **31**:400-402.

14. Sigrist CJA, Cerutti L, de Castro E, Langendijk-Genevaux PS, Bulliard V, Bairoch A, Hulo N: **Prosite, a protein domain database for functional characterization and annotation**. *Nucleic Acids Res* 2010, **38(Database Issue)**:D161-D166.

15. Bru C, Courcelle E, Carrre S, Beausse Y, Dalmar S, Kahn D: **The prodom database of protein domain families: more emphasis on 3d**. *Nucleic Acids Res* 2005, **33(Database Issue)**:D212-D215.

16. Murzin AG, Brenner SE, Hubbard T, Chothia C: **Scop: a structural classification of proteins database for the investigation of sequences and structures**. *J Mol Biol* 1995, **247**:536-540.

17. Andreeva A, Howorth D, Chandonia JM, Brenner SE, Hubbard TJ, Chothia C, Murzin AG: **Data growth and its impact on the scop database: new developments**. *Nucleic Acids Res* 2008, **36(Database issue)**:D419-D425.

18. Andreeva A, Howorth D, Chothia C, Kulesha E, Murzin AG: **Scop2 prototype: a new approach to protein structure mining**. *Nucleic Acids Res* 2014, **42(D1)**:D310-D314.

19. Orengo CA, Bray JE, Buchan DW, Harrison A, Lee D, Pearl FM, Sillitoe I, Todd AE, Thornton JM: **The cath protein family database: a resource for structural and functional annotation of genomes**. *Proteomics* 2002, **2(1)**:11-21.

20. Cuff AL, Sillitoe I, Lewis T, Clegg AB, Rentzsch R, Furnham N, Pellegrini-Calace M, Jones D, Thornton J, Orengo CA: **Extending cath: increasing coverage of the protein structure universe and linking structure with function**. *Nucleic Acids Res* 2011, **39(Database issue)**:D420-D426.

21. Sillitoe I, Lewis TE, Cuff A, Das S, Ashford P, Dawson NL, Furnham N, Laskowski RA, Lee D, Lees JG *et al.*: **Cath: comprehensive structural and functional annotations for genome sequences**. *Nucleic Acids Res* 2015, **43(Database issue)**:D376-D381.

22. Gough J, Chothia C: **Superfamily: hmms representing all proteins of known structure. Scop sequence searches, alignments and genome assignments**. *Nucleic Acids Res* 2002, **30**:268-272.

23. Wilson D, Pethica R, Zhou Y, Talbot C, Vogel C, Madera M, Chothia C, Gough J: **Superfamily — sophisticated comparative genomics, data mining, visualization and phylogeny**. *Nucleic Acids Res* 2009, **37(Database issue)**:D380-D386.

24. Yeats C, Lees J, Carter P, Sillitoe I, Orengo C: **The gene3d web services: a platform for identifying, annotating and comparing structural domains in protein sequences**. *Nucleic Acids Res* 2011, **39(Web Server issue)**:W546-W550.

25. Lees JG, Lee D, Studer RA, Dawson NL, Sillitoe I, Das S, Yeats C, Dessailly BH, Rentzsch R, Orengo CA: **Gene3d: multi-domain annotations for protein sequence and comparative genome analysis**. *Nucleic Acids Res* 2014, **42(Database issue)**:D240-D245.

26. Cheng H, Schaeffer RD, Liao Y, Kinch LN, Pei J, Shi S, Kim BH, Grishin NV: **Ecod: an evolutionary classification of protein domains**. *PLoS Comput Biol* 2014, **10**:e1003926.

27. Suhrer SJ, Wiederstein M, Gruber M, Sippl MJ: **Cops — a novel workbench for explorations in fold space**. *Nucleic Acids Res* 2009, **37(Web Server issue)**:W539-W544.

28. Marchler-Bauer A, Anderson JB, Derbyshire MK, DeWeese-Scott C, Gonzales NR, Gwadz M, Hao LN, He SQ, Hurwitz DI, Jackson JD *et al.*: **Cdd: a conserved domain database for interactive domain family analysis**. *Nucleic Acids Res* 2007, **35(Database issue)**:D237-D240.

29. Marchler-Bauer A, Zheng C, Chitsaz F, Derbyshire MK, Geer LY, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Lanczycki CJ *et al.*: **Cdd: conserved domains and protein three-dimensional structure**. *Nucleic Acids Res* 2013, **41(Database issue)**: D348-D352.

30. Geer LY, Domrachev M, Lipman DJ, Bryant SH: **Cdart: protein homology by domain architecture**. *Genome Res* 2002, **12**:1619-1623.

31. Mitchell A, Chang HY, Daugherty L, Fraser M, Hunter S, Lopez R, McAnulla C, McMenamin C, Nuka G, Pesseat S *et al.*: **The interpro protein families database: the classification resource after 15 years**. *Nucleic Acids Res* 2015, **43(Database issue)**: D213-D221.

