THOMSON REUTERS

Current Opinion in Drug Discovery & Development 2009 **12**(3):408-419
© Thomson Reuters (Scientific) Ltd ISSN 2040-3437

REVIEW

New in protein structure and function annotation: Hotspots, single nucleotide polymorphisms and the 'Deep Web'

Yana Bromberg[1,2,6]*, Guy Yachdav[1,2,3,6], Yanay Ofran[4], Reinhard Schneider[5] & Burkhard Rost[1,2,3,6]

Addresses
[1]Department of Biochemistry and Molecular Biophysics, Columbia University,
630 West 168th Street, New York, NY 10032, USA
Email: bromberg@rostlab.org

[2]Columbia University Center for Computational Biology and Bioinformatics (C2B2),
1130 St Nicholas Ave, New York, NY 10032, USA

[3]NorthEast Structural Genomics Consortium (NESG),
1130 St Nicholas Ave, New York, NY 10032, USA

[4]The Mima and Everard Goodman Faculty of Life Science,
Bar Ilan University, Ramat-Gan, Israel

[5]European Molecular Biology Laboratory (EMBL),
Meyerhofsstrasse 1, 69117 Heidelberg, Germany

[6]Biosof LLC, 215W 95th Street,
Suite 4E, New York, NY 10032, USA

*To whom correspondence should be addressed

The rapidly increasing quantity of protein sequence data continues to widen the gap between available sequences and annotations. Comparative modeling suggests some aspects of the 3D structures of approximately half of all known proteins; homology- and network-based inferences annotate some aspect of function for a similar fraction of the proteome. For most known protein sequences, however, there is detailed knowledge about neither their function nor their structure. Comprehensive efforts towards the expert curation of sequence annotations have failed to meet the demand of the rapidly increasing number of available sequences. Only the automated prediction of protein function in the absence of homology can close the gap between available sequences and annotations in the foreseeable future. This review focuses on two novel methods for automated annotation, and briefly presents an outlook on how modern web software may revolutionize the field of protein sequence annotation. First, predictions of protein binding sites and functional hotspots, and the evolution of these into the most successful type of prediction of protein function from sequence will be discussed. Second, a new tool, comprehensive in silico mutagenesis, which contributes important novel predictions of function and at the same time prepares for the onset of the next sequencing revolution, will be described. While these two new sub-fields of protein prediction represent the breakthroughs that have been achieved methodologically, it will then be argued that a different development might further change the way biomedical researchers benefit from annotations: modern web software can connect the worldwide web in any browser with the 'Deep Web' (ie, proprietary data resources). The availability of this direct connection, and the resulting access to a wealth of data, may impact drug discovery and development more than any existing method that contributes to protein annotation.

Keywords Binding site prediction, function and structure prediction, functional site prediction, *in silico* (computational) mutagenesis

Abbreviations

nsSNP non-synonymous single nucleotide polymorphism, **PPI** protein-protein interaction

Introduction

Modern biology increasingly relies on high-throughput techniques. One particular experiment no longer precisely answers one question; instead, partial answers are combined in different ways to answer many questions. This high-throughput trend challenges computational biologists to quickly extract as much useful information from the data as possible. Furthermore, novel drugs arrive on the market if and only if they undergo detailed analyses that investigate their micro-molecular (binding sites) as well as macro-molecular (pathways) actions. Generally, the challenge to computational biology is to annotate the 3D structure and function of as many proteins as possible at reasonable accuracy levels.

