2012-03-01

Proteins: form and function

Roy D. Sleator

BioengineeredBugs

Taylor & Francis
Taylor & Francis Group

Bioengineered

ISSN: 2165-5979 (Print) 2165-5987 (Online) Journal homepage: https://www.tandfonline.com/loi/kbie20

Proteins

Form and function

Roy D. Sleator

To cite this article: Roy D. Sleator (2012) Proteins, Bioengineered, 3:2, 80-85, DOI: 10.4161/bbug.18303

To link to this article: https://doi.org/10.4161/bbug.18303

Published online: 01 Mar 2012.

Bioengineered Bugs 3:2, 80–85; March/April 2012; © 2012 Landes Bioscience

Proteins
Form and function

Roy D. Sleator

Department of Biological Sciences; Cork Institute of Technology; Cork, Ireland

Keywords: protein function, homology-based transfer, ontologies, sequence and structure motifs, evolution, protein domains, gene duplication, divergence, combination, circular permutation

An overwhelming array of structural variants has evolved from a comparatively small number of protein structural domains; which has in turn facilitated an expanse of functional derivatives. Herein, I review the primary mechanisms which have contributed to the vastness of our existing, and expanding, protein repertoires. Protein function prediction strategies, both sequence and structure based, are also discussed and their associated strengths and weaknesses assessed.

Introduction

While Louis Sullivan's assertion that "form follows function" is true of most man-made structures; in protein science the reverse is true—function follows form.

Data from the most recent large scale sequencing projects has facilitated detailed descriptions of the constituent protein repertoires of more than 600 distinct organisms.[1] Taking protein domains (clusters of 50–200 conserved residues) to represent units of evolution, as well as their more usual designation as structural/functional motifs, it is possible to accurately trace the evolutionary relationships of approximately half of these proteins.[2]

Until recently, in the absence of any experimental evidence, homology-based transfer remained the gold standard for ascribing a functional role to such newly identified proteins.[3] Based on this approach, if a query protein shares significant sequence similarity (suggesting a common evolutionary origin) to a protein of known function, then the function of the latter (subject) may be transferred to the former (query protein). However, as the databases continue to expand at an exponential rate, the utility of homology based prediction methods continues to contract, with fewer query proteins registering significant hits to known proteins.

Herein, I review the current knowledge on protein evolution with a specific focus on how gene duplications, sequence divergence and domain combinations have shaped protein evolution. Furthermore, the most recent advances in the field of automated function prediction (AFP) are discussed, along with the future challenges and outstanding questions which still remain to be answered.

Correspondence to: Roy D. Sleator; Email: roy.sleator@cit.ie
Submitted: 09/19/11; Accepted: 10/04/11
http://dx.doi.org/10.4161/bbug.18303

What is Shaping Protein Structure?

Duplication. Of the animal genomes sequenced to date, the proportion of matched domains which are the result of duplications is estimated at between 93% and 97%.[4] Indeed, the hemoglobins, which were the first homologous proteins to have their structure determined, are perhaps the best example of how duplication (and subsequent mutational events) has given rise to subtle structural and functional variations such as oxygen binding profiles.[5] Furthermore, in addition to the generation of whole protein homologs, partial gene duplications resulting in domain duplication and elongation are also common features of protein evolution.[6] In many cases such enlargements have resulted from the addition of sub-domains, variability in loop length and/or changes to the structural core, such as β-sheet extensions. Examples of such protein duplication events include cutinase and bovine bile-salt activated cholesterol esterase (**Fig. 1**). While cutinase is the smallest enzyme of the α/β hydrolases, with five strands in the main β-sheet,[7] bovine bile-salt activated cholesterol esterase has 11 strands and loop structures up to 79 residues in length.[8]

Divergence. There are essentially two types of protein structural divergence: changes to the proteins surface or peripheral regions (e.g., surface loops, surfaces helices and strands on the edges of β-sheets) and the less common but far more detrimental modifications to the proteins interior or core.[9] Indeed, it has been demonstrated that mutations in the protein surface are four times more biologically acceptable than those in the interior.[1] In support of this is the observation that pairs of homologous proteins with identities of approximately 20% have been shown to exhibit up to 50% divergence in the peripheral regions alone.[10]

