Chapter 5

Computational Methods for Annotation Transfers from Sequence

Domenico Cozzetto and David T. Jones

Abstract

Surveys of public sequence resources show that experimentally supported functional information is still completely missing for a considerable fraction of known proteins and is clearly incomplete for an even larger portion. Bioinformatics methods have long made use of very diverse data sources alone or in combination to predict protein function, with the understanding that different data types help elucidate complementary biological roles. This chapter focuses on methods accepting amino acid sequences as input and producing GO term assignments directly as outputs; the relevant biological and computational concepts are presented along with the advantages and limitations of individual approaches.

Key words Protein function prediction, Homology-based annotation transfers, Phylogenomics, Multi-domain architecture, De novo function prediction

1 Introduction

For decades experimentalists have been painstakingly probing a range of functional aspects of individual proteins. This steady but slow acquisition of functional data is in stark contrast to the results of next-generation sequencing technologies, which can survey gene expression regulation, genomic organization, and variation on a large scale [1]. Similarly, parallel efforts aim to map the networks of interactions between proteins, nucleic acids, and metabolites that regulate biological processes [2–4]. Nonetheless, comprehensive studies of protein function are hindered, because the combinations of gene products, biological roles, and cellular conditions are too numerous and because many experimental protocols cannot be applied to all proteins. Furthermore, the results need to be critically interpreted, integrated with existing knowledge, and translated into machine-readable formats—such as Gene Ontology (GO) [5] terms—for further analyses.

Christophe Dessimoz and Nives Škunca (eds.), *The Gene Ontology Handbook*, Methods in Molecular Biology, vol. 1446,
DOI 10.1007/978-1-4939-3743-1_5, © The Author(s) 2017

Manual curation requires substantial time and effort too; therefore the exponential growth in the number of sequences in UniProtKB [6] has only been matched by a linear increase in the number of entries with experimentally supported GO terms. Moreover, only 0.03 % of the sequences have received annotations for all three GO domains and the level of annotation detail can also fall far short of the maximum possible—e.g., there is direct evidence that some *E. coli K12* proteins act as transferases with no additional information about the chemical group relocated from the donor to the acceptor. Automated protein function prediction has consequently represented the only viable way to bridge some of these gaps, and indeed UniProtKB already exploits some computational tools (Fig. 1).

Given the lack of a general theory which can link protein sequences and environmental conditions directly to biological functions from physicochemical properties, current methods for protein function prediction implement knowledge-based heuristics that transfer functional information from already annotated proteins to unannotated ones. This chapter reviews sequence-based approaches to GO term prediction, which are the most popular, well understood, and easily accessible to a wide range of users. The

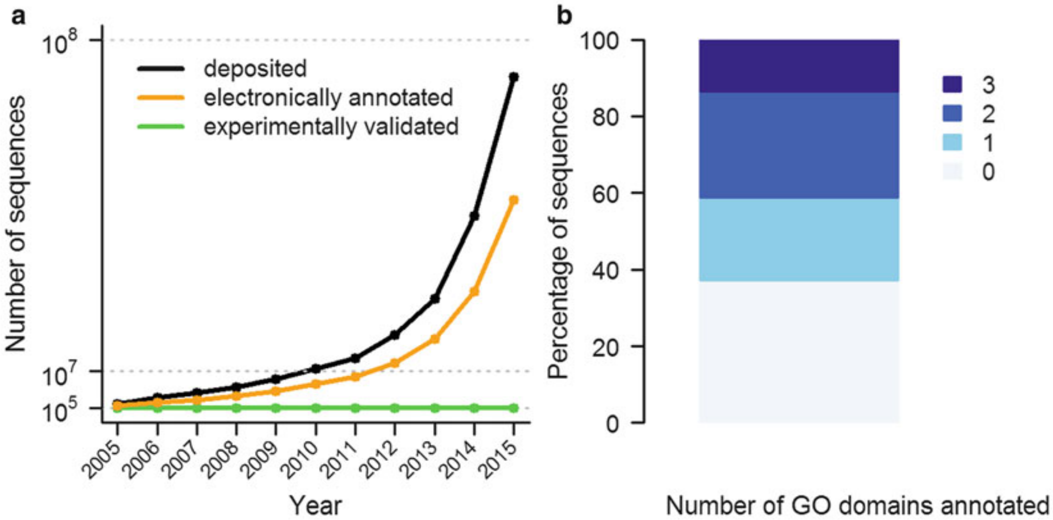

Fig. 1 Function annotation coverage of proteins in UniprotKB. (**a**) Over the past decade, the number of amino acid chains deposited in UniProtKB has grown exponentially (*black line*), while those with experimentally supported GO term assignments has only increased linearly (*green line*). This core subset however has allowed to assign GO terms to a substantial fraction of sequences (*orange line*). (**b**) Even with electronically inferred annotations, more than 80 % of sequences in UniProtKB release 2015_01 lack assignments for at least one of the molecular function, biological process, or cellular component GO sub-ontologies. Plots and statistics are based on the first release of each year

