DRUG DISCOVERY TODAY BIOSILICO

Beyond annotation transfer by homology: novel protein-function prediction methods to assist drug discovery

Yanay Ofran, Marco Punta, Reinhard Schneider and Burkhard Rost

Every entirely sequenced genome reveals 100s to 1000s of protein sequences for which the only annotation available is 'hypothetical protein'. Thus, in the human genome and in the genomes of pathogenic agents there could be 1000s of potential, unexplored drug targets. Computational prediction of protein function can play a role in studying these targets. We shall review the challenges, research approaches and recently developed tools in the field of computational function-prediction and we will discuss the ways these issues can change the process of drug discovery.

Yanay Ofran*
Marco Punta
Burkhard Rost
CUBIC,
Department of Biochemistry and Molecular Biophysics,
Columbia University,
650 West 168th Street BB217,
New York, NY 10032, USA
Yanay Ofran
Marco Punta
Burkhard Rost
Columbia University Center for Computational Biology and Bioinformatics (C2B2),
Russ Berrie Pavilion,
1150 St. Nicholas Avenue,
New York, NY 10032, USA
Burkhard Rost
Northeast Structural Genomics Consortium (NESG),
Department of Biochemistry and Molecular Biophysics,
Columbia University,
650 West 168th Street BB217,
New York, NY 10032, USA
Reinhard Schneider
Structural and Computational Biology Unit,
EMBL,
Meyerhofstr. 1,
69117 Heidelberg, Germany
*e-mail:
ofran@cubic.bioc.columbia.edu

Haemophilus influenzae was the first entire organism to be sequenced a decade ago [1]; baker's yeast (*Saccharomyces cerevisiae*), the first single-cell eukaryote to be sequenced, followed just a year later [2]; the first animal genome to be sequenced two years after that was the worm *Caenorhabditis elegans* [3] and, four years ago, the completion of the human genome added the first vertebrate sequence to this list [4,5]. In March 2005, entirely sequenced genomes were available for >250 organisms and an estimated 1000 organisms are being sequenced at present. In addition, environmental sequencing provides new sequences at a previously unseen pace [6]. This explosion of sequence information has begun to change the face of molecular biology completely as it triggers the development of novel, experimental high-throughput techniques. For example, instead of arbitrarily focusing on a particular kinase or receptor, researchers can now target all kinases and/or receptors (that have a certain signature and/or phenotype) from the entire human genome. Similarly, structural genomics attempts to solve the structure of at least one representative protein for each existing structural fold, thus targeting one protein for each family of sequence-related proteins. Structural genomics consortia simultaneously pursue targets from all of the available organisms [7] and less than five years of pilot projects (funded by the Protein Structure Initiative from the National Institutes of Health) have generated >1000 high-resolution structures [8].

This deluge of new protein sequences and structures provides unprecedented possibilities in drug discovery. Comparative genomics paved the way forward, identifying the genetic basis of disease and the ability to interfere with it [9]. Now, the wealth of genomic data available presents the drug discovery community with new challenges. Traditionally, when attempting to interfere or tamper with pathological processes, drug discovery focused on the 'usual suspects' such as kinases or G-protein coupled receptors (GPCRs). These targets have high therapeutic relevance and have been analyzed exhaustively from the perspectives of basic science and lead discovery. Drug discovery for these popular targets, therefore, relies on a meticulous understanding of their function. By contrast, potential targets provided by genome projects and

1359-6446/05/$ – see front matter PII: S1359-6446(05)03621-4

environmental sequencing are not usually endowed with such elaborate background knowledge. Although it was typical to find 100 literature references for each target identified ten years ago, this number was down to approximately ten references for the targets identified today [10]. Sub-cellular localization, enzymatic or transport action and protein–ligand interactions can all be unknown variables for these proteins. The increased number of targets and the decreased amount of information is generating a bottleneck in the target validation process.

Changing paradigms in drug discovery

Over the past few years we have experienced some significant changes to the pharmaceutical drug discovery process. One change that is impacting strongly on the early-stage development process is the shift from focusing on improving and optimizing the chemistry of a small number of targets to focusing on the validation process of a few thousand druggable targets. Two other new and widely applied strategies, geared to filling the drying pipelines of big pharmaceutical companies, are the search for new molecular entities directed at diseases with large unmet needs (instead of the small incremental improvements in efficacy resulting in 'me too' drugs) and the significant efforts being made to implement innovative strategies for finding alternative medical uses for drugs that have already been developed. The motto of drug repurposing is 'don't tell me about tomorrow's drugs; tell me what I can do with my drugs today'. Thus, the aim is to find a better or different target for an existing compound, which leads to a change in approach because companies compete on the target level instead of on the compound level.

The goal for the coming decades will be to explore the overlap between chemistry space and protein space (Figure 1). This will require a new conceptual framework in which drug developers investigate the applicability of the known compounds, not only to known targets but also to every node in the biological system that affects the therapeutic response. This process will require a complete map of the target space. At present, the total number of druggable proteins in humans is not known but we certainly can expect this number to be dramatically larger than the few hundred target classes we have explored so far.