In addition to subtle changes resulting from missense point mutations leading to single amino acid substitutions and the resulting gradual divergence in structure and function, more radical divergence of structure, mediated by domain shuffling (recombination or permutation) has also been reported.[11] Circular permutations (CPs) in particular represent a specific form of recombination event which is characterized by the presence of the same protein sub-sequences in the same linear order but different positions of the N and C termini,[12] in essence CP of a protein can be visualized as if its original termini were linked and new ones created elsewhere (**Fig. 2**). First observed in plant lectins,[14] a substantial number of natural examples of CP have been reported;

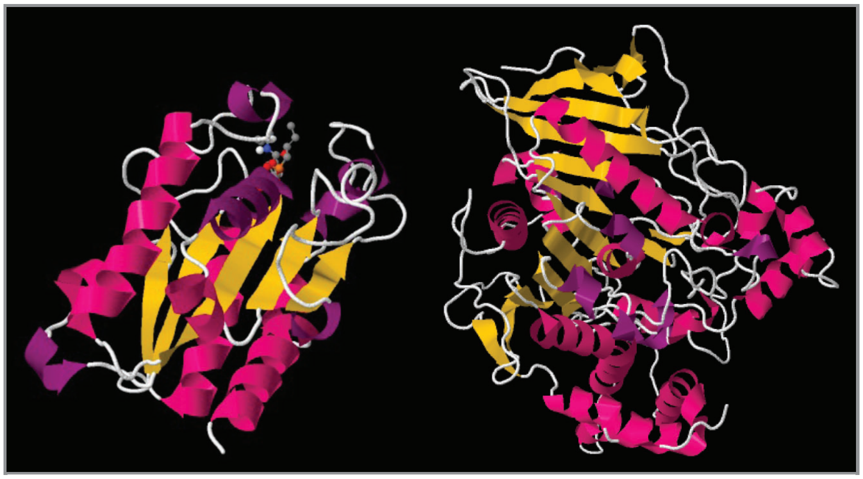

Figure 1. Protein duplication. Partial gene duplications resulting in domain duplication and elongation are common features of protein evolution. An example of such a protein duplication event is observed between cutinase and bovine bile-salt activated cholesterol esterase. While cutinase (A) is the smallest enzyme of the α/β hydrolases, with five strands in the main β-sheet, bovine bile-salt activated cholesterol esterase (B) has 11 strands, and loop structures up to 79 residues in length.

indeed, some 120 protein clusters which appear to have segments of their sequences in different sequential order are reported in the Circular Permutation Database.[15] In addition to natural evolutionary processes, artificial CPs have been engineered in an effort to study protein folding properties as well as the design of more efficient enzymes.[16] A circularly permuted streptavidin for example has been designed to remove the flexible polypeptide loop that undergoes an open to closed conformational change when biotin is bound. The original termini have been joined by a tetrapeptide linker, and four loop residues have been removed, resulting in the creation of new N- and C-termini.[17]

While domain shuffling may have dramatic effects on protein structure, protein homologs usually conserve their catalytic mechanisms, i.e., the relative positions of their functional active sites or catalytic residues may shift but they retain their functional activity. This usually occurs when divergence induces structural changes in the catalytic region, thus necessitating a reconfiguration of the position of the catalytic residues in order to maintain function.[18] In several cases, while the functionally equivalent residues are located at non-homologous positions on the protein's 3D structure, the catalytic residues themselves are identical. An example of this is chloramphenicol acetyltransferase (PaXAT) and UDP-*N*-acetylglucosamine acyltransferase (LpxA) both of which contain an essential histidine residue thought to be involved in deprotonation of a hydroxyl group in their individual substrates. However, these residues are located at different points within the protein fold; in LpxA, the histidine is located in the core of the domain,[19] whereas in PaXAT, it occurs in a loop extending from the solenoid structure.

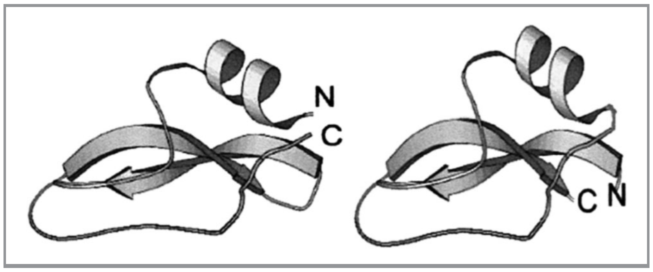

Figure 2. A schematic example of a circular permutation. The original termini (left) are fused to form a continuous part of the chain and new termini are formed by cutting the polypeptide chain elsewhere (right). Reproduced with permission from Uliel et al.[13]

Thus, two proteins may have quite divergent structures and/or sequences while retaining similar function; such proteins are said to be functional analogs. Such analogs may also arise as a result of convergent evolution; that is, they do not diverge from a common ancestor but instead arise independently and converge on the same active configuration as a result of natural selection for a particular biochemical function. L-aspartate aminotransferase and D-amino acid aminotransferase provide excellent examples of convergently evolved functional analogs. Despite having a strikingly similar arrangement of residues in their active sites, the two proteins have completely different architectures; differing in size, amino acid sequence and in the fold of the protein domains.