At the beginning of 2009, the genomes of more than 900 organisms had been completely sequenced [1], and more than 7 million sequences were stored in protein databases (*www.ebi.ac.uk/swissprot/sptr_stats/index.html*) [2]. Many large-scale efforts have aimed to provide annotations for these sequences, for example, SWISS-PROT [3], Gene Ontology (GO) [4], the Human Proteomics Initiative (HPI; *www.expasy.org/sprot/hpi/hpi_stat.html*) [3], and the Protein Structure Initiative (PSI) [5]. GO, one of the major achievements of the last decade, systematically describes biological function using ontologies that encompass molecular function, cellular localization and biological processes. However, not even the best ontologies can rely solely on expert annotations, because the experts are unable to keep up with the rapid influx of new data. From among the ~ 7 million proteins of known sequence to date, only ~ 87,000 proteins have been manually annotated with a GO term (*www.ebi.ac.uk/GOA/uniprot_release.html*) [4] and ~ 5500 of the human proteins have GO numbers manually assigned by UniProt (*www.ebi.ac.uk/GOA/human_release.html*). Most methods that predict function use classifications such as GO, and base their inferences on sequence similarity to proteins of experimentally characterized function [6-8]. SWISS-PROT undoubtedly constitutes the most comprehensive source of expert-curated annotations of protein function. However, even this excellent resource infers as many annotations through sequence similarity as it does through explicit experimental support.

For some of the proteins for which some aspects of function have been probed experimentally, there are also experimentally determined 3D structures. However, for most proteins, experimental structures are not available [8,9]. Experimental structures are available for ~ 55,000 known proteins (< 1%), of which less than half are distinct, at 95% sequence identity (*www.rcsb.org/pdb/statistics/clusterStatistics.do*) [10]. For any detailed experiment that probes a particular protein, it is ultimately necessary to identify the mechanistic details describing the protein structure. Structural genomics efforts seek to experimentally determine the 3D structures for most protein families in a manner that optimizes the impact of each experimental structure for modeling [11-13]. The particular combination of many structures with many sequences enables novel inferences about protein function to be made [14,15]. The PSI, which drives structural genomics in the US [5], relies on computational biology to enrich the experimental data by many orders of magnitude [12]. This enrichment is currently mostly confined to the application of comparative modeling. Protein structure prediction methods have improved significantly over the last decade [16,17]. Methods such as Rosetta [18] and I-TASSER [19] can now generate good models for proteins that are similar to proteins of known structure, even if this similarity is not easily detectable. However, it remains unclear to what extent these improved prediction methods aid functional annotation.

Structural genomics radically altered many assumptions; for example, it was discovered that knowledge of the details of 3D protein structures does not automatically provide information about protein function [20,21]. However, knowledge of these structural details may help in the inference of function. This realization spawned the development of many methods for the prediction of protein function from 3D structure [22]. Generally, the choice of computational tools for the prediction of function depends on the type of desired prediction and on the amount of information available for the particular protein [23]. Currently available methods for protein function prediction cover a large number of techniques, from annotation transfer models using sequence, structure, and/or local motifs [24-26] to automatic text mining [27] and function predictions using machine learning [28-32]. Several approaches to the problem of obtaining sufficient data for protein function prediction combine various sources of information to improve the accuracy of the function predictions [33-35].

From molecular detail to the ambiguity of the system

More recently, the trends in protein function prediction have split into two opposite directions. The first is at the level of the system, annotating the relevance of a certain protein to a phenotype [36,37], disease [38], module [39] or pathway [40]. This approach frequently focuses on coarse-grained aspects of function, analyzing similarities between entire modules/pathways or the similarity in the responses of proteins to particular experimental conditions (eg, correlated coexpression of responses to knockouts) [41,42]. This perspective ultimately ceases to be concerned with the physicochemical details of any particular protein. The second direction that protein function prediction has taken pertains to the increasingly detailed level of predicting molecular function from as much information as possible, including 3D structures [22,43], models, and stability assays [44]. This review will focus on some methods that have evolved in this direction over the last 2 years. In particular, two particular types of approach will be discussed: (i) those that predict binding sites; and (ii) those that identify patterns of functionally important residues based on the analysis of effects of non-synonymous SNPs (nsSNPs) or point mutants. These two novel types of functional annotation are relevant for drug design because they directly translate into discovering ways to alter a specific protein activity.

Worldwide web surfing software to complement experimental methods

The final focus of this review is on a new development that, initially, may appear to be unrelated to methods predicting aspects of protein function and structure. This novel advance connects two very different types of repositories: the 'general' worldwide web (typically interfaced through the Google search engine on any internet-enabled machine anywhere in the world), and the even larger underlying repository of data that is not visible to current search engines (such as data contained within publications or public databases that are not indexed for regular web searches). The automatic extraction of otherwise 'hidden' functional and structural protein information from these (usually experimentally derived) resources is, effectively, at least equivalent in information gain to making *de novo* predictions or attempting to transfer annotations between similar molecules, if not more useful. Thus, accessing these supplies of data, as a novel method for the retrieval of molecular information, creates new perspectives on protein annotation.

