Arch Microbiol (2010) 192:151–155
DOI 10.1007/s00203-010-0549-9

MINI-REVIEW

An overview of in silico protein function prediction

Roy D. Sleator · Paul Walsh

Received: 2 December 2009 / Revised: 8 January 2010 / Accepted: 10 January 2010 / Published online: 3 February 2010

Abstract As the protein databases continue to expand at an exponential rate, fed by daily uploads from multiple large scale genomic and metagenomic projects, the problem of assigning a function to each new protein has become the focus of significant research interest in recent times. Herein, we review the most recent advances in the field of automated function prediction (AFP). We begin by defining what is meant by biological "function" and the means of describing such functions using standardised machine readable ontologies. We then focus on the various function-prediction programs available, both sequence and structure based, and outline their associated strengths and weaknesses. Finally, we conclude with a brief overview of the future challenges and outstanding questions in the field, which still remain unanswered.

Keywords Protein function · Homology-based transfer · Ontologies · Sequence and structure Motifs

Communicated by Erko Stackebrandt.

R. D. Sleator (✉)
Department of Biological Sciences,
Cork Institute of Technology, Rossa Avenue,
Bishopstown, Cork, Ireland
e-mail: roy.sleator@cit.ie

P. Walsh
Department of Computing,
Cork Institute of Technology, Rossa Avenue,
Bishopstown, Cork, Ireland

R. D. Sleator · P. Walsh
CIT Bioinformatics Group,
Cork Institute of Technology, Rossa Avenue,
Bishopstown, Cork, Ireland

Introduction

The recent explosion in the number and diversity of novel proteins identified by genomic and metagenomic sequencing projects poses a new and important question to the blossoming field of systems biology—What do all these proteins do?

Until recently, in the absence of any experimental evidence, homology-based transfer remained the gold standard for in silico analysis of protein function. 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 may be transferred to the former (referred to as the 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.

To compensate, several non-homology computational-based approaches to protein function prediction, based on sequence, structure, evolution, biochemical and genetic and genomic knowledge, have begun to immerge (Fig. 1).

Herein, we review the most recent advances in the field of automated function prediction (AFP) and discuss the future challenges and outstanding questions, which still remain unanswered.

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

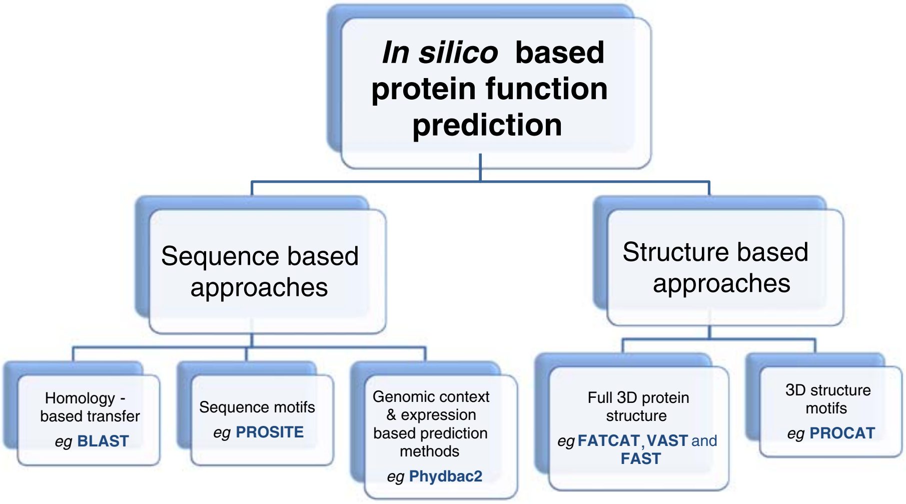
In silico based protein function prediction
Sequence based approaches
Structure based approaches
Homology - based transfer
eg BLAST
Sequence motifs
eg PROSITE
Genomic context & expression based prediction methods
eg Phydbac2
Full 3D protein structure
eg FATCAT, VAST and FAST
3D structure motifs
eg PROCAT