Conversely, certain proteins share significant sequence and/or structure similarity but differ in terms of substrate specificity or indeed catalytic function. An example of such structural analogs, which arise by means of divergent evolution from a single ancestor, include Human IL-10 (hIL-10); a cytokine that modulates diverse immune responses and the Epstein-Barr virus (EBV) IL-10 homolog (vIL-10). Although vIL-10 suppresses inflammatory responses like hIL-10, it cannot activate many other immune-stimulatory functions performed by the cellular cytokine.[20]

Combination. While the evolutionary impact of duplication and divergence on protein sequence, structure and function is obvious, multi-domain proteins are, for the most part, the result of gene combinations.[21] Such combinations can give rise to domain recruitment and enlargement and can significantly affect both protein structure/stability and function. For example, in the case of domain recruitment the addition of an accessory domain may affect protein function by modulating substrate selectivity; achieved either by the addition of a binding site, or, by playing a purely structural role, shaping the existing active site to accommodate substrates of different shapes and/or sizes.[18] For example, prokaryotic methionine aminopeptidase exists as a monomeric single-domain protein while creatinase, is a two-domain protein. The additional domain of the second subunit of creatinase caps the active site allowing the binding of the small molecule creatine.[22]

What is Protein Function?

Before commencing any discussion on protein function prediction we must first consider what is meant by "function." Biological function is highly contextual; different aspects of the function of a given protein may be viewed as occurring in different scales of space and time from the almost instantaneous enzymatic reactions to the much slower overall biological process.[23] Knowing which functional aspect is being investigated is thus extremely important and can only properly be achieved by the establishment of a standardized machine readable vocabulary.

Fortunately, significant progress has been made in the computer science arena in developing the theory and application of structured machine readable vocabularies, known as ontologies, which provide a formal explicit specification of a commonly used abstract model of the world.[24] Ontologies not only allow formal definition of concepts, but also enable the creation of software tools capable of reasoning about the properties and relationships of a domain. Formats such as the Resource Description Framework (RDF) and the Web Ontology Language (OWL) have been devised that allow ontological concepts to be persisted and communicated. RDF, for example, allows the creation of statements about a particular domain by the use of triples in the form of subject-predicate-object expressions. The subject and object represents a concept, whereas the predicate defines the relationship between them.

Detailed ontologies can be created by composing further defining concepts and relationships that model the domain of interest. Ontologies that define different aspects of proteins could be used to annotate biological data with functional facets and provide the basis of a framework for machine based reasoning.

The Gene Ontology (GO)[25] goes some way to achieving this goal, formulizing a definition of functional context and providing machine—legible functional annotation. GO has three "ontology trees" describing three aspects of gene product function: Molecular function, biological process and cellular location. By providing a standard vocabulary and defining relationships between terms, annotations can be computationally processed,[26] thus providing a standard approach for programs to output their functional predictions.

Having defined biological "function" and the means of describing such functions we can now turn our attention to the various function prediction programs, and their associated strengths and weaknesses.

Protein Function Prediction Methods

Protein function prediction methods can be loosely divided into sequence and structure based approaches. Herein, I outline the current state of the art for sequence and structure based protein function prediction.

Sequence based approaches. *Homology-based transfer.* Homology-based transfer, using programs such as BLAST,[27] is perhaps the most widely used form of computational function prediction method; assigning un-annotated proteins with the function of their annotated homologs. The rationale for this approach is based on the assumption that two sequences with a high degree of similarity most likely evolved from a common ancestor and thus must have similar functions.