focus is primarily on the underpinning concepts and assumptions, as well as on the known advantages and pitfalls, which are all applicable to other controlled vocabularies, such as those described in the Chap. 19 [7] "KEGG, EC and other sources of functional data". How well current function prediction methods perform and how prediction accuracy can be measured are topics extensively covered in the Chap. 8 [8] "Evaluating GO annotations", Chap. 9 [9] "Evaluating functional annotations in enzymes", and Chap. 10 [10] "Community Assessment".

2 Annotation Transfers from Homologous Proteins

The most common way to annotate uncharacterized proteins consists in finding homologues—that is, proteins sharing common ancestry—of known function, and inheriting the information available for them under the assumption that function is evolutionarily conserved. BLAST [11] or PSI-BLAST [12] are routinely used to search for homologous sequences, and tools that compare sequences against hidden Markov models (HMMs), or pairs of profiles or of HMMs can be useful to extend the coverage of the protein sequence universe thanks to the increased sensitivity for remote homologues. A detailed presentation of sequence comparison methods is beyond the scope of this chapter and is available elsewhere [13]. In the simplest case, transfers can be made from the sequence with experimentally validated annotations and the lowest *E*-value—and this represents a useful baseline to benchmark the effectiveness of more advanced methods. This approach can produce erroneous results when key functional residues are mutated, or when the alignment doesn't span the whole length of the proteins—possibly indicating changes in domain architecture [14]. Iterative transfers of computationally generated functional assignments can lead to uncontrolled propagation of such errors; the average error rate of molecular function annotations is estimated to approach 0 % only in the manually curated UniProtKB/SwissProt database, while it is substantially higher in un-reviewed resources [15].

Several studies have consequently attempted to estimate sequence similarity thresholds that would generate predictions with a guaranteed level of accuracy, and have suggested that 80 % global sequence identity should be generally sufficient for safe annotation transfers [16–20]. However, this rule of thumb can either be too stringent or too lax, because biological sequences evolve at differing rates due to the need to maintain physiological function on the one hand, and to avoid deregulated gene expression, protein translation, folding, or physical interactions on the other [21]. Ideally, these cutoff values should be specific to individual families or even functional categories, but usually the number of labelled examples is not sufficient to allow reliable calibration. To circumvent these issues, it is possible to trade annotation specificity for accuracy, because broad functional aspects—e.g., about ligand binding and enzymatic or transporter activities—diverge at lower rates than the fine details—such as the specific metal ions bound or the molecules and chemical groups that are recognized and processed.

GOtcha [22] was the first tool to make predictions representing the enrichment of the GO terms assigned to BLAST hits in the hierarchical context of GO. It first calculates weights for each GO term, taking into account the number of similar sequences annotated with it and the statistical significance of the observed similarities. The program then considers the semantic relationships among the terms to update the tallies and reflect increasing confidence in more general annotations. PFP [23] follows a similar approach, but targets more difficult annotation cases, too, by leveraging information from PSI-BLAST hits with unconventionally high E-values. Furthermore, the scoring scheme exploits data about the co-occurrence of GO term pairs in UniProtKB entries, which allows safer annotations to be produced. Other methods fall in this category too, and interested readers are referred to the primary literature [24–27]. More sophisticated approaches rely on machine learning [28] rather than statistical analyses, and use experimental data to train classifiers that predict GO terms based on an array of alignment-derived features—such as sequence similarity scores, E-values, the coverage of the sequences, or the scores that GOtcha calculates for each GO category [29–31].

3 Annotation Transfers from Orthologous Proteins

Simple homology-based predictors are quick but error prone because they don’t try to distinguish functionally equivalent relatives from those that have functionally diverged. In phylogenetic terms, this problem can be cast as classifying orthologues—homologue pairs evolved after speciation—and paralogues—homologue pairs derived from gene duplication. It is widely accepted that duplicated genes lack selective pressure to maintain their original biological roles, so they can easily undergo nucleotide changes ultimately leading to functional divergence [32]. The realization that genetic diversity arises from gene losses and horizontal transfers, too, makes phylogenetic reconstruction even more complex.