Predicting binding sites

Prediction of binding sites differs by type of interaction and amount of available information

Some types of binding site can be recognized from 3D protein structures with high levels of success [24,45,46]. Moreover, a recent study has illustrated the ability to differentiate proteins, based on the type of small molecule ligands that they bind, using only structural information about atoms in the binding site [47]. As high-accuracy experimental structures are not available for the majority of proteins, some methods predict protein binding sites [48] from predicted structures [26], or directly from protein sequences [30,49].

The binding sites of small substrates are very specific; computationally, such sites have to be inferred from sequence motifs (ie, by homology). Although largely successful, such methods are confined to a tiny subset of all binding sites. Generic prediction methods that identify yet-unknown binding sites have either not been successful to date, or they have to be specifically developed for particular substrates. One example of a successful method of this class pertains to the prediction of metal-binding sites [30,31]. Catalytic active sites are also extremely difficult to predict by methods other than motif-based approaches [49,50]. Arguably, the most successful binding site predictions from sequence alone pertain to the prediction of DNA- and RNA-binding sites [28,29].

Protein-protein binding arguably makes up the largest class of natural protein interactions [51]. Not surprisingly, a whole spectrum of methods exists that predict such binding sites [52,53]. Protein-protein interaction (PPI) sites are rather different from sites that bind small ligands, nucleic acids, metal ions and even small peptides. Interfaces between proteins and smaller substrates are typically cavities and concave clefts [54,55]. However, proteins tend to bind to each other through much larger and more structurally intricate surfaces [56,57].

The most accurate identification of PPI sites is accomplished by analyzing high-resolution experimental 3D structures of the protein-protein complex. In the absence of experimental structures, computational models for complexes have been used to identify PPI sites [58]. However, protein-docking methods are still far from accurate enough to accomplish this feat for an abstract pair of molecules [59]. In fact, while docking methods could help in the identification of interaction sites, it is more common for docking solutions to rely on other methods for the prediction of PPI sites (such as those described below).

There are several different types of PPI. Some PPIs are obligatory (eg, between chains that do not function separately), and others are transient (eg, between chains that have a molecular function in both their bound and unbound states). Each type of interaction may be stabilized by different mechanisms, and therefore the interaction sites involved may have different characteristics [60].

Early binding-site prediction methods were developed using the little experimental data available, and were therefore mostly based on general, and often theoretical, parameters. For example, the method introduced by Kini and Evans relied on the observation of an abundance of proline in regions flanking PPI sites [61]. Jones and Thornton introduced a method that used experimental 3D protein structures to predict whether a particular patch on the surface was likely to be an interaction site based on its topology, solvent accessibility and hydrophobicity [62]. Some of the more recent binding site prediction methods used a similar concept to that used by Jones and Thornton, and appeared to improve performance by using larger datasets with machine learning (eg, Bayesian networks) [63]. Other methods have also used the concept of patches on the protein surface to identify generic binding sites through evolutionary profiles; this idea was pioneered by the developers of the ConSurf tool [64] and was also used in HotPatch [55]. More recent methods, however, replace the concept of a 'patch' on the surface of the protein with the analysis of individual residues. Some of these methods use only sequence and sequence-derived features to predict interaction sites [60,65,66], but most require knowledge of the full experimental 3D structure of a protein [67-69]. Another category of protein binding site prediction methods relies on external sources of information, such as knowledge of specific PPIs, in an attempt to identify sequence motifs that may define PPI sites [70,71]. Similarly, it has been suggested that PPI data may be used to search for pairs of positions in multiple alignment matrices that have co-evolved in interacting proteins, and to identify these as putative interaction sites [72].