While sequence similarity is undoubtedly correlated with functional similarity, exceptions have been observed on both ends of the similarity scale. Rost,[28] for example, showed that even at high sequence similarity rates, enzymatic function may not necessarily be conserved, while Galperin et al.[29] observed that enzymes that are analogous on the basis of sequence dissimilarity are in fact homologous. While such errors are the exception rather than the rule, they may set the seed for further annotation errors; as more sequences enter the databases, more are annotated by homology-based transfer, thus helping to propagate and amplify the original single erroneous annotation.[30,31]

Furthermore, as the databases continue to expand the utility of the homology-based transfer approach begins to beak-down. The recent explosion of large scale metagenomic sequencing projects[32] has resulted in an unprecedented amount of novel sequence being deposited in the databases. As a direct consequence of this sequence expansion, the number of clustered similar proteins for which no single annotated reference sequence exists is expanding rapidly, eroding the foundations of the homology-based transfer approach. Indeed, it has been estimated that $< 35\%$ of all proteins could be annotated automatically when accepting errors of $\leq 5\%$, while even allowing for error rates of $> 40\%$ there is no annotation for $> 30\%$ of all proteins.[33]

Sequence Motifs. Typically of the 100–300 amino acids in a functional protein domain $< 10\%$ constitute the protein's active sites.[34] Therefore, homology-based transfer from a complete protein is often not necessary to predict a protein's function. All that is required is a sequence (or structure) based signature which is associated with a particular function. Such signatures may occur at a single position on the sequence or as a "fingerprint" composed of several such patters. A few databases are dedicated to motif searching; PROSITE[35] for example is composed of manually selected biologically important motifs and has three types of signatures: patterns, rules and profiles. Each signature represents a different automated method for searching motifs; while patterns and rules typically span only a few residues (e.g., A typical entry in PROSITE would be [ST]-x(2)-[DE], i.e. a serine or threonine, followed by any two residues, followed by aspartate or glutamate—the consensus sequence of a Casein kinase II phosphorylation site) profiles extend the similarity to the level of entire domains. Other well-known motif databases include BLOCKS[36] and PRINTS.[37]

Genomic context and expression based prediction methods. Genomic context based prediction, also referred to as phylogenomic profiling is a method for predicting protein function based on the observation that proteins with similar pedigrees (inter-genomic profiles) are believed to have evolved in tandem and as such are likely to share a common function.[38] Furthermore, in prokaryote genomes the loci of functionally related proteins tend to be co-located on the chromosome (**Fig. 3A**). Combining co-evolution and co-location (chromosomal proximity) has given rise to a new generation of function-prediction algorithms such as Phydbac2.[44]

As an extension of co-location, genes involved in similar cellular functions also tend to be co-transcribed (**Fig. 3B**). Following this logic unknown genes co-expressed with known genes may be functionally annotated by virtue of association. This "guilt by association" approach has given rise to an algorithm of the same name, developed by Walker et al.,[45] for the analysis of gene expression arrays. Unlike the sequence motif based approach which focuses on molecular function annotation; expression microarray based predictions are useful for annotation of the cellular aspect of protein function. Furthermore, given that most cellular processes are performed by groups of physically interacting proteins, it is fair to assume that such interacting proteins have similar overall cellular functions. Thus, protein-protein interaction (PPI) data may also facilitate protein function annotation and several PPI databases are now available such as STRING—a database of known and predicted PPIs.[46]

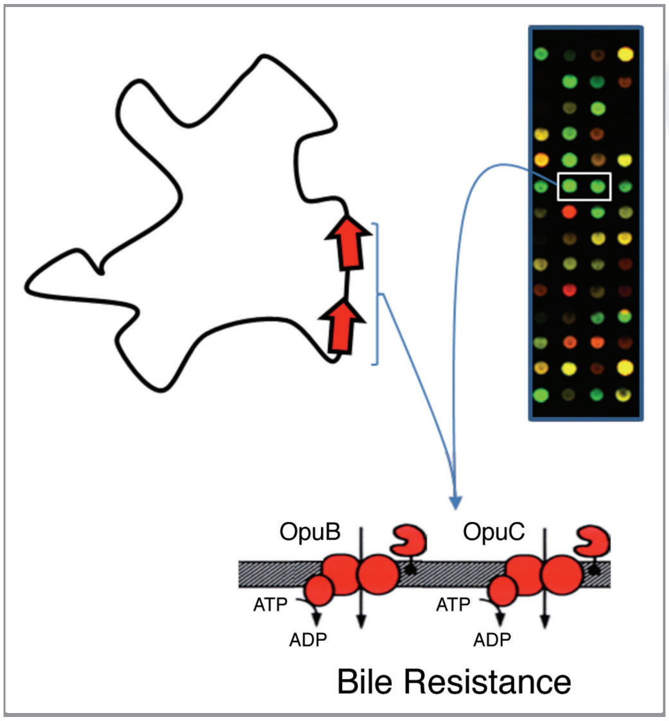

Figure 3. Genomic context based prediction. In prokaryote genomes the loci of functionally related proteins tend to be co-located on the chromosome; an example of this are the membrane bound ABC transport proteins OpuC and OpuB(BilE) of *Listeria monocytogenes* which are separated by only 2.4 Kb[39] on the listerial chromosome and contribute to bile resistance in this gastrointestinal pathogen.[40-43] As an extension of chromosomal proximity, genes involved in similar cellular functions also tend to be co-transcribed as has also been shown to be the case with *opuC* and *opuB*(*bilE*).[40]