In this setup, annotations can be transferred with varying levels of confidence depending on how many orthologues there are and how closely related they are. This can partly account for the observation that orthologues can diverge functionally, particularly over long evolutionary distances or after duplication events in at least one of the lineages [33]. However, experimental studies have also shown that paralogues can retain functional equivalence, even long after the duplication event [34, 35]. Recent studies have consequently tested how useful the distinction between orthologues and paralogues is for protein function prediction and have drawn different conclusions [36–39]. The latest findings suggest that the functional similarity between orthologues is slightly higher than that between paralogues at the same level of sequence divergence, and that the signal is stronger for cellular components than for biological processes or molecular functions [38].

The traditional approach to orthologue detection involves computationally intensive calculations to build phylogenetic trees and then identify gene duplication and loss events [40]. SIFTER [41] builds on this framework to transfer the most specific experimentally supported molecular function terms available from the annotated sequences to all nodes in the tree using a Bayesian approach. The propagation algorithm captures the notion that functional transitions are more likely to occur after duplication than after speciation events, and when the terms are similar—i.e., the corresponding nodes are close in the GO graph. In order to speed up the computation, the authors have recently suggested limiting the number of GO term annotations that can be assigned to each protein [42], and they are providing pre-calculated predictions for a vast set of sequences from different species, including multi-domain proteins [43]. The semiautomated Phylogenetic Annotation and Inference Tool (PAINT) [44] recently adopted by the GO consortium provides a more flexible framework, which tries to keep functional change events uncoupled, so that the gain of one function does not imply the loss of another and vice versa—a desirable feature for annotating biological processes and for dealing with multifunctional proteins in general. Furthermore, unlike SIFTER, PAINT makes no assumption about how function diverges over evolutionary distance and whether its conservation is higher within orthologous groups than between them.

The increasing availability of completely sequenced genomes has promoted the development of alternative algorithms for orthologue detection. These first categorize pairs of orthologues in any two species, and then cluster the results across organisms, which helps recognize and fix spurious assignments [40]. The results are usually made publicly available in the form of specialized databases such as EggNOG [45], Ensembl Compara [46], Inparanoid [47], PANTHER [48], PhylomeDB [49], and OMA [50], and the clustering results provide the basis for GO term annotation transfers, under the assumption that the members of an orthologous group are functionally equivalent.

4 Annotation Transfers from Protein Families

Even when the sequence similarities between proteins of interest and those that have previously been characterized are limited to specific sites, such as individual domains or motifs, they can still be useful for function prediction. Some biological activities such as molecular recognition, protein targeting, and pathway regulation have long been mechanistically linked to short linear motifs—stretches of 10–20 consecutive amino acids exposed on protein surfaces [51]. Furthermore, some well-known protein families can be described by specific arrangements of multiple, possibly discontinuous, linear motifs, or by more general models of their domain sequences, namely sequence profiles [52] or hidden Markov models [53]. Many public databases now give access to groups of evolutionarily related proteins, coding for individual domains or multi-domain architectures. Even though these resources cannot directly assign GO terms to the input amino acid sequences, they can produce valuable assignments to know protein families.

InterPro [54] collates such results from 11 specialized and complementary resources, which differ by the types of patterns used for family assignment, by the amount of manual curation of their contents, and by the use of additional data such as 3-D structure or phylogenetic trees. InterPro entries combine available data and organize them in a hierarchical way, which mirrors the biological relationships between families and subfamilies of proteins. The curators also enrich these annotations with supporting biological information from the scientific literature and with links to external resources such as the PDB [55] and GO. InterPro provides function predictions for the input sequences based on the InterPro2GO mapping, which links each protein domain family to the most specific GO terms that apply to all its members [56]. These annotations form a large bulk of the electronically inferred functional assignments in UniProtKB, where they are integrated with associations generated from other controlled vocabularies, e.g., about subcellular localization and enzymatic activity.

CATH-Gene3D [57] and SUPERFAMILY [58] are two databases that store domain assignments for known protein sequences based on the CATH [59] and SCOP [60] protein structure classification schemes, respectively. CATH-Gene3D data are clustered into functional families which include relatives with highly similar sequences, structures, and functions, as to highlight the strong conservation of important regions such as specificity-determining residues. GO terms are associated probabilistically to each functional family based on how often they occur in the UniProtKB annotations of the whole sequences. The recent CATH FunFHMMer web server automates the search procedure for input sequences, resolves multi-domain architectures, assigns each predicted domain to its functional family, and finally inherits the GO term annotations found in the library [61]. The dcGO—short for domain centric—method follows a similar route, but with some key differences [62]. HMM models are built for both individual domains and supra-domains, i.e., sets of consecutive domains that are defined according to the SCOP structural definition and the evolutionary one in Pfam [63]. Given the annotations in the GOA database [64] and the GO hierarchical structure, each domain and supra-domain is labelled with a set of GO terms that are associated with it in a statistically significant way. The strength of each association is then empirically converted into a confidence score. To facilitate the analysis of the results by non-specialists, the predicted GO terms are divided into four classes according to how specific and informative they are using their information content.