This review discusses the ability of computational tools to address all of these challenges. In particular, we will focus on the emerging field of protein-function prediction. Whereas several recent reviews offer a comprehensive coverage of the challenges and accomplishments [11,12], the purpose of this review is to examine this field in the context of drug discovery.

Drug discovery and protein-function prediction

The development of a drug is a long and difficult process that involves numerous steps: target identification and validation; lead compound screening or design; compound optimization; compound purification or synthesis; and clinical trials. Whereas various bioinformatics and biophysics techniques are routinely used to assist with lead compound screening or design processes [13,14], function prediction can have an impact on target selection and on various stages of lead compound design. The most widely known and used *in silico* protein function prediction method is homology-based annotation transfer (i.e. the transfer of a function from one protein to another on the basis of their common evolutionary origin). This method, although powerful, has severe limitations and it has been claimed that annotation transfer is one of the main sources of incorrect functional annotations that occur in databases [15–17]. The first and most obvious limitation is that this procedure cannot be used on proteins that lack annotated homologues. As an alternative, numerous non-homology based methods for protein-function prediction have been developed over the past few years. They take advantage of sequence, structure, evolution and biochemical and genetic knowledge. We will discuss the applicability of protein-function prediction to different stages of drug discovery.

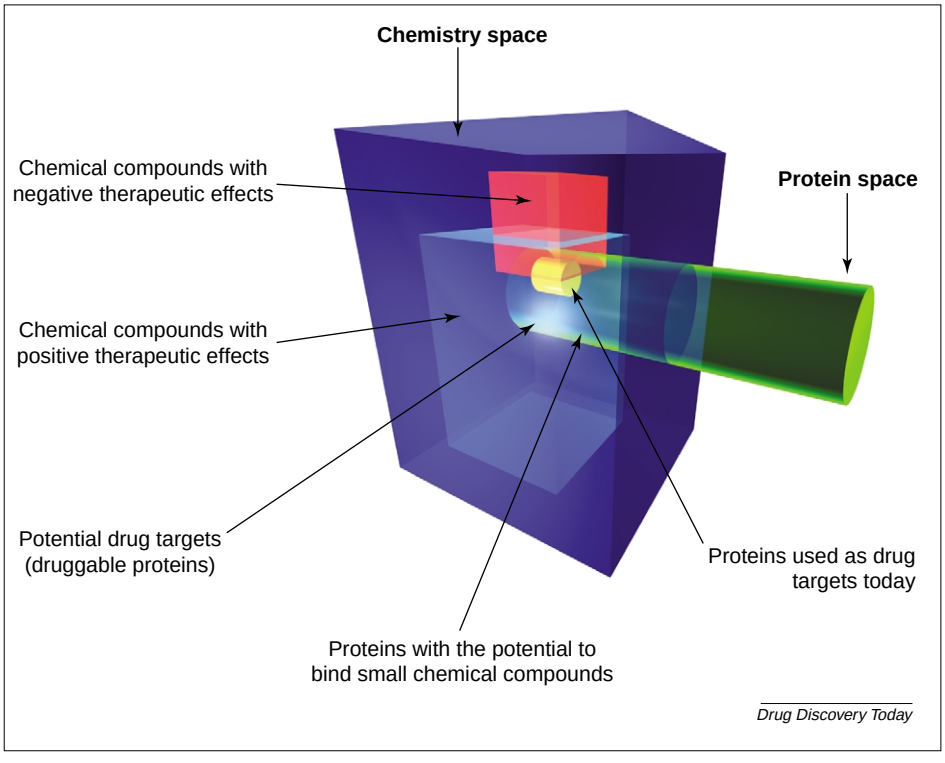

FIGURE 1

The overlap between the chemistry space and the protein space. Only a small fraction of proteins are used as drug targets today. The goal for the next few decades will be to explore the space of the potential drug targets (druggable proteins) and to reveal the relationship between them and the chemistry space.

A detailed explanation of the methods we refer to can be found in the second half of this review.

It is becoming apparent that, in the future, the traditional approach of analyzing the interaction between lead compounds and targets of interest will not be the only routine way to discover new drugs. When applied to a complete proteome, each compound could have a varying degree of affinity for several different targets. The selected target will, therefore, be the protein that is involved in the desired biological process and has the highest (or at least one of the highest) affinity for the compound. In this respect the emphasis on multi-target compounds (or the development of multi-component therapeutics) will probably increase significantly as our understanding of biological systems improves. The goal will be to identify and select the disease-modifying nodes in a biochemical network that will have the desired therapeutic effect when they are triggered. Functional knowledge about each potential target and their roles in the disease-modifying effect will be crucial to this process. Given the poor coverage of these targets by experimental annotations, in most cases functional information from *in silico* methods might provide the only clues available. Several tools might be applicable at this stage, carefully applying homology-based annotation transfer (from sequence or structure) could narrow down the number of targets to screen and *de novo* prediction methods that predict functional class [e.g. in the form of gene ontology (GO) annotation] could add novel targets to the list.