Structure based approaches. Given that protein structure is far more conserved than sequence, many proteins which exhibit little or no sequence similarities, due to evolutionary constraints still retain significant structure similarity.[47] In this respect structure is a useful indicator of function; indeed most known protein folds are associated with a particular function or functional milieu.[18] Programs that scan the Protein Data Bank (PDB) for structural similarity given a query sequence include, among others, FATCAT,[48] PAST[49] and VAST.[50] However, knowledge of 3D protein structure alone is not always sufficient to accurately infer function. Indeed, it is estimated that functional hypotheses can be made from 3D structures for only ~20–50% of hypothetical proteins.[51,52]

Rather than focusing on the protein as a whole, it is possible, and in some instances more desirable, to target 3D motifs associated with specific functions (e.g., binding sites or active sites). The rational for analyzing structure motifs (or patterns) is analogous to that of sequence patterns—to identify unique signatures indicative of a particular function. Libraries of 3D motifs with known function have begun to evolve,[53] one example of which is PROCAT,[54] a database of 3D enzyme active sites that can be queried for specific functional signatures. In addition, hybrid motifs incorporating information from sequence and structure, as well as from the literature, have also been used to predict protein function.[55]

Conclusions and Future Prospects

Herein, I have discussed how mechanisms such as gene duplication, sequence divergence and domain combinations[56] have shaped protein evolution and how the retention of sequence and/or structural domains has facilitated the tracking of this evolutionary process through the millennia. I have also introduced the far more complex issue of protein function elucidation wherein, in contrast to protein structure in which the data are either known or easily predicted, the multifaceted and ambiguous nature of biological function makes its elucidation a far more complex endeavor. The complexity of the problem is perhaps best illustrated by Jeffrey's[57] so called "moonlighting proteins" which perform several contextually different functions, ranging from the molecular to the cellular level. Thus, given the aggregate nature of protein function prediction, perhaps the best outcome will be achieved by adopting a multifaceted approach. For example, while biochemical function prediction is likely best served by focusing on sequence motifs, resolution of physiological function is better addressed at the genomic level, based for example on microarray expression data. Therefore, composite methods, employing a diversity of features to assess different functional aspects, are most likely to succeed. Examples of such aggregate functional prediction programs include InterPro, ProKnow and ProFunc, which utilize several data sources and/or algorithms to predict function.

However, despite the emergence of ever more sophisticated and versatile function prediction algorithms; the proper assessment of such programs still remains a significant limitation to the development of the field. Unlike assessment of protein structure, function prediction methods still lack a viable blind benchmark for which to assess program efficacy. This obstacle may eventually be overcome by emulating successful collaborative efforts of computational and experimental structural biologists in the form of CASP (Critical Assessment of Structure Prediction) for the benchmarking of protein structure.

Note

R.D.S. is an ESCMID Research Fellow. This article is based on a Chapter entitled *Prediction of Protein Functions* in *Functional Genomics: Methods and Protocols*, edited by Michael Kaufmann and Claudia Klinger.