5 De Novo Function Annotation Using Biological Features

The function annotation methods described so far make use of homology to transfer GO terms to a target protein from other previously characterized proteins. In some cases, however, no useful functional annotations can be found for any of the detectable homologues, or in the most extreme case no homologous sequences can be found at all. In this case a de novo method is required which can infer GO terms directly from amino acid sequence in the absence of evolutionary relatedness. This is a very hard problem, and only a few tools have been developed which can handle these situations. The most successful approaches to date employ the basic idea of first transforming the target sequence into a set of component features. These features are then related to particular broad functional classes by means of supervised machine learning techniques. In this way the methods address the question of what kinds of functions can proteins perform with the given set of protein features. As a trivial example, proteins which are predicted to have particular numbers of transmembrane helices as component features will be more likely to have transmembrane transporter activity.

ProtFun, which makes use of neural networks, was the first widely used method for transferring functional annotations between human proteins through similarity of biochemical attributes, such as the occurrence of charged amino acids, low-complexity regions, signal peptides, trans-membrane helices, and posttranslationally modified residues [65, 66]. In the original ProtFun method, only the broad functional classes originally compiled by Monica Riley [67] were considered, but later the authors extended their approach to predicting a representative set of GO terms. FFPred, which is based on support vector machines, has taken this approach further by considering the observed strong correlation between the lengths and positions of intrinsically disordered protein regions with certain molecular functions and biological processes [68, 69]. As with

ProtFun, FFPred was initially developed specifically for annotating human proteins, but the results have been shown to extend reasonably well to other vertebrate proteomes too.

Feature-based protein function assignment offers both advantages and disadvantages over sequence similarity-based approaches. The main advantage is fairly obvious: feature-based methods can work in the absence of homology to characterized proteins, and thus can even be used to assign GO terms to orphan proteins. A further advantage is that feature-based prediction is also able to provide insight into functional changes that occur after alternative splicing, as the input features are likely to reflect sequence deletions relative to the main transcript, e.g., the loss of a signal peptide or disordered region. Probably the main disadvantage is that classification models can only be built for GO terms where there are sufficient examples with experimentally validated assignments. This generally means that assignments can only be made for terms fairly high up in the overall GO graph, and thus highly specific predictions are generally not possible using this kind of approach. Of course, as datasets become larger, these methods will be able to overcome such limitation.

6 Conclusions and Outlook

The widening gap between the number of known sequences and those experimentally characterized has stimulated the development and refinement of a wide array of computational methods for protein function prediction. The scope of this survey has been limited to four classes of sequence-based approaches for GO term annotation transfers, but several other routes could be followed. If the 3-D structure of a protein has been solved or accurately modelled, it is possible to search for global or local structural similarities and predict binding regions and catalytic sites [70, 71]. Comparison of multiple complete genomes can help detect not only orthologous genes as described above, but also further patterns indicative of functional linkages between gene pairs such as fusion events, conserved chromosomal proximity, and co-occurrence/absence in a group of species [72]. Phylogenetic profiling posits that coevolving protein families are functionally coupled, e.g., because they encode for proteins assembling into obligate complexes or participating in the same biological process. Since its inception, this “guilt-by-association” method has been implemented in several different ways [73], and tools able to make GO term assignments are also emerging [74]. Involvement in the same biological process or co-localization can also be inferred from the analysis of protein-protein interaction maps, gene expression profiles, and phenotypic variations following engineered genetic mutations [75]. Finally, integrative strategies combine all such heterogeneous data sources and hold the potential to produce more confident predictions, reduce errors, and overcome the intrinsic limitations of individual algorithms [31, 76–78]. For instance, protein sequence and structure data appear to be better suited to predict terms in the molecular function category, while genome-wide datasets can shed light on biological processes and protein subcellular localization. In the future, these methods will become increasingly valuable to generate testable hypotheses about protein function as they improve in accuracy – thanks to additional experimental data and to better ways of using them – as well as in user-friendliness to experimentalists and nonspecialists in general.

Acknowledgements

This work was partially supported by the UK Biotechnology and Biological Sciences Research Council. Open Access charges were funded by the University College London Library, the Swiss Institute of Bioinformatics, the Agassiz Foundation, and the Foundation for the University of Lausanne.

Open Access This chapter is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, duplication, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, a link is provided to the Creative Commons license and any changes made are indicated.