Fig. 1 Schematic overview of in silico-based protein function prediction methods

almost instantaneous enzymatic reactions (molecular function) to the much slower overall biological process (Godzik et al. 2007). Knowing which functional aspect is being investigated is thus extremely important and can only properly be achieved by the establishment of a standardised 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 (Losko and Heumann 2009). 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) (Ashburner and Lewis 2002) goes some way to achieving this goal of 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 (Smith et al. 2005), 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, we 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 (Altschul et al. 1997), is perhaps the most widely used form of computational function-prediction method; assigning unannotated proteins with the function of their annotated homologues. 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 to functional similarity, exceptions have been observed on both ends of the similarity scale. Rost (2002), for example, showed that even at high sequence similarity rates, enzymatic function may not necessarily be conserved, while Galperin et al. (1998) observed that certain enzymes with high levels of functional homology may be classed as analogous on the basis of sequence dissimilarity. 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 (Bork 2000; Gilks et al. 2005).

Furthermore, as the databases continue to expand the utility of the homology-based transfer approach begins to breakdown. The recent explosion of large scale metagenomic sequencing projects (Sleator et al. 2008) has resulted in an unprecedented amount of novel sequence data 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 ≤5%, while even allowing for error rates of >40%, there is no annotation for >30% of all proteins (Rost et al. 2003).

Sequence motifs

Typically of the 100–300 amino acids in a functional protein domain <10% constitute the protein's active sites (Friedberg 2006). 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 (Hulo et al. 2008) 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 (Henikoff et al. 2000) and PRINTS (Attwood et al. 2003).

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 (Eisenberg et al. 2000). Furthermore, in prokaryote genomes the loci of functionally related proteins tend to be co-located on the chromosome. Combining co-evolution and co-location (chromosomal proximity) has given rise to a new generation of function-prediction algorithms such as Phydbac2 (Enault et al. 2005).

As an extension of co-location, genes involved in similar cellular functions also tend to be co-transcribed. 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. (1999), 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 carried out 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, including STRING (Zhao et al. 2008), DIP (Lehne and Schlitt 2009) and GRID (Breitkreutz et al. 2003). The availability of protein interaction networks for model species has facilitated the development of effective computational approaches to interpret the data and rapidly elucidate protein function as outlined by Sharan et al. (2007).

Structure-based approaches

Given that protein structure is far more conserved than sequence, many proteins that exhibit little or no sequence similarities, due to evolutionary constraints still retain significant structure similarity (Watson et al. 2005). In this respect, structure is a useful indicator of function; indeed most known protein folds are associated with a particular function or functional milieu (Todd et al. 2001). Programs that scan the Protein Data Bank (PDB) for structural similarity given a query sequence include, amongst others, FATCAT (Ye and Godzik 2004), PAST (Taubig et al. 2006) and VAST (Gibrat et al. 1996). 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 (Goldsmith-Fischman and Honig 2003; Laskowski et al. 2003).

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 analysing 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 (Jones and Thornton 2004), one example of which is PROCAT (Wallace et al. 1996), 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 (Di Gennaro et al. 2001).

Conclusion

In contrast to sequence and structure information in which the data are either known (as is the case for amino acid sequence) or easily predicted (e.g. loops in structure prediction) the multifaceted and ambiguous nature of biological function makes its elucidation a far more complex endeavour. The complexity of the problem is perhaps best illustrated by Jeffery (2003) 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 (which classifies sequences at superfamily, family and subfamily levels, predicting the occurrence of functional domains, repeats and important sites), ProFunc (which identifies the likely biochemical function of a protein from its three-dimensional structure; using fold matching, residue conservation, surface cleft analysis and functional 3D templates, to identify both the protein's likely active site and possible homologues in the PDB) and ProKnow (which annotates proteins with Gene Ontology functional terms; extracting features from the protein such as 3D fold, sequence, motif and functional linkages) .

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.

Acknowledgments The authors wish to acknowledge the financial assistance of the Faculty of Engineering and Science, the Department of Biological Sciences and the Department of Computing at Cork Institute of Technology.