References

1. Chothia C, Gough J. Genomic and structural aspects of protein evolution. Biochem J 2009; 419:15-28; PMID:19272021; http://dx.doi.org/10.1042/BJ20090122
2. Sleator RD. An overview of the processes shaping protein evolution. Sci Prog 2010; 93:1-6; PMID:20222353; http://dx.doi.org/10.3184/003685009X12605492662844
3. Sleator RD, Walsh P. An overview of in silico protein function prediction. Arch Microbiol 2010; 192:151-5; PMID:20127480; http://dx.doi.org/10.1007/s00203-010-0549-9
4. Wilson D, Pethica R, Zhou Y, Talbot C, Vogel C, Madera M, et al. SUPERFAMILY–sophisticated comparative genomics, data mining, visualization and phylogeny. Nucleic Acids Res 2009; 37:D380-6; PMID:19036790; http://dx.doi.org/10.1093/nar/gkn762
5. Blanchetot A, Wilson V, Wood D, Jeffreys AJ. The seal myoglobin gene: an unusually long globin gene. Nature 1983; 301:732-4; PMID:6828156; http://dx.doi.org/10.1038/301732a0
6. Moore AD, Bjorklund AK, Ekman D, Bornberg-Bauer E, Elofsson A. Arrangements in the modular evolution of proteins. Trends Biochem Sci 2008; 33:444-51; PMID:18656364; http://dx.doi.org/10.1016/j.tibs.2008.05.008
7. Longhi S, Czjzek M, Lamzin V, Nicolas A, Cambillau C. Atomic resolution (1.0 A) crystal structure of Fusarium solani cutinase: stereochemical analysis. J Mol Biol 1997; 268:779-99; PMID:9175860; http://dx.doi.org/10.1006/jmbi.1997.1000
8. Chen JC, Miercke LJ, Krucinski J, Starr JR, Saenz G, Wang X, et al. Structure of bovine pancreatic cholesterol esterase at 1.6 A: novel structural features involved in lipase activation. Biochemistry 1998; 37:5107-17; PMID:9548741; http://dx.doi.org/10.1021/bi972989g
9. Gerstein M, Sonnhammer EL, Chothia C. Volume changes in protein evolution. J Mol Biol 1994; 236:1067-78; PMID:8120887; http://dx.doi.org/10.1016/0022-2836(94)90012-4
10. Chothia C, Lesk AM. The relation between the divergence of sequence and structure in proteins. EMBO J 1986; 5:823-6; PMID:3709526
11. Kawashima T, Kawashima S, Tanaka C, Murai M, Yoneda M, Putnum NH, Rokhsar DS, Kanehisa M, Satoh N, Wada H. Domain shuffling and the evolution of vertebrates. Genome Res 2009; 19:1393-403; PMID:19443856; http://dx.doi.org/10.1101/gr.087072.108
12. Vogel C, Morea V. Duplication, divergence and formation of novel protein topologies. Bioessays 2006; 28:973-8; PMID:16998824; http://dx.doi.org/10.1002/bies.20474
13. Uliel S, Fliess A, Unger R. Naturally occurring circular permutations in proteins. Protein Eng 2001; 14:533-42; PMID:11579221; http://dx.doi.org/10.1093/protein/14.8.533
14. Lindqvist Y, Schneider G. Circular permutations of natural protein sequences: structural evidence. Curr Opin Struct Biol 1997; 7:422-7; PMID:9204286; http://dx.doi.org/10.1016/S0959-440X(97)80061-9
15. Lo WC, Lee CC, Lee CY, Lyu PC. CPDB: a database of circular permutation in proteins. Nucleic Acids Res 2009; 37:D328-32; PMID:18842637; http://dx.doi.org/10.1093/nar/gkn679
16. Heinemann U, Ay J, Gaiser O, Muller JJ, Ponnuswamy MN. Enzymology and folding of natural and engineered bacterial beta-glucanases studied by X-ray crystallography. Biol Chem 1996; 377:447-54; PMID:8922278
17. Chu V, Freitag S, Le Trong I, Stenkamp RE, Stayton PS. Thermodynamic and structural consequences of flexible loop deletion by circular permutation in the streptavidin-biotin system. Protein Sci 1998; 7:848-59; PMID:9568892; http://dx.doi.org/10.1002/pro.5560070403
18. Todd AE, Orengo CA, Thornton JM. Evolution of function in protein superfamilies, from a structural perspective. J Mol Biol 2001; 307:1113-43; PMID:11286560; http://dx.doi.org/10.1006/jmbi.2001.4513
19. Wyckoff TJ, Raetz CR. The active site of Escherichia coli UDP-N-acetylglucosamine acyltransferase. Chemical modification and site-directed mutagenesis. J Biol Chem 1999; 274:27047-55; PMID:10480918; http://dx.doi.org/10.1074/jbc.274.38.27047
20. Yoon SI, Jones BC, Logsdon NJ, Walter MR. Same structure, different function crystal structure of the Epstein-Barr virus IL-10 bound to the soluble IL-10R1 chain. Structure 2005; 13:551-64; PMID:15837194; http://dx.doi.org/10.1016/j.str.2005.01.016
21. Apic G, Gough J, Teichmann SA. Domain combinations in archaeal, eubacterial and eukaryotic proteomes. J Mol Biol 2001; 310:311-25; PMID:11428892; http://dx.doi.org/10.1006/jmbi.2001.4776