Protein-protein binding sites as new frontiers in drug design and development

In the past, PPI sites that are not targeted by native small-molecule ligands have been avoided in the context of drug design [57,73]. However, recent studies have argued strongly for the design of molecules that target these sites [74]. Moreover, several drugs that are currently in development target PPI sites. Benzodiazepinedione and nutlin are two examples of drug classes that target PPI sites, and anticancer drugs belonging to both classes are currently under investigation. Although benzodiazepinedione and nutlin are structurally dissimilar, both target the interface between the protein HDM2 (human double minute 2) and the tumor suppressor p53. The interaction between HDM2 and p53 is believed to inhibit the tumor suppressing activity of p53. The specific binding of benzodiazepinedione or nutlin to the p53-binding site on HDM2 prevents the interaction between the two proteins and enables tumor suppression by p53 [75,76]. Because the targeting of PPI sites is becoming increasingly popular in drug development [74,77], the importance of methods that predict potential interaction sites is increasing.

Another advantage of targeting PPI sites for therapeutic purposes is that the binding of the drug can specifically interfere with a damaging interaction without necessarily disrupting any other essential function of the binding partners; that is, these drugs are likely to reduce unwanted side effects, such as those associated with complete expression knockouts. Furthermore, given the size of PPIs (~ 2000 Å^2) and the relative ease with which steric hindrance can disrupt an interaction, many different small molecules could be developed, each of which may be able to manipulate the same PPI (as demonstrated by benzodiazepinedione and nutlin, which bear no structural similarity to each other but abrogate the same interaction).

Binding hotspots accurately identified from sequences

Not all residues at an interface contribute equally to the binding energy. In fact, most of the binding energy of a given interaction is associated with just a few residues, the so-called 'hotspots' [78,79]. Although all of the residues in a binding site contribute to the binding energy to some extent, those closer to the substrate contribute more. Is there anything more to the concept of hotspots than this? Protein-protein binding sites that are known in detail from experimental high-resolution structures are extremely diverse [60], and the only obvious commonality between them is hotspots [80]. This finding clearly extends beyond the triviality of 'some residues are more important than others' and thereby underlines the importance of targeting such generic sites.

A common method used to explore the importance of a residue to a particular interaction involves mutating it, typically to alanine, and measuring the effect of this substitution on the interaction [79]. This is often done sequentially on a large scale in a procedure known as an 'alanine scan'. Many experiments have demonstrated that most interface residues can be mutated without affecting the affinity of the protein to its partners [81]. Those few residues that, upon mutation, change the affinity of the protein for its substrate are often defined as hotspots [78]. Overall, less than 5% of the residues in a typical 1200 to 2000 Å^2 interface contribute more than 2 kcal/mol to the binding energy. In small interfaces, this can correspond to as little as a single residue [78]. This fact interestingly coincides with the performance results reported by several binding site prediction methods: high levels of precision (accuracy) but low levels of recall (coverage). That is, when a residue is identified as part of the interaction site, this is usually correct, however, many of the residues in the interface are not identified at all. It has been suggested that the poor recall of interface-predicting methods should be attributed to the fact that some of these methods are actually predicting hotspots, rather than identifying all of the interface residues [80]. Several new methods, databases and analyses, have therefore attempted to identify hotspots explicitly, rather than all interface residues [68,82].

For example, consider the complex of the bacterial ribonuclease barnase and its inhibitor barstar (Figure 1A) [83]. The interface between the two proteins consists of more than 50 residues, of which 26 are on the barnase (Figure 1B). However, in an alanine scan only five of these residues were determined to be critical to the stability of the barnase-barstar complex (Figure 1C). The sequence of one chain of the barnase was used to predict the interaction sites using a sequence-based method (ISIS [65]; Figure 1D), and the structure of the same chain was used to predict interaction sites using a structure-based method (ProMate [46]; Figure 1E). The interaction sites predicted using these two methods were remarkably similar to those covered by the hotspots that were identified through the experimental alanine scans.