References

Altschul SF et al (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Ashburner M, Lewis S (2002) On ontologies for biologists: the Gene Ontology–untangling the web. Novartis Found Symp 247: 66–80; discussion 80–63, 84–90, 244–252

Attwood TK et al (2003) PRINTS and its automatic supplement, prePRINTS. Nucleic Acids Res 31:400–402

Bork P (2000) Powers and pitfalls in sequence analysis: the 70% hurdle. Genome Res 10:398–400

Breitkreutz BJ, Stark C, Tyers M (2003) The GRID: the general repository for interaction datasets. Genome Biol 4:R23

Di Gennaro JA et al (2001) Enhanced functional annotation of protein sequences via the use of structural descriptors. J Struct Biol 134:232–245

Eisenberg D, Marcotte EM, Xenarios I, Yeates TO (2000) Protein function in the post-genomic era. Nature 405:823–826

Enault F, Suhre K, Claverie JM (2005) Phydbac "Gene Function Predictor": a gene annotation tool based on genomic context analysis. BMC Bioinformatics 6:247

Friedberg I (2006) Automated protein function prediction–the genomic challenge. Brief Bioinform 7:225–242

Galperin MY, Walker DR, Koonin EV (1998) Analogous enzymes: independent inventions in enzyme evolution. Genome Res 8:779–790

Gibrat JF, Madej T, Bryant SH (1996) Surprising similarities in structure comparison. Curr Opin Struct Biol 6:377–385

Gilks WR, Audit B, de Angelis D, Tsoka S, Ouzounis CA (2005) Percolation of annotation errors through hierarchically structured protein sequence databases. Math Biosci 193:223–234

Godzik A, Jambon M, Friedberg I (2007) Computational protein function prediction: are we making progress? Cell Mol Life Sci 64:2505–2511

Goldsmith-Fischman S, Honig B (2003) Structural genomics: computational methods for structure analysis. Protein Sci 12:1813–1821

Henikoff JG, Greene EA, Pietrokovski S, Henikoff S (2000) Increased coverage of protein families with the blocks database servers. Nucleic Acids Res 28:228–230

Hulo N et al (2008) The 20 years of PROSITE. Nucleic Acids Res 36:D245–D249

Jeffery CJ (2003) Moonlighting proteins: old proteins learning new tricks. Trends Genet 19:415–417

Jones S, Thornton JM (2004) Searching for functional sites in protein structures. Curr Opin Chem Biol 8:3–7

Laskowski RA, Watson JD, Thornton JM (2003) From protein structure to biochemical function? J Struct Funct Genomics 4:167–177

Lehne B, Schlitt T (2009) Protein-protein interaction databases: keeping up with growing interactomes. Hum Genomics 3:291–297

Losko S, Heumann K (2009) Semantic data integration and knowledge management to represent biological network associations. Methods Mol Biol 563:241–258

Rost B (2002) Enzyme function less conserved than anticipated. J Mol Biol 318:595–608

Rost B, Liu J, Nair R, Wrzeszczynski KO, Ofran Y (2003) Automatic prediction of protein function. Cell Mol Life Sci 60:2637–2650

Sharan R, Ulitsky I, Shamir R (2007) Network-based prediction of protein function. Mol Syst Biol 3:88

Sleator RD, Shortall C, Hill C (2008) Metagenomics. Lett Appl Microbiol 47:361–366

Smith CL, Goldsmith CA, Eppig JT (2005) The mammalian phenotype ontology as a tool for annotating, analyzing and comparing phenotypic information. Genome Biol 6:R7

Taubig H, Buchner A, Griebsch J (2006) PAST: fast structure-based searching in the PDB. Nucleic Acids Res 34:W20–W23

Todd AE, Orengo CA, Thornton JM (2001) Evolution of function in protein superfamilies, from a structural perspective. J Mol Biol 307:1113–1143

Walker MG, Volkmuth W, Sprinzak E, Hodgson D, Klingler T (1999) Prediction of gene function by genome-scale expression analysis: prostate cancer-associated genes. Genome Res 9:1198–1203

Wallace AC, Laskowski RA, Thornton JM (1996) Derivation of 3D coordinate templates for searching structural databases: application to Ser-His-Asp catalytic triads in the serine proteinases and lipases. Protein Sci 5:1001–1013

Watson JD, Laskowski RA, Thornton JM (2005) Predicting protein function from sequence and structural data. Curr Opin Struct Biol 15:275–284

Ye Y, Godzik A (2004) FATCAT: a web server for flexible structure comparison and structure similarity searching. Nucleic Acids Res 32:W582–W585

Zhao XM, Chen L, Aihara K (2008) Protein function prediction with high-throughput data. Amino Acids 35:517–530