22. Hoeffken HW, Knof SH, Bartlett PA, Huber R, Moellering H, Schumacher G. Crystal structure determination, refinement and molecular model of creatine amidinohydrolase from Pseudomonas putida. J Mol Biol 1988; 204:417-33; PMID:3221393; http://dx.doi.org/10.1016/0022-2836(88)90586-4
23. Godzik A, Jambon M, Friedberg I. Computational protein function prediction: are we making progress? Cell Mol Life Sci 2007; 64:2505-11; PMID:17611711; http://dx.doi.org/10.1007/s00018-007-7211-y
24. Losko S, Heumann K. Semantic data integration and knowledge management to represent biological network associations. Methods Mol Biol 2009; 563:241-58; PMID:19597789; http://dx.doi.org/10.1007/978-1-60761-175-2_13
25. Ashburner M, Lewis S.. On ontologies for biologists: the Gene Ontology–untangling the web. Novartis Found Symp 2002; 247:66-80; discussion -3, 4-90, 244-52. PMID:12539950; http://dx.doi.org/10.1002/0470857897.ch6
26. Smith CL, Goldsmith CA, Eppig JT. The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information. Genome Biol 2005; 6:R7; PMID:15642099; http://dx.doi.org/10.1186/gb-2004-6-1-r7
27. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997; 25:3389-402; PMID:9254694; http://dx.doi.org/10.1093/nar/25.17.3389
28. Rost B. Enzyme function less conserved than anticipated. J Mol Biol 2002; 318:595-608; PMID:12051862; http://dx.doi.org/10.1016/S0022-2836(02)00016-5
29. Galperin MY, Walker DR, Koonin EV. Analogous enzymes: independent inventions in enzyme evolution. Genome Res 1998; 8:779-90; PMID:9724324
30. Bork P. Powers and pitfalls in sequence analysis: the 70% hurdle. Genome Res 2000; 10:398-400; PMID:10779480; http://dx.doi.org/10.1101/gr.10.4.398
31. Gilks WR, Audit B, de Angelis D, Tsoka S, Ouzounis CA. Percolation of annotation errors through hierarchically structured protein sequence databases. Math Biosci 2005; 193:223-34; PMID:15748731; http://dx.doi.org/10.1016/j.mbs.2004.08.001
32. Sleator RD, Shortall C, Hill C. Metagenomics. Lett Appl Microbiol 2008; 47:361-6; PMID:19146522; http://dx.doi.org/10.1111/j.1472-765X.2008.02444.x
33. Rost B, Liu J, Nair R, Wrzeszczynski KO, Ofran Y. Automatic prediction of protein function. Cell Mol Life Sci 2003; 60:2637-50; PMID:14685688; http://dx.doi.org/10.1007/s00018-003-3114-8
34. Friedberg I. Automated protein function prediction–the genomic challenge. Brief Bioinform 2006; 7:225-42; PMID:16772267; http://dx.doi.org/10.1093/bib/bbl004
35. Hulo N, Bairoch A, Bulliard V, Cerutti L, Cuche BA, de Castro E, et al. The 20 years of PROSITE. Nucleic Acids Res 2008; 36:D245-9; PMID:18003654; http://dx.doi.org/10.1093/nar/gkm977
36. Henikoff JG, Greene EA, Pietrokovski S, Henikoff S. Increased coverage of protein families with the blocks database servers. Nucleic Acids Res 2000; 28:228-30; PMID:10592233; http://dx.doi.org/10.1093/nar/28.1.228
37. Attwood TK, Bradley P, Flower DR, Gaulton A, Maudling N, Mitchell AL, et al. PRINTS and its automatic supplement, prePRINTS. Nucleic Acids Res 2003; 31:400-2; PMID:12520033; http://dx.doi.org/10.1093/nar/gkg030
38. Eisenberg D, Marcotte EM, Xenarios I, Yeates TO. Protein function in the post-genomic era. Nature 2000; 405:823-6; PMID:10866208; http://dx.doi.org/10.1038/35015694
39. Sleator RD, Gahan CG, Hill C. A postgenomic appraisal of osmotolerance in Listeria monocytogenes. Appl Environ Microbiol 2003; 69:1-9; PMID:12513970; http://dx.doi.org/10.1128/AEM.69.1.1-9.2003
40. Sleator RD, Watson D, Hill C, Gahan CG. The interaction between Listeria monocytogenes and the host gastrointestinal tract. Microbiology 2009; 155:2463-75; PMID:19542009; http://dx.doi.org/10.1099/mic.0.030205-0
41. Watson D, Sleator RD, Casey PG, Hill C, Gahan CG. Specific osmolyte transporters mediate bile tolerance in Listeria monocytogenes. Infect Immun 2009; 77:4895-904; PMID:19737907; http://dx.doi.org/10.1128/IAI.00153-09
42. Sleator RD, Hill C. Compatible solutes: A listerial passe-partout? Gut Microbes 2010; 1:77-9; PMID:21326913; http://dx.doi.org/10.4161/gmic.1.2.10968
43. Sleator RD, Hill C. Compatible solutes: the key to Listeria's success as a versatile gastrointestinal pathogen? Gut Pathog 2010; 2:20; PMID:21143981; http://dx.doi.org/10.1186/1757-4749-2-20