The images or other third party material in this chapter are included in the work's Creative Commons license, unless indicated otherwise in the credit line; if such material is not included in the work's Creative Commons license and the respective action is not permitted by statutory regulation, users will need to obtain permission from the license holder to duplicate, adapt or reproduce the material.

References

1. Soon WW, Hariharan M, Snyder MP (2013) High-throughput sequencing for biology and medicine. Mol Syst Biol 9:640. doi:10.1038/msb.2012.61
2. Mitra K, Carvunis AR, Ramesh SK, Ideker T (2013) Integrative approaches for finding modular structure in biological networks. Nat Rev Genet 14(10):719–732. doi:10.1038/nrg3552
3. Mahony S, Pugh BF (2015) Protein-DNA binding in high-resolution. Crit Rev Biochem Mol Biol:1–15. doi:10.3109/10409238.2015.1051505
4. McHugh CA, Russell P, Guttman M (2014) Methods for comprehensive experimental identification of RNA-protein interactions. Genome Biol 15(1):203. doi:10.1186/gb4152
5. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. The Gene Ontology

Consortium. Nat Genet 25(1):25–29. doi:10.1038/75556

6. UniProt C (2015) UniProt: a hub for protein information. Nucleic Acids Res 43(Database issue):D204–D212. doi:10.1093/nar/gku989
7. Furnham N (2016) Complementary sources of protein functional information: the far side of GO. In: Dessimoz C, Škunca N (eds) The gene ontology handbook. Methods in molecular biology, vol 1446. Humana Press. Chapter 19
8. Škunca N, Roberts RJ, Steffen M (2016) Evaluating computational gene ontology annotations. In: Dessimoz C, Škunca N (eds) The gene ontology handbook. Methods in molecular biology, vol 1446. Humana Press. Chapter 8
9. Holliday GL, Davidson R, Akiva E, Babbitt PC (2016) Evaluating functional annotations of enzymes using the gene ontology. In: Dessimoz C, Škunca N (eds) The gene ontology handbook. Methods in molecular biology, vol 1446. Humana Press. Chapter 9
10. Friedberg I, Radivojac P (2016) Community-wide evaluation of computational function prediction. In: Dessimoz C, Škunca N (eds) The gene ontology handbook. Methods in molecular biology, vol 1446. Humana Press. Chapter 10
11. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215(3):403–410. doi:10.1016/S0022-2836(05)80360-2
12. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25(17):3389–3402
13. Soding J, Remmert M (2011) Protein sequence comparison and fold recognition: progress and good-practice benchmarking. Curr Opin Struct Biol 21(3):404–411. doi:10.1016/j.sbi.2011.03.005
14. Rost B (2002) Enzyme function less conserved than anticipated. J Mol Biol 318(2):595–608. doi:10.1016/S0022-2836(02)00016-5
15. Schnoes AM, Brown SD, Dodevski I, Babbitt PC (2009) Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. PLoS Comput Biol 5(12): e1000605. doi:10.1371/journal.pcbi.1000605
16. Devos D, Valencia A (2000) Practical limits of function prediction. Proteins 41(1):98–107
17. Wilson CA, Kreychman J, Gerstein M (2000) Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. J Mol Biol 297(1):233–249. doi:10.1006/jmbi.2000.3550
18. Tian W, Skolnick J (2003) How well is enzyme function conserved as a function of pairwise sequence identity? J Mol Biol 333(4):863–882
19. Sangar V, Blankenberg DJ, Altman N, Lesk AM (2007) Quantitative sequence-function relationships in proteins based on gene ontology. BMC Bioinformatics 8:294. doi:10.1186/1471-2105-8-294
20. Addou S, Rentzsch R, Lee D, Orengo CA (2009) Domain-based and family-specific sequence identity thresholds increase the levels of reliable protein function transfer. J Mol Biol 387(2):416–430. doi:10.1016/j.jmb.2008.12.045
21. Zhang J, Yang JR (2015) Determinants of the rate of protein sequence evolution. Nat Rev Genet 16(7):409–420. doi:10.1038/nrg3950
22. Martin DM, Berriman M, Barton GJ (2004) GOtcha: a new method for prediction of protein function assessed by the annotation of seven genomes. BMC Bioinformatics 5:178. doi:10.1186/1471-2105-5-178
23. Hawkins T, Chitale M, Luban S, Kihara D (2009) PFP: automated prediction of gene ontology functional annotations with confidence scores using protein sequence data. Proteins 74(3):566–582. doi:10.1002/prot.22172
24. Chitale M, Hawkins T, Park C, Kihara D (2009) ESG: extended similarity group method for automated protein function prediction. Bioinformatics 25(14):1739–1745. doi:10.1093/bioinformatics/btp309
25. Vinayagam A, Konig R, Moormann J, Schubert F, Eils R, Glatting KH, Suhai S (2004) Applying support vector machines for gene ontology based gene function prediction. BMC Bioinformatics 5:116. doi:10.1186/1471-2105-5-116
26. Gotz S, Garcia-Gomez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, Robles M, Talon M, Dopazo J, Conesa A (2008) High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res 36(10):3420–3435. doi:10.1093/nar/gkn176
27. Piovesan D, Martelli PL, Fariselli P, Zauli A, Rossi I, Casadio R (2011) BAR-PLUS: the Bologna Annotation Resource Plus for functional and structural annotation of protein sequences. Nucleic Acids Res 39(Web Server issue):W197–W202. doi:10.1093/nar/gkr292
28. Duda RO, Hart PE, Stork DG (2012) Pattern classification. Wiley, New York
29. Sokolov A, Ben-Hur A (2010) Hierarchical classification of gene ontology terms using the