In chemistry space, competition and intellectual property (IP) considerations play an important role in lead selection and optimization. Often, a compound (other than the optimum compound) is brought into the development process simply because it does not have any IP limitations. It is also not unusual for compounds to be selected and optimized around competitors' IP gaps and we are starting to see a similar situation for the target space. Targets will be routinely selected because IP does not already cover them (not because they are understood and would be the primary choice for the underlying disease model). As in the previous case, *in silico* function prediction methods can prove instrumental in providing functional insights for proteins where scarce or no experimental data are available.

Structural knowledge is a crucial element in drug design. Protein-structure determination is generally more difficult in eukaryotes than in prokaryotes and archaea, although most drug targets are eukaryotic proteins. One way to solve this problem is to steer the structure determination effort towards a prokaryotic protein that is homologous to the eukaryotic proteins of interest. Solving the structure of the homologue would provide a low resolution scaffold for drug design studies on the target protein. In order for the process to work we have to make sure that the functions of the two proteins are similar. A spectacular example of this procedure is the work on potassium channels performed in Roderick MacKinnon's group [18,19]. Voltage-gated potassium channels in eukaryotes mediate neuronal signal transmission by participating in shaping the action potential that travels along the axon. Structural determination of the eukaryotic channel had proved to be difficult and elusive. As a consequence, MacKinnon and his group decided to turn to a homologue that they found in the genome of the archaebacterium *Aeropyrum pernyx*. Although the authors went on to perform electrophysiology experiments to assess the actual function of their archaebacterial homologue, the first stages of their analysis were based on computational tools such as sequence similarity searches and motif conservation analysis. The resolved crystal structure of the archaebacterial voltage-gated potassium channel was the outcome [18]. MacKinnon and his co-workers went even further and demonstrated that toxins known to bind to the eukaryotic channel also bind to the archaebacterial channel. This suggests that, despite the evolutionary distance between eukaryotes and archaea, homologous structures can offer invaluable insights for lead design. Such structural knowledge could be a crucial first step in the lead design of novel targets, bypassing the traditional and sometimes painful way of solving the structure of the target itself. Once this indirect structural knowledge about the target is available lead design could utilize potential binding sites on the surface of the target. This could be assisted and enhanced by various computational methods that identify the active site (or binding site) of a protein.

The problem that has the highest economic impact on drug development is that of adverse drug reactions (ADRs), detected during clinical trials or with drugs that have already been marketed, which can potentially lead to severe side effects. The presence of a competing binding site, on proteins other than the target, can limit the drug efficacy or even cause drug interference with a metabolic pathway that is unrelated to the target. Although some proteins competing for the same drug might be detected by homology others might not be identified. On the whole, ligand binding can be regarded as a local property of the protein, not usually requiring overall sequence or structure conservation. Methods that predict binding sites based on local conservations or biophysical properties could be very important in early screening of lead compounds, giving some indication of a possible drug interaction with competing targets [20].

Homology-based annotation transfer

SWISS-PROT [21] is the most comprehensive public database having large-scale and detailed annotations of protein function. A team of experts continuously adds annotations that are primarily extracted from scientific literature [22]. In March 2005 SWISS-PROT contained >172,000 different proteins but in other publicly available databases there is at least ten times this number of protein sequences. For many proteins that are not in SWISS-PROT we can detect an evolutionary relationship to proteins that are in SWISS-PROT by their sequence similarities.

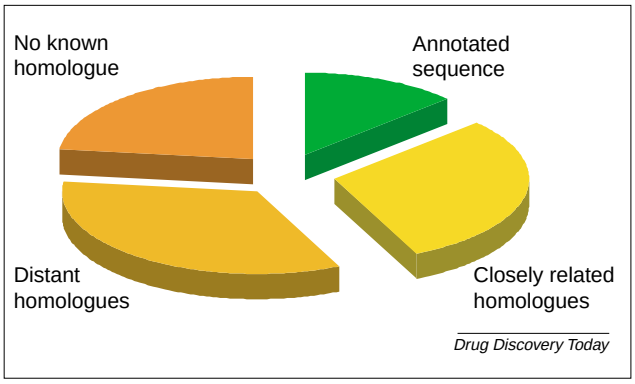

FIGURE 2

The distribution of known sequences. Out of approximately two million known protein sequences, <15% have comprehensive, experimental, functional characterization (green). The challenge of *in silico* function prediction in this context is to devise tools that will reveal the function of the remaining 85% of proteins that are of potential interest in drug discovery. Sequence homology is a powerful tool for this exploration and ~33% of these sequences have a very high sequence homology to one of the annotated sequences that enables an accurate annotation transfer. Another 37% have distant homology to an annotated sequence that does not allow automatic annotation transfer but could enable manual, expert analysis of function. Finally, ~25% of known protein sequences have no known homologue, thus, the prediction of their function has to be done *de novo*.

Concept of homology-based transfer

Proteins that are evolutionarily related (i.e. have a common ancestor) are commonly referred to as homologues and very close homologues often have a similar function [12]. Homology-based transfer of functional annotations is a naïve prediction method that assigns proteins that have not been annotated with the function of their annotated homologues. One significant problem of this approach is that it is not clear what level of sequence similarity ascertains that two proteins have the same function [23–28]. There are essentially three sources of error for homology-based annotation transfer [29–33]: mistakes in the method that identifies the homologues as similar enough to transfer the annotation (often this could result from the fact that most proteins have multiple domains [7]); mistakes in the original database annotation of the homologue; and mistakes caused by evolutionary divergence (i.e. the homologue has lost the function and/or acquired another one). Another obvious limitation of homology-based transfer is with proteins for which we cannot find any annotated homologues (Figure 2).