44. Enault F, Suhre K, Claverie JM. Phydbac "Gene Function Predictor": a gene annotation tool based on genomic context analysis. BMC Bioinformatics 2005; 6:247; PMID:16221304; http://dx.doi.org/10.1186/1471-2105-6-247

45. Walker MG, Volkmuth W, Sprinzak E, Hodgson D, Klingler T. Prediction of gene function by genome-scale expression analysis: prostate cancer-associated genes. Genome Res 1999; 9:1198-203; PMID:10613842; http://dx.doi.org/10.1101/gr.9.12.1198

46. Zhao XM, Chen L, Aihara K. Protein function prediction with high-throughput data. Amino Acids 2008; 35:517-30; PMID:18427717; http://dx.doi.org/10.1007/s00726-008-0077-y

47. Watson JD, Laskowski RA, Thornton JM. Predicting protein function from sequence and structural data. Curr Opin Struct Biol 2005; 15:275-84; PMID:15963890; http://dx.doi.org/10.1016/j.sbi.2005.04.003

48. Ye Y, Godzik A. FATCAT: a web server for flexible structure comparison and structure similarity searching. Nucleic Acids Res 2004; 32:W582-5; PMID:15215455; http://dx.doi.org/10.1093/nar/gkh430

49. Täubig H, Buchner A, Griebsch J. PAST: fast structure-based searching in the PDB. Nucleic Acids Res 2006; 34:W20-3; PMID:16844992; http://dx.doi.org/10.1093/nar/gkl273

50. Gibrat JF, Madej T, Bryant SH. Surprising similarities in structure comparison. Curr Opin Struct Biol 1996; 6:377-85; PMID:8804824; http://dx.doi.org/10.1016/S0959-440X(96)80058-3

51. Laskowski RA, Watson JD, Thornton JM. From protein structure to biochemical function? J Struct Funct Genomics 2003; 4:167-77; PMID:14649301; http://dx.doi.org/10.1023/A:1026127927612

52. Goldsmith-Fischman S, Honig B. Structural genomics: computational methods for structure analysis. Protein Sci 2003; 12:1813-21; PMID:12930981; http://dx.doi.org/10.1110/ps.0242903

53. Jones S, Thornton JM. Searching for functional sites in protein structures. Curr Opin Chem Biol 2004; 8:3-7; PMID:15036149; http://dx.doi.org/10.1016/j.cbpa.2003.11.001

54. Wallace AC, Laskowski RA, Thornton JM. Derivation of 3D coordinate templates for searching structural databases: application to Ser-His-Asp catalytic triads in the serine proteinases and lipases. Protein Sci 1996; 5:1001-13; PMID:8762132; http://dx.doi.org/10.1002/pro.5560050603

55. Di Gennaro JA, Siew N, Hoffman BT, Zhang L, Skolnick J, Neilson LI, et al. Enhanced functional annotation of protein sequences via the use of structural descriptors. J Struct Biol 2001; 134:232-45; PMID:11551182; http://dx.doi.org/10.1006/jsbi.2001.4391

56. Chothia C, Gough J, Vogel C, Teichmann SA. Evolution of the protein repertoire. Science 2003; 300:1701-3; PMID:12805536; http://dx.doi.org/10.1126/science.1085371

57. Jeffery CJ. Moonlighting proteins: old proteins learning new tricks. Trends Genet 2003; 19:415-7; PMID:12902157; http://dx.doi.org/10.1016/S0168-9525(03)00167-7