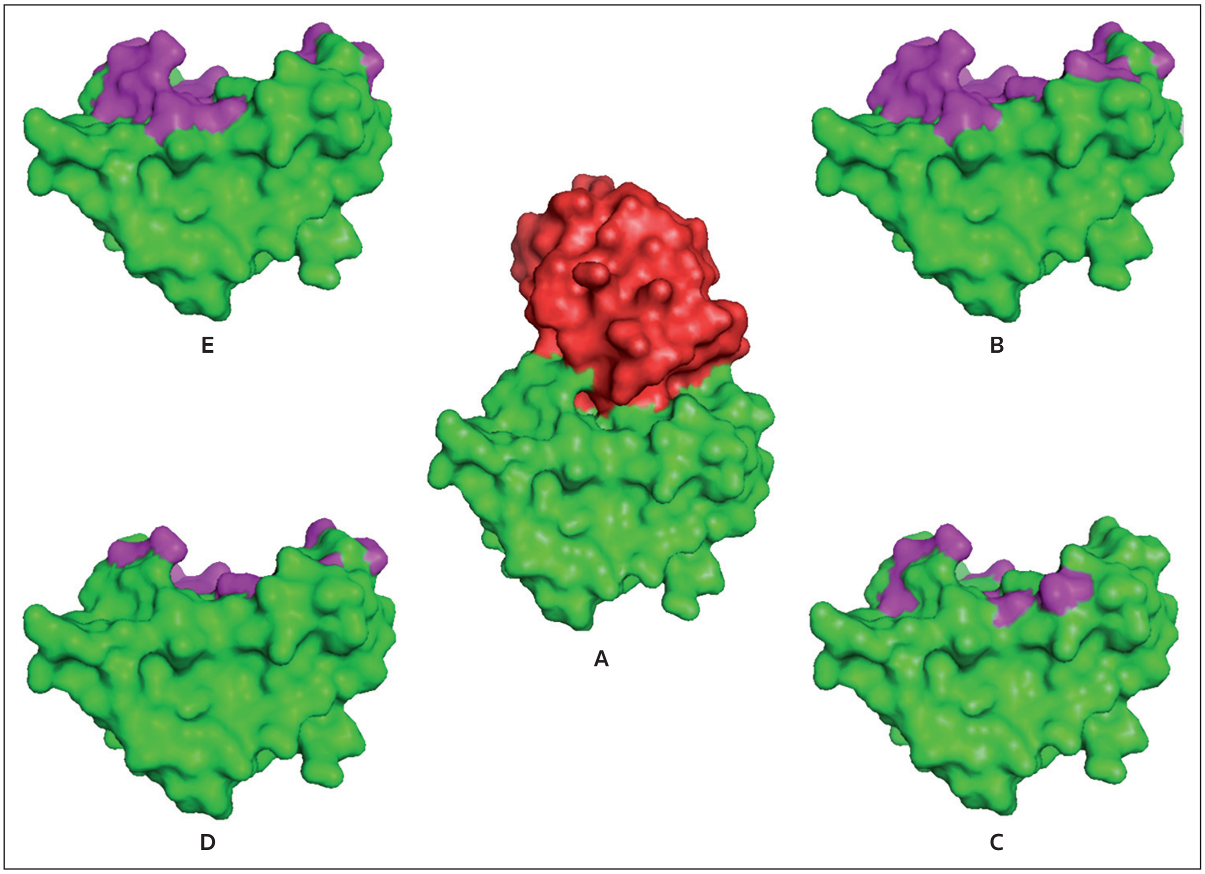
Computational methods aid in interaction modeling

Computational 3D models of target proteins and studies of their interactions with suggested ligands have become useful tools in the development of novel drugs. The process of interaction modeling, however, has been compared to 'solving a three-dimensional jigsaw puzzle' where the pieces 'are the crystal structure of a scaffold protein, a ligand structure, and a set of amino-acid positions that will be mutated to create the binding site' [84]. The procedure involved in the design of a new drug includes iteratively searching through thousands of possible amino-acid sequences to be mutated. The number of suggested sequences to be searched for and mutated could be significantly limited based on information provided by binding site predictions. Additionally, *in silico* mutagenesis studies (discussed below) could contribute to eliminating unlikely candidate sequences and prioritizing those mutations that have the most functional promise.