Although powerful, homology-based transfer is inaccurate and limited

Because the idea behind annotation transfer is so simple and appealing it is widely used. Recently, a few studies attempted to establish the accuracy of annotation transfer on a large scale [34,35]. They concluded that annotation transfer is only an accurate method for deducing the function for pairs of proteins that are highly similar in sequence. By analyzing 105 entire proteomes [36], it was estimated that <35% of all proteins could be annotated automatically when accepting errors ≤5%. Even when we allow error rates >40% we still have no annotation for >30% of all proteins [11].

One example of unsuccessful sequence homology-based transfer was found by Keller *et al.* [37]. They report the case of CbiT, a protein involved in the biosynthesis of vitamin B12. Sequence-based annotation transfer identified it as a decarboxylase. When the structure of CbiT became available Keller and his co-workers were able to find structural similarities to methyltransferases. This structure-based function prediction was later corroborated by experiments. Thus, predictions that are made on the basis of sequence homology should be handled with caution.

Other annotation transfer methods

Sequence motifs and patterns increase the power of annotation transfer

Examining entire sequences, two proteins might not appear to be very similar but they could have a short sequence motif in common that is specific to a particular function. A few databases are dedicated to identifying such motifs. Many of them include searching tools that compare every protein submitted with all known motifs (Figure 3a). For example, PROSITE [38] contains manually selected biologically important motifs. It has three types of signatures: patterns, rules and profiles and each signature represents a different automated method for searching motifs. Although the two most local signatures (patterns and rules) might span over just a few residues, profiles extend the similarity to the level of entire domains. Another popular library is Pfam [39], in which motifs typically span over entire domains (~100 residues [7]). It is based on a combination of expert curation and automated analysis. The annotation in Pfam includes a description of each family and links to other resources and literature references. Other well-known motif databases include BLOCKS [40] and PRINTS [41].

Information on 3D structure can refine annotation transfer

Knowledge of the 3D structure of a protein allows for the inference of evolutionary and functional relations that are not apparent from the sequence [12,42]. However, having knowledge of the 3D structure is not enough to infer function. Functional hypotheses can be made from 3D structures for ~20–50% of the hypothetical proteins, for which structural genomics determined the structures [43,44].

3D motifs widen the perspective even more

Sometimes it is possible to find 3D motifs that are associated with specific functions (e.g. active sites or binding sites). This is similar to the case involving sequence motifs and patterns. Although the concept of 3D motifs is not new, the advent of structural genomics has boosted the development of methods that infer function from structure and, over the past few years, it has become a

major research field. Libraries of these 3D motifs with known function have begun to evolve (Figure 3b) [45]. One example is PROCAT [46], a database of 3D enzyme active sites that can be queried for specific functional signatures. In general, local structural similarities can be established by all atoms [47], pseudo atoms [48], RMSD calculations or surface patch comparisons [49]. In addition, hybrid motifs incorporating information from sequence and structure, as well as from the literature, have also been used to predict function [50,51].

De novo prediction of functional features

All the approaches discussed so far require the presence of an experimentally annotated homologue. If no homologue exists or if the particular functional phenotype is not identifiable or experimentally tractable can we predict function *de novo* (i.e. without transferring annotation from other proteins)? Several methods have been developed recently with this aim as the ultimate goal. The bad news is that automated methods will probably never predict function at high levels of detail without errors. The good news is that some methods achieve levels of accuracy and detail that are useful and might even be good enough to provide evidence as reliable as more costly high-throughput experiments.

Subcellular localization

The identification of the native compartment (i.e. its localization) of a hypothetical protein is one crucial step closer to identifying its role. Detecting the presence of a well-known disease-related protein in an unsuspected cell compartment can provide new directions in the study of that disease [52]. Despite large-scale experiments that determine the subcellular localization in yeast, homology-based inferences are available for less than a third of all human proteins because of the lack of annotated homologues. Methods that predict subcellular localization are a prominent example for the success of *de novo* methods. The best methods use either signal sequences [53] or more generalized features, such as overall amino acid composition and predicted structural features [54]. Methods have improved continuously over the past 5–10 years to reach levels of performance similar to those for high-throughput experiments [54]. One task that remains to be addressed is refining the resolution of the predictions (reliable predictions are only available for some of the major compartments). For example, one particular aspect of compartmentalization is the identification of integral membrane proteins. Although the best methods reliably distinguish membrane from non-membrane proteins [55–58] the prediction of a particular membrane the protein is inserted into (e.g. Golgi or extracellular) has yet to be solved satisfactorily.