32. Levitt M: **Nature of the protein universe**. *Proc Natl Acad Sci U S A*
•• 2009, **106**:11079-11084.
A meticulous study of all known sequences in terms of families that have single-domain or multi-domain architectures reveals single-domain families grow very slow and are mostly shared by the major groups of organisms while multi-domain families grow as fast as the number of sequences and account for species diversity.

33. Ben-Tal N, Kolodny R: **Representation of the protein universe**
• **using classifications, maps, and networks**. *Israel J Chem* 2014, **54**:1286-1292.
The authors surveyed three alternative global representations of protein space, which provide a global picture of the protein universe: classifications, maps, and networks.

34. Rekapalli B, Wuichet K, Peterson GD, Zhulin IB: **Dynamics of domain coverage of the protein sequence universe**. *BMC Genomics* 2012, **13**:634.

35. Nepomnyachiy S, Ben-Tal N, Kolodny R: **Global view of**
•• **the protein universe**. *Proc Natl Acad Sci U S A* 2014, **111**:11691-11696.
The authors used domain and motif networks built from evolutionary relationships among a representative set of SCOP domains to reveal that protein space is both discrete and continuous and that recurring motifs, especially the all-alpha and alpha/beta domains are very abundant.

36. Kummerfeld SK, Teichmann SA: **Protein domain organisation: adding order**. *BMC Bioinform* 2009, **10**:39.

37. Terrapon N, Weiner J, Grath S, Moore AD, Bornberg-Bauer E: **Rapid similarity search of proteins using alignments of domain arrangements**. *Bioinformatics* 2014, **30**:274-281.

38. Kemena C, Bitard-Feildel T, Bornberg-Bauer E: **Mdat-aligning multiple domain arrangements**. *BMC Bioinformatics* 2015, **16**:19.

39. Syamaladevi DP, Joshi A, Sowdhamini R: **An alignment-free domain architecture similarity search (adass) algorithm for**

inferring homology between multi-domain proteins. *Bioinformation* 2013, **9**:491-499.

40. Lhota J, Hauptman R, Hart T, Ng C, Xie L: **A new method to improve network topological similarity search: applied to fold recognition**. *Bioinformatics* 2015, **31**:2106-2114.

41. Tamuri AU, Laskowski RA: **Archschema: a tool for interactive graphing of related pfam domain architectures**. *Bioinformatics* 2010, **26**:1260-1261.

42. Moore AD, Held A, Terrapon N, Weiner J 3rd, Bornberg-Bauer E: **Domosaics: software for domain arrangement visualization and domain-centric analysis of proteins**. *Bioinformatics* 2014, **30**:282-283.

43. Adebali O, Ortega DR, Zhulin IB: **Cdvist: a webserver for identification and visualization of conserved domains in protein sequences**. *Bioinformatics* 2015, **31**:1475-1477.

44. Haynie DT, Xue B: **Superdomains in the protein structure hierarchy: the case of ptp-c2**. *Protein Sci* 2015, **24**:874-882.

45. Vogel C, Berzuini C, Bashton M, Gough J, Teichmann SA: **Supra-domains: evolutionary units larger than single protein domains**. *J Mol Biol* 2004, **336**:809-823.

46. Bornberg-Bauer E: **Signals: tinkering with domains**. *Sci Signal* 2010, **3**.

47. Weiner J, Moore AD, Bornberg-Bauer E: **Just how versatile are domains?** *Bmc Evol Biol* 2008, **8**.

48. Buljan M, Frankish A, Bateman A: **Quantifying the mechanisms of domain gain in animal proteins**. *Genome Biol* 2010, **11**.

49. Nasir A, Kim KM, Caetano-Anolles G: **Global patterns of protein
• domain gain and loss in superkingdoms**. *PLoS Comput Biol* 2014, **10**:e1003452.

The authors describe the evolutionary dynamics of protein domains and model the effects of domain gain and loss in the proteomes of fully sequenced from Archaea, Bacteria, and Eukarya.

50. Bjorklund AK, Ekman D, Light S, Frey-Skott J, Elofsson A: **Domain rearrangements in protein evolution**. *J Mol Biol* 2005, **353**:911-923.

51. Xie XY, Jin J, Mao YY: **Evolutionary versatility of eukaryotic protein domains revealed by their bigram networks**. *Bmc Evol Biol* 2011, **11**.

52. Basu MK, Carmel L, Rogozin IB, Koonin EV: **Evolution of protein domain promiscuity in eukaryotes**. *Genome Res* 2008, **18**:449-461.

53. Barrera A, Alastruey-Izquierdo A, Martin MJ, Cuesta I, Vizcaino JA: **Analysis of the protein domain and domain architecture content in fungi and its application in the search of new antifungal targets**. *PLoS Comput Biol* 2014, **10**:e1003733.

54. Basu MK, Poliakov E, Rogozin IB: **Domain mobility in proteins: functional and evolutionary implications**. *Brief Bioinform* 2009, **10**:205-216.

55. Hsu CH, Chen CK, Hwang MJ: **The architectural design of networks of protein domain architectures**. *Biol Lett* 2013, **9**:20130268.