GOstruct method. J Bioinforma Comput Biol 8(02):357–376

30. Clark WT, Radivojac P (2011) Analysis of protein function and its prediction from amino acid sequence. Proteins 79(7):2086–2096
31. Cozzetto D, Buchan DW, Bryson K, Jones DT (2013) Protein function prediction by massive integration of evolutionary analyses and multiple data sources. BMC Bioinformatics 14(Suppl 3):S1. doi:10.1186/1471-2105-14-S3-S1
32. Gabaldon T, Koonin EV (2013) Functional and evolutionary implications of gene orthology. Nat Rev Genet 14(5):360–366. doi:10.1038/nrg3456
33. Kachroo AH, Laurent JM, Yellman CM, Meyer AG, Wilke CO, Marcotte EM (2015) Evolution. Systematic humanization of yeast genes reveals conserved functions and genetic modularity. Science 348(6237):921–925. doi:10.1126/science.aaa0769
34. Dean EJ, Davis JC, Davis RW, Petrov DA (2008) Pervasive and persistent redundancy among duplicated genes in yeast. PLoS Genet 4(7): e1000113. doi:10.1371/journal.pgen.1000113
35. Tischler J, Lehner B, Chen N, Fraser AG (2006) Combinatorial RNA interference in Caenorhabditis elegans reveals that redundancy between gene duplicates can be maintained for more than 80 million years of evolution. Genome Biol 7(8):R69. doi:10.1186/gb-2006-7-8-R69
36. Nehrt NL, Clark WT, Radivojac P, Hahn MW (2011) Testing the ortholog conjecture with comparative functional genomic data from mammals. PLoS Comput Biol 7(6):e1002073. doi:10.1371/journal.pcbi.1002073
37. Chen X, Zhang J (2012) The ortholog conjecture is untestable by the current gene ontology but is supported by RNA sequencing data. PLoS Comput Biol 8(11):e1002784. doi:10.1371/journal.pcbi.1002784
38. Altenhoff AM, Studer RA, Robinson-Rechavi M, Dessimoz C (2012) Resolving the ortholog conjecture: orthologs tend to be weakly, but significantly, more similar in function than paralogs. PLoS Comput Biol 8(5):e1002514. doi:10.1371/journal.pcbi.1002514
39. Rogozin IB, Managadze D, Shabalina SA, Koonin EV (2014) Gene family level comparative analysis of gene expression in mammals validates the ortholog conjecture. Genome Biol Evol 6(4):754–762. doi:10.1093/gbe/evu051
40. Altenhoff AM, Dessimoz C (2012) Inferring orthology and paralogy. Methods Mol Biol 855:259–279. doi:10.1007/978-1-61779-582-4_9
41. Engelhardt BE, Jordan MI, Muratore KE, Brenner SE (2005) Protein molecular function prediction by Bayesian phylogenomics. PLoS Comput Biol 1(5):e45. doi:10.1371/journal.pcbi.0010045
42. Engelhardt BE, Jordan MI, Srouji JR, Brenner SE (2011) Genome-scale phylogenetic function annotation of large and diverse protein families. Genome Res 21(11):1969–1980. doi:10.1101/gr.104687.109
43. Sahraeian SM, Luo KR, Brenner SE (2015) SIFTER search: a web server for accurate phylogeny-based protein function prediction. Nucleic Acids Res. doi:10.1093/nar/gkv461
44. Gaudet P, Livstone MS, Lewis SE, Thomas PD (2011) Phylogenetic-based propagation of functional annotations within the Gene Ontology consortium. Brief Bioinform 12(5):449–462. doi:10.1093/bib/bbr042
45. Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC, Rattei T, Mende DR, Sunagawa S, Kuhn M, Jensen LJ, von Mering C, Bork P (2016) eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Res 44(D1):D286–D293. doi:10.1093/nar/gkv1248
46. Flicek P, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S, Gil L, Giron CG, Gordon L, Hourlier T, Hunt S, Johnson N, Juettemann T, Kahari AK, Keenan S, Kulesha E, Martin FJ, Maurel T, McLaren WM, Murphy DN, Nag R, Overduin B, Pignatelli M, Pritchard B, Pritchard E, Riat HS, Ruffier M, Sheppard D, Taylor K, Thormann A, Trevanion SJ, Vullo A, Wilder SP, Wilson M, Zadissa A, Aken BL, Birney E, Cunningham F, Harrow J, Herrero J, Hubbard TJ, Kinsella R, Muffato M, Parker A, Spudich G, Yates A, Zerbino DR, Searle SM (2014) Ensembl 2014. Nucleic Acids Res 42(Database issue):D749–D755. doi:10.1093/nar/gkt1196
47. Sonnhammer EL, Ostlund G (2015) InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. Nucleic Acids Res 43(Database issue):D234–D239. doi:10.1093/nar/gku1203
48. Mi H, Poudel S, Muruganujan A, Casagrande JT, Thomas PD (2016) PANTHER version 10: expanded protein families and functions, and analysis tools. Nucleic Acids Res 44(D1): D336–D342. doi:10.1093/nar/gkv1194
49. Huerta-Cepas J, Capella-Gutierrez S, Pryszcz LP, Marcet-Houben M, Gabaldon T (2014) PhylomeDB v4: zooming into the plurality of evolutionary histories of a genome. Nucleic