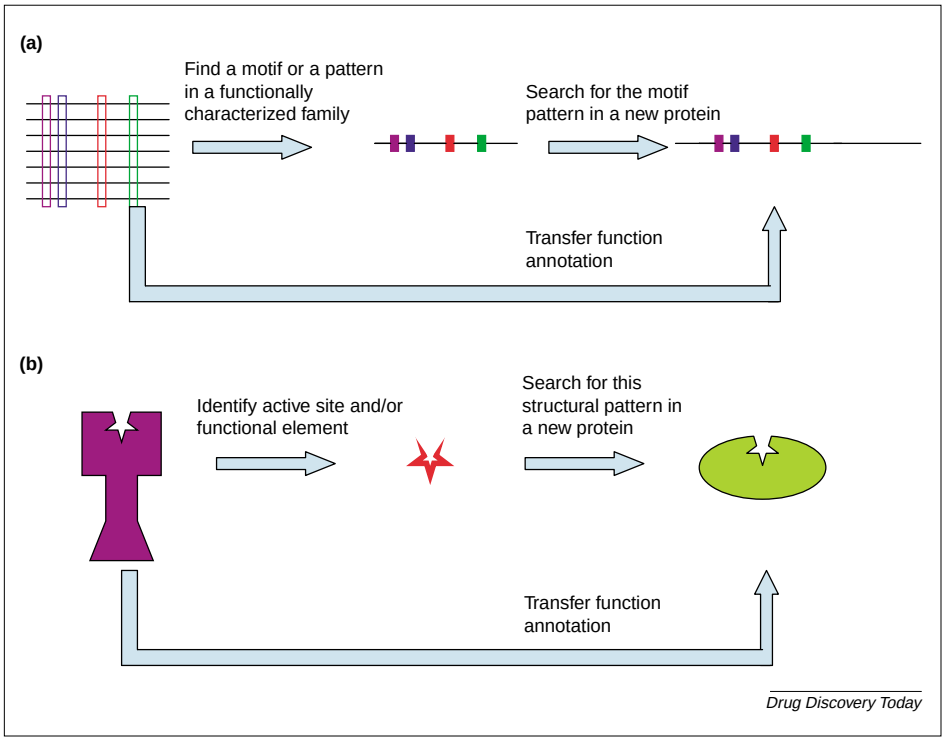

FIGURE 3

Using motifs and patterns to analyze function. Identifying distinct motifs and patterns in annotated sequences could enable the prediction of function for non-annotated proteins that contain the same motifs. **(a)** Multiple sequence alignment of a functionally annotated protein family can lead to the identification of motifs. The discovery of such motifs in a target protein could allow annotation transfer even in the absence of a significant level of overall sequence similarity. **(b)** Structural motifs, analyzing experimentally characterized structures can reveal a structural determinant that is strongly relevant for a particular function. Finding such a determinant in a protein of newly determined 3D structure might enable the transfer of the functional annotation from the original structure to the new one.

Posttranslational modifications

Posttranslational modifications are other features relevant for drug discovery that can be reliably predicted. Over 325 structural and regulatory posttranslational modifications in proteins are known [59] but prediction methods are currently constrained to a few important modifications. The best results are achieved by combining *de novo* methods with features such as highly conserved sequence motifs and more-complex sequence patterns or predicted structural properties such as solvent accessibility [60]. Prominent methods include targeting N-terminal signal peptide cleavage sites, proteolytic cleavage and, more specifically, proteasome cleavage sites, phosphorylation sites, lipid modification and *N*- and *O*-glycosylations [60].

Binding sites and functional residues

Methods that identify ligand binding sites will probably have a major impact on drug discovery strategies [61]. Comprehensive databases of ligand binding sites have been compiled to assist predictions [62]. Most of the available methods focus either on the evolutionary conservation [63] or on the electrostatic properties for binding site characterization [64,65] (in this case

knowledge of the protein structure is required). Another type of a successful *de novo* method identifies generic features in residues that are not specific to any particular group of related proteins and do not require the knowledge of any experimentally characterized motif. Predictions are typically based on the identification of particular biophysical traits. For example, residues that bind to other proteins have distinct biophysical, structural and evolutionary characteristics [66]. ISIS [67], a system of neural networks trained on evolutionary information and predicted structural features, identifies the residues within a query sequence that probably bind to other proteins.

Protein–protein interactions

Every protein has a biological function, although most cellular processes are carried out by groups of proteins that physically interact with each other. Having knowledge of the interaction partner(s) helps to unveil the role of a new protein. Therefore, an extensive research effort is invested in both experimental and computational methods designed to unravel protein–protein interactions [68,69]. The focus over the past few years has been on collecting maps of interactions for entire proteomes that are as complete as possible [70–72]. To date, *de novo* methods have not been sufficiently generic (i.e. applicable to all protein–protein interactions, all protein families, all pathways or processes and all species) or successful in predicting unknown interacting partners *in silico*. However, predictions can already help in experimental design [68].

Association through networks

Pairs of physically interacting proteins are grouped into networks of associations (using various computational and manual methods) that trivially share some features with other systems subject to population growth (e.g. Internet, growth of settlements, growth of protein families). These network structures bring about groups of proteins that are associated in local clusters such as pathways and the associations are targeted in microarray studies. Being related through the same or a similar network is a feature that could refine the annotation transfer. A recent study has shown that combining different methods for predicting protein associations could generate functional hypotheses through links to known functional modules [73]. Additionally, all proteins mapped to the same pathway constitute interesting targets for functional studies.