56. Andreeva A: **Classification of proteins: available structural space for molecular modeling**. In *Homology Modeling, Methods and Protocols 857.* Edited by Andrew JW, Orry RA. Springer; 2012:1-31.

57. Mi HY, Muruganujan A, Thomas PD: **Panther in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees**. *Nucleic Acids Res* 2013, **41**:D377-D386.

58. Abhiman S, Sonnhammer ELL: **Funshift: a database of function shift analysis on protein subfamilies**. *Nucleic Acids Res* 2005, **33(Database issue)**:D197-D200.

59. Akiva E, Brown S, Almonacid DE, Barber AE, Custer AF, Hicks MA, Huang CC, Lauck F, Mashiyama ST, Meng EC *et al.*: **The structure–function linkage database**. *Nucleic Acids Res* 2014, **42**:D521-D530.

60. Hernandez S, Ferragut G, Amela I, Perez-Pons J, Pinol J, Mozo-Villarias A, Cedano J, Querol E: **Multitaskprotdb: a database of multitasking proteins**. *Nucleic Acids Res* 2014, **42**:D517-D520.

61. Huberts DHEW, van der Klei IJ: **Moonlighting proteins: an intriguing mode of multitasking**. *Bba-Mol Cell Res* 2010, **1803**:520-525.

62. Mani M, Chen C, Amblee V, Liu H, Mathur T, Zwicke G, Zabad S, Patel B, Thakkar J, Jeffery CJ: **Moonprot: a database for proteins that are known to moonlight**. *Nucleic Acids Res* 2015, **43(Database issue)**:D277-D282.

63. Fang H, Gough J: **Dcgo: database of domain-centric ontologies on functions, phenotypes, diseases and more**. *Nucleic Acids Res* 2013, **41(Database issue)**:D536-D544.

64. Radivojac P, Clark WT, Oron TR, Schnoes AM, Wittkop T,
•• Sokolov A, Graim K, Funk C, Verspoor K, Ben-Hur A *et al.*: **A large-scale evaluation of computational protein function prediction**. *Nat Methods* 2013, **10**:221-227.

The first large-scale community-based crucial assessment of protein function annotation experiment, which assessed 54 methods for protein function prediction on a target set of 866 proteins from 11 organisms reveals a great deal of improvement on the latest protein function prediction algorithms, however, there is considerable need for improvement of currently available tools.

65. Swadesh M: **Unaaliq and proto eskimo v: comparative vocabulary**. *Int J Am Linguist* 1952, **18**:241-256.

66. Apic G, Gough J, Teichmann SA: **Domain combinations in archaeal, eubacterial and eukaryotic proteomes**. *J Mol Biol* 2001, **310**:311-325.

67. Zmasek CM, Godzik A: **This deja vu feeling-analysis of
•• multidomain protein evolution in eukaryotic genomes**. *PLoS Comput Biol* 2012, **8**.

The authors compared repertoires of domain architectures of 172 complete eukaryotic genomes and suggest the existence of a core set of domain combinations that keeps reemerging in different species, which are accompanied by a smaller number of unique domain combinations that do not appear anywhere else.

68. Moore AD, Grath S, Schuler A, Huylmans AK, Bornberg-Bauer E: **Quantification and functional analysis of modular protein evolution in a dense phylogenetic tree**. *Biochim Biophys Acta* 2013, **1834**:898-907.

69. Forslund K, Henricson A, Hollich V, Sonnhammer EL: **Domain tree-based analysis of protein architecture evolution**. *Mol Biol Evol* 2008, **25**:254-264.

70. Gough J: **Convergent evolution of domain architectures (is rare)**. *Bioinformatics* 2005, **21**:1464-1471.

71. Suen S, Lu HH, Yeang CH: **Evolution of domain architectures and catalytic functions of enzymes in metabolic systems**. *Genome Biol Evol* 2012, **4**:976-993.

72. Linkeviciute V, Rackham OJ, Gough J, Oates ME, Fang H: **Function-selective domain architecture plasticity potentials in eukaryotic genome evolution**. *Biochimie* 2015 http://dx.doi.org/10.1016/j.biochi.2015.05.003. in press.

73. Vogel C, Teichmann SA, Pereira-Leal J: **The relationship between domain duplication and recombination**. *J Mol Biol* 2005, **346**:355-365.

74. Bornberg-Bauer E, Alba MM: **Dynamics and adaptive benefits
•• of modular protein evolution**. *Curr Opin Struct Biol* 2013, **23**:459-466.

An exceptional review of the mechanisms by which new protein domain arrangements occur and how new domains dynamically spread across genomes possibly as a consequence of their high adaptive potential.

75. Forslund K, Sonnhammer EL: **Evolution of protein domain
•• architectures**. *Methods Mol Biol* 2012, **856**:187-216.

An extensive review on protein domain architecture evolution mechanisms which focuses on phylogenetic studies, studies relating domain family size to occurrence, and studies showing evidence for selective pressure for expansion of certain domain families.