Acids Res 42(Database issue):D897–D902. doi:10.1093/nar/gkt1177

50. Altenhoff AM, Skunca N, Glover N, Train CM, Sueki A, Pilizota I, Gori K, Tomiczek B, Muller S, Redestig H, Gonnet GH, Dessimoz C (2015) The OMA orthology database in 2015: function predictions, better plant support, synteny view and other improvements. Nucleic Acids Res 43(Database issue):D240–D249. doi:10.1093/nar/gku1158
51. Van Roey K, Uyar B, Weatheritt RJ, Dinkel H, Seiler M, Budd A, Gibson TJ, Davey NE (2014) Short linear motifs: ubiquitous and functionally diverse protein interaction modules directing cell regulation. Chem Rev 114(13):6733–6778. doi:10.1021/cr400585q
52. Gribskov M, McLachlan AD, Eisenberg D (1987) Profile analysis: detection of distantly related proteins. Proc Natl Acad Sci U S A 84(13):4355–4358
53. Eddy SR (1998) Profile hidden Markov models. Bioinformatics 14(9):755–763
54. Mitchell A, Chang HY, Daugherty L, Fraser M, Hunter S, Lopez R, McAnulla C, McMenamin C, Nuka G, Pesseat S, Sangrador-Vegas A, Scheremetjew M, Rato C, Yong SY, Bateman A, Punta M, Attwood TK, Sigrist CJ, Redaschi N, Rivoire C, Xenarios I, Kahn D, Guyot D, Bork P, Letunic I, Gough J, Oates M, Haft D, Huang H, Natale DA, Wu CH, Orengo C, Sillitoe I, Mi H, Thomas PD, Finn RD (2015) The InterPro protein families database: the classification resource after 15 years. Nucleic Acids Res 43(Database issue):D213–D221. doi:10.1093/nar/gku1243
55. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. Nucleic Acids Res 28(1):235–242
56. Burge S, Kelly E, Lonsdale D, Mutowo-Muellenet P, McAnulla C, Mitchell A, Sangrador-Vegas A, Yong SY, Mulder N, Hunter S (2012) Manual GO annotation of predictive protein signatures: the InterPro approach to GO curation. Database (Oxford) 2012:bar068. doi:10.1093/database/bar068
57. Sillitoe I, Lewis TE, Cuff A, Das S, Ashford P, Dawson NL, Furnham N, Laskowski RA, Lee D, Lees JG, Lehtinen S, Studer RA, Thornton J, Orengo CA (2015) CATH: comprehensive structural and functional annotations for genome sequences. Nucleic Acids Res 43(Database issue):D376–D381. doi:10.1093/nar/gku947
58. Oates ME, Stahlhacke J, Vavoulis DV, Smithers B, Rackham OJ, Sardar AJ, Zaucha J, Thurlby N, Fang H, Gough J (2015) The SUPERFAMILY 1.75 database in 2014: a doubling of data. Nucleic Acids Res 43(Database issue):D227–D233. doi:10.1093/nar/gku1041
59. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM (1997) CATH--a hierarchic classification of protein domain structures. Structure 5(8):1093–1108
60. Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247(4): 536–540. doi:10.1006/jmbi.1995.0159
61. Das S, Sillitoe I, Lee D, Lees JG, Dawson NL, Ward J, Orengo CA (2015) CATH FunFHMMer web server: protein functional annotations using functional family assignments. Nucleic Acids Res. doi:10.1093/nar/gkv488
62. Fang H, Gough J (2013) DcGO: database of domain-centric ontologies on functions, phenotypes, diseases and more. Nucleic Acids Res 41(Database issue):D536–D544. doi:10.1093/nar/gks1080
63. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, Sonnhammer EL, Tate J, Punta M (2014) Pfam: the protein families database. Nucleic Acids Res 42(Database issue):D222–D230. doi:10.1093/nar/gkt1223
64. Huntley RP, Sawford T, Mutowo-Meullenet P, Shypitsyna A, Bonilla C, Martin MJ, O'Donovan C (2015) The GOA database: gene Ontology annotation updates for 2015. Nucleic Acids Res 43(Database issue):D1057–D1063. doi:10.1093/nar/gku1113
65. Jensen LJ, Gupta R, Blom N, Devos D, Tamames J, Kesmir C, Nielsen H, Staerfeldt HH, Rapacki K, Workman C, Andersen CA, Knudsen S, Krogh A, Valencia A, Brunak S (2002) Prediction of human protein function from post-translational modifications and localization features. J Mol Biol 319(5):1257–1265. doi:10.1016/S0022-2836(02)00379-0
66. Jensen LJ, Gupta R, Staerfeldt HH, Brunak S (2003) Prediction of human protein function according to Gene Ontology categories. Bioinformatics 19(5):635–642
67. Riley M (1993) Functions of the gene products of Escherichia coli. Microbiol Rev 57(4):862–952
68. Lobley A, Swindells MB, Orengo CA, Jones DT (2007) Inferring function using patterns of native disorder in proteins. PLoS Comput Biol 3(8):e162. doi:10.1371/journal.pcbi.0030162
69. Minneci F, Piovesan D, Cozzetto D, Jones DT (2013) FFPred 2.0: improved homology-independent prediction of gene ontology