Assigning gene ontology terms from sequence and 3D structure

Hierarchies, schemata and ontologies are often used in an attempt to capture some of the complexity of protein function. Several groups and consortia have introduced relatively comprehensive schemata for protein function. Enzyme classification hierarchy classifies enzymes through four digits [74]. The first digit represents the particular type of enzyme (e.g. ligase), the second digit is the group or bond on which the enzyme acts and the following digits identify more details about the enzymatic reaction [23]. The GO consortium [75] attempts to realize ontology for all proteins. GO distinguishes three categories: molecular function (at the molecular level the protein can catalyze a metabolic reaction or transmit a signal); biological process (a set of many cooperating proteins is responsible for achieving broad biological goals, e.g. mitosis, purine metabolism or signal transduction cascades); and cellular component (this category includes the organization into subcellular compartments, the localization of proteins and macromolecular complexes, e.g. nucleus, telomere, and origin recognition complex).

Groups have recently begun to develop methods that predict the GO classification (for the more-populated GO classes) directly. This breakthrough in the field originated from a collaboration of the groups of Søren Brunak (Center for Biological Sequence Analysis, Copenhagen, Denmark) and Alfonso Valencia (Centro Nacional de Biotecnología, Madrid, Spain). Their objective was to predict classes of cellular function (originally introduced by Monika Riley [76]) from the sequence alone. The method they introduced, dubbed ProtFun, uses complex and hierarchical systems of neural networks [77]. The authors applied their method to annotating functional classes for all human proteins and, recently, to the identification of novel archeal enzymes [78]. The most impressive news from these groundbreaking methods is that detailed aspects of function can be predicted without relying on sequence homology (i.e. for completely uncharacterized proteins).

Whereas ProtFun [77] predicts function from the sequence alone, ProKnow [79] integrates various tools to predict function from the 3D structure. In addition to 3D structure information, ProKnow also exploits sequence alignments, motifs and functional links. The result is the prediction of a GO classification for the queried structure. Several new methods that assign a GO classification to queried sequences have been introduced recently (http://ffas.burnham.org/AFP/Challenges/servers/) [80,81]. Predicting GO classification is now becoming the standard in the field of automated function prediction.

Conclusions

Although genome projects and environmental genomics have the potential to provide a better understanding of disease processes and help to identify new and better targets, the lack of functional annotation for most of the newly sequenced proteins hampers the ability to exploit these data for the faster and more widespread development of new drugs. Predicting protein function can help to identify new targets for known drugs, new functional analogues of known targets in different organisms or cellular compartments, or to detect proteins homologous to the original targets that are more suitable to experimental analysis. As experimental determination of protein function is still costly and slow *in silico* methods can

sometimes give the only useful clues. For many years, homology-based annotation transfer has been the only means of predicting function with some accuracy. People realized the limits of this approach and so they also started to develop new methods that do not have homology as a requisite to infer protein function. In this review, we have presented a quick survey of these methods, from subcellular localization predictions to posttranslational modifications and from predictions of active or binding sites to functional residues and protein–protein interactions. These methods use evolutionary analysis, biophysical properties, geometry or a combination of these features. Altogether, they allow us to refine and to extend annotation of protein function, assisting drug development.

Acknowledgements

Thanks to Jinfeng Liu, Rajesh Nair, Andrew Kernytsky and Kazimierz Wrzeszczynski (Columbia University, USA) for their help in preparing this manuscript. This work was supported by the grants (RO1-GM64633–01) from the National Institutes of Health (NIH), and (RO1-LM07329–01) from the National Library of Medicine (NLM). Last, but not least, thanks to the GeneOntology team of Michael Ashburner (Cambridge, UK) for their gargantuan effort, to Amos Bairoch (SIB, Geneva, Switzerland), Rolf Apweiler (EBI, Hinxton, UK), Phil Bourne (San Diego University, USA) and their crews for maintaining excellent databases and to all experimentalists who enable computational biology by making their data publicly available.

References

1 Fleischmann, R.D. *et al.* (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* 269, 496–512
2 Goffeau, A. *et al.* (1996) Life with 6000 genes. *Science* 274, 546–567
3 *C. elegans* Sequencing Consortium. (1998) Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* 282, 2012–2018
4 Lander, E.S. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature* 409, 860–921
5 Venter, J.C. *et al.* (2001) The human genome. *Science* 291, 1304–1351
6 Venter, J.C. *et al.* (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science* 304, 66–74
7 Liu, J. *et al.* (2004) Automatic target selection for structural genomics on eukaryotes. *Proteins* 56, 188–200
8 Westbrook, J. *et al.* (2003) The protein data bank and structural genomics. *Nucleic Acids Res.* 31, 489–491
9 Drews, J. (2003) Stategic trends in the drug industry. *Drug Discov. Today* 8, 411–420
10 Lehman Bothers, McKinsey and Company (2001) The fruits of genomics: drug pipelines face indigestion until the new biology ripens.
11 Rost, B. *et al.* (2003) Automatic prediction of protein function. *Cell. Mol. Life Sci.* 60, 2637–2650
12 Whisstock, J.C. and Lesk, A.M. (2003) Prediction of protein function from protein sequence and structure. *Q. Rev. Biophys.* 36, 307–340
13 Cao, Z.W. *et al.* (2005) Computer prediction of drug resistance mutations in proteins. *Drug Discov. Today* 10, 521–529
14 Koch, M.A. and Waldmann, H. (2005) Protein structure similarity clustering and natural product structure as guiding principles in drug discovery. *Drug Discov. Today* 10, 471–483
15 Bork, P. and Koonin, E.V. (1998) Predicting functions from protein sequences–where are the bottlenecks? *Nat. Genet.* 18, 313–318
16 Iliopoulos, I. *et al.* (2003) Evaluation of annotation strategies using an entire genome sequence. *Bioinformatics* 19, 717–726
17 Karp, P.D. (1998) What we do not know about sequence analysis and sequence databases. *Bioinformatics* 14, 753–754
18 Jiang, Y. *et al.* (2003) X-ray structure of a voltage-dependent K+ channel. *Nature* 423, 33–41
19 Ruta, V. *et al.* (2003) Functional analysis of an archaebacterial voltage-dependent K+ channel. *Nature* 422, 180–185
20 Shulman-Peleg, A. *et al.* (2004) Recognition of functional sites in protein structures. *J. Mol. Biol.* 339, 607–633
21 Boeckmann, B. *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* 31, 365–370
22 Junker, V. *et al.* (2000) The role SWISS-PROT and TrEMBL play in the genome research environment. *J. Biotechnol.* 78, 221–234
23 Todd, A.E. *et al.* (2001) Evolution of function in protein superfamilies, from a structural perspective. *J. Mol. Biol.* 307, 1113–1143
24 Shah, I. and Hunter, L. (1997) Predicting enzyme function from sequence: a systematic appraisal. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 5, 276–283
25 Ouzounis, C. *et al.* (1998) Are binding residues conserved? *Pac. Symp. Biocomput.* 401–412
26 Rost, B. (2002) Enzyme function less conserved than anticipated. *J. Mol. Biol.* 318, 595–608
27 Nair, R. and Rost, B. (2002) Sequence conserved for sub-cellular localization. *Protein Sci.* 11, 2836–2847
28 Wrzeszczynski, K.O. and Rost, B. (2004) Cataloguing proteins in cell cycle control. *Methods Mol. Biol.* 241, 219–233
29 Fraser, C.M. *et al.* (2000) Microbial genome sequencing. *Nature* 406, 799–803
30 Kyrpides, N.C. and Ouzounis, C.A. (1998) Errors in genome reviews. *Science* 281, 1457
31 Devos, D. and Valencia, A. (2001) Intrinsic errors in genome annotation. *Trends Genet.* 17, 429–431
32 Koonin, E.V. *et al.* (2002) The structure of the protein universe and genome evolution. *Nature* 420, 218–223
33 Iyer, L.M. *et al.* (2001) Quod erat demonstrandum? The mystery of experimental validation of apparently erroneous computational analyses of protein sequences. *Genome Biology* 2 DOI: 10.1186/gb-2001-2-12-research0051 (http://genomebiology.com)
34 Devos, D. and Valencia, A. (2000) Practical limits of function prediction. *Proteins* 41, 98–107
35 Hegyi, H. and Gerstein, M. (2001) Annotation transfer for genomics: measuring functional divergence in multi-domain proteins. *Genome Res.* 11, 1632–1640
36 Carter, P. *et al.* (2003) PEP: Predictions for entire proteomes. *Nucleic Acids Res.* 31, 410–413
37 Keller, J.P. *et al.* (2002) The crystal structure of MT0146/CbiT suggests that the putative precorrin-8w decarboxylase is a methyltransferase. *Structure (Camb)* 10, 1475–1487
38 Sigrist, C.J. *et al.* (2002) PROSITE: a documented database using patterns and profiles as motif descriptors. *Brief. Bioinform.* 3, 265–274
39 Bateman, A. *et al.* (2004) The Pfam protein families database. *Nucleic Acids Res.* 32, 138–141
40 Henikoff, S. *et al.* (1999) Blocks+: a non-redundant database of protein alignment blocks derived from multiple compilations. *Bioinformatics* 15, 471–479
41 Attwood, T.K. *et al.* (2003) PRINTS and its automatic supplement, prePRINTS. *Nucleic Acids Res.* 31, 400–402
42 Thornton, J.M. *et al.* (2000) From structure to function: approaches and limitations. *Nat. Struct. Biol.* 7(*Suppl*), 991–994
43 Laskowski, R.A. *et al.* (2003) From protein structure to biochemical function? *J. Struct. Funct. Genomics* 4, 167–177
44 Goldsmith-Fischman, S. and Honig, B. (2003) Structural genomics: computational methods for structure analysis. *Protein Sci.* 12, 1813–1821
45 Jones, S. and Thornton, J.M. (2004) Searching for functional sites in protein structures. *Curr. Opin. Chem. Biol.* 8, 3–7
46 Wallace, A.C. *et al.* (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
47 Stark, A. and Russell, R.B. (2003) Annotation in three dimensions. PINTS: patterns in non-homologous tertiary structures. *Nucleic Acids Res.* 31, 3341–3344
48 Kleywegt, G.J. (1999) Recognition of spatial motifs in protein structures. *J. Mol. Biol.* 285, 1887–1897
49 Ferre, F. *et al.* (2004) SURFACE: a database of protein surface regions for functional annotation. *Nucleic Acids Res.* 32, 240–244
50 Di Gennaro, J.A. *et al.* (2001) Enhanced functional annotation of protein sequences via the use of structural descriptors. *J. Struct. Biol.* 134, 232–245
51 Fetrow, J.S. and Skolnick, J. (1998) Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and T1 ribonucleases. *J. Mol. Biol.* 281, 949–968
52 Quesada, I. and Soria, B. (2004) Intracellular location of KATP channels and sulphonylurea receptors in the pancreatic beta-cell: new targets for oral antidiabetic agents. *Curr. Med. Chem.* 11,