terms for eukaryotic protein sequences. PLoS One 8(5):e63754. doi:10.1371/journal.pone.0063754

70. Jacobson MP, Kalyanaraman C, Zhao S, Tian B (2014) Leveraging structure for enzyme function prediction: methods, opportunities, and challenges. Trends Biochem Sci 39(8):363–371. doi:10.1016/j.tibs.2014.05.006

71. Petrey D, Chen TS, Deng L, Garzon JI, Hwang H, Lasso G, Lee H, Silkov A, Honig B (2015) Template-based prediction of protein function. Curr Opin Struct Biol 32C:33–38. doi:10.1016/j.sbi.2015.01.007

72. Galperin MY, Koonin EV (2014) Comparative genomics approaches to identifying functionally related genes. In: Algorithms for computational biology. Springer, Berlin, pp 1–24

73. Pellegrini M (2012) Using phylogenetic profiles to predict functional relationships. Methods Mol Biol 804:167–177. doi:10.1007/978-1-61779-361-5_9

74. Skunca N, Bosnjak M, Krisko A, Panov P, Dzeroski S, Smuc T, Supek F (2013) Phyletic profiling with cliques of orthologs is enhanced by signatures of paralogy relationships. PLoS Comput Biol 9(1):e1002852. doi:10.1371/journal.pcbi.1002852

75. Yu D, Kim M, Xiao G, Hwang TH (2013) Review of biological network data and its applications. Genomics Inform 11(4):200–210. doi:10.5808/GI.2013.11.4.200

76. Ma X, Chen T, Sun F (2014) Integrative approaches for predicting protein function and prioritizing genes for complex phenotypes using protein interaction networks. Brief Bioinform 15(5):685–698. doi:10.1093/bib/bbt041

77. Wass MN, Barton G, Sternberg MJ (2012) CombFunc: predicting protein function using heterogeneous data sources. Nucleic Acids Res 40(Web Server issue):W466–W470. doi:10.1093/nar/gks489

78. Piovesan D, Giollo M, Leonardi E, Ferrari C, Tosatto SC (2015) INGA: protein function prediction combining interaction networks, domain assignments and sequence similarity. Nucleic Acids Res. doi:10.1093/nar/gkv523