2707–2716
53 Bendtsen, J.D. *et al.* (2004) Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol.* 340, 783–795
54 Nair, R. and Rost, B. (2005) Mimicking cellular sorting improves prediction of subcellular localization. *J. Mol. Biol.* 348, 85–100
55 Chen, C.P. *et al.* (2002) Transmembrane helix predictions revisited. *Protein Sci.* 11, 2774–2791
56 Melen, K. *et al.* (2003) Reliability measures for membrane protein topology prediction algorithms. *J. Mol. Biol.* 327, 735–744
57 Jacoboni, I. *et al.* (2001) Prediction of the transmembrane regions of beta-barrel membrane proteins with a neural network-based predictor. *Protein Sci.* 10, 779–787
58 Bigelow, H. *et al.* (2004) Prediction of transmembrane beta-barrels for entire proteomes. *Nucleic Acids Res.* 32, 2566–2577
59 Garavelli, J.S. (2003) The RESID Database of Protein Modifications: 2003 developments. *Nucleic Acids Res.* 31, 499–501
60 Nakai, K. (2001) Prediction of *in vivo* fates of proteins in the era of genomics and proteomics. *J. Struct. Biol.* 134, 103–116
61 Campbell, S.J. *et al.* (2003) Ligand binding: functional site location, similarity and docking. *Curr. Opin. Struct. Biol.* 13, 389–395
62 Stuart, A.C. *et al.* (2002) LigBase: a database of families of aligned ligand binding sites in known protein sequences and structures. *Bioinformatics* 18, 200–201
63 Lichtarge, O. and Sowa, M.E. (2002) Evolutionary predictions of binding surfaces and interactions. *Curr. Opin. Struct. Biol.* 12, 21–27
64 Elcock, A.H. (2001) Prediction of functionally important residues based solely on the computed energetics of protein structure. *J. Mol. Biol.* 312, 885–896
65 Ondrechen, M.J. *et al.* (2001) THEMATICS: a simple computational predictor of enzyme function from structure. *Proc. Natl. Acad. Sci. U. S. A.* 98, 12473–12478
66 Ofran, Y. and Rost, B. (2003) Analysing six types of protein-protein interfaces. *J. Mol. Biol.* 325, 377–387
67 Ofran, Y. and Rost, B. (2003) Predict protein-protein interaction sites from local sequence information. *FEBS Lett.* 544, 236–239
68 Valencia, A. and Pazos, F. (2002) Computational methods for the prediction of protein interactions. *Curr. Opin. Struct. Biol.* 12, 368–373
69 Salwinski, L. and Eisenberg, D. (2003) Computational methods of analysis of protein-protein interactions. *Curr. Opin. Struct. Biol.* 13, 377–382
70 Xenarios, I. *et al.* (2002) DIP, the database of interacting proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.* 30, 303–305
71 Bader, G.D. *et al.* (2003) BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res.* 31, 248–250
72 von Mering, C. *et al.* (2005) STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Research* 33, 433–437
73 Bowers, P.M. *et al.* (2004) Use of logic relationships to decipher protein network organization. *Science* 306, 2246–2249
74 Webb, E.C. (1992) *Enzyme Nomenclature 1992. Recommendations of the Nomenclature committee of the International Union of Biochemistry and Molecular Biology*, Academic Press
75 Ashburner, M. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* 25, 25–29
76 Riley, M. (1993) Function of the gene products in *Escherichia coli*. *Microbiol. Rev.* 57, 862–952
77 Jensen, L.J. *et al.* (2003) Prediction of human protein function according to Gene Ontology categories. *Bioinformatics* 19, 635–642
78 Jensen, L.J. *et al.* (2002) Prediction of novel archaeal enzymes from sequence-derived features. *Protein Sci.* 11, 2894–2898
79 Pal, D. and Eisenberg, D. (2005) Inference of protein function from protein structure. *Structure (Camb)* 13, 121–130
80 Enault, F. *et al.* (2003) Annotation of bacterial genomes using improved phylogenomic profiles. *Bioinformatics* 19 (*Suppl 1*), i105–i107
81 Szafron, D. *et al.* (2004) Proteome analyst: custom predictions with explanations in a web-based tool for high-throughput proteome annotations. *Nucleic Acids Res.* 32, 365–371