In silico mutagenesis and protein function prediction

Mutagenesis studies are one experimental means of annotating functionally important residues. For example, site-directed mutagenesis is performed to confirm or reject theories regarding functional involvement of specific residues [85,86]. Alanine scans (described in the previous section) are frequently employed to identify binding hotspots [78,79]. The systematic mutagenesis of large protein fragments, or even of entire proteins, has produced maps of protein function [87,88]. Experimental data are increasingly being complemented by the discovery of natural variations obtained from large-scale sequencing [89,90].

Figure 1. The prediction of protein interaction sites.



(**A**) The complex between the bacterial ribonuclease barnase (green) and its inhibitor barstar (red) (PDB code: 1BRS). (**B**) When barstar is removed from barnase, the residues on the receptors at the interface are revealed. (**C**) Only a few of the residues at the interface – the hotpsots – were found to be critical for stabilizing the complex by experimental alanine scanning. (**D**) and (**E**) Two different prediction methods were used in an attempt to identify the hotspot residues in the area of the interface: (**D**) a sequence-based method (ISIS) [65,80], and (**E**) a structure-based method (ProMate) [46].

ISIS interaction sites identified from sequence

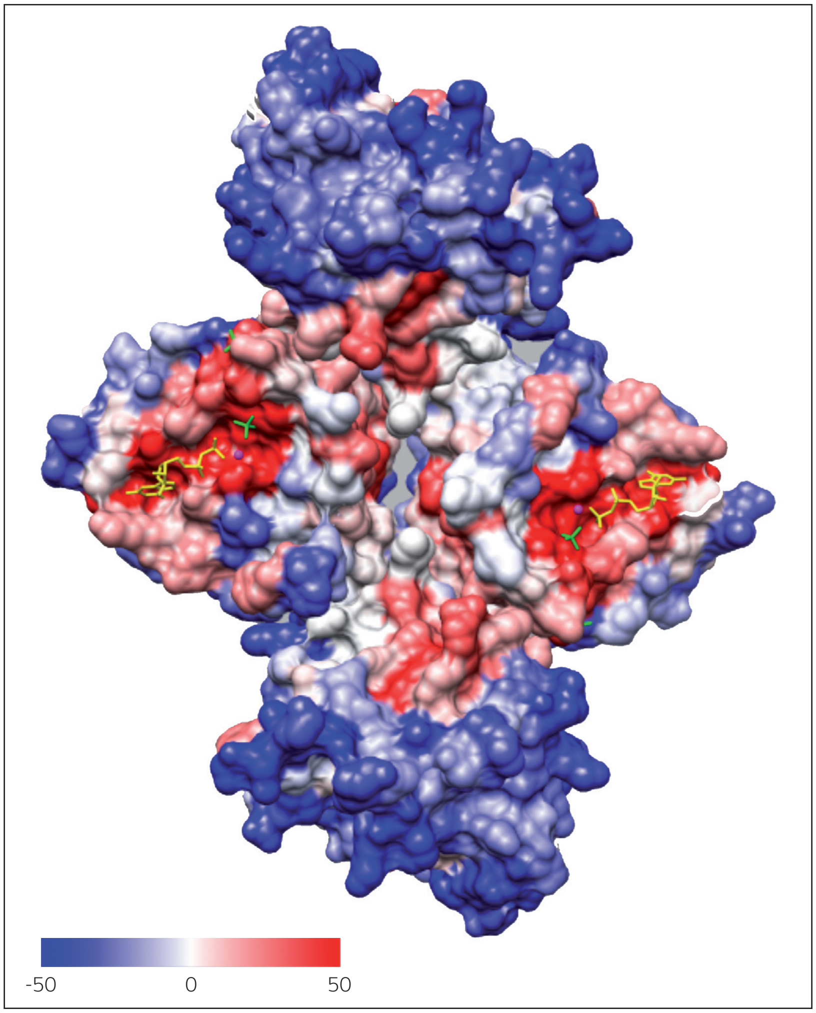
(*Adapted with permission from Bromberg Y, Yachdav G, Ofran Y, Schneider R and Rost B © 2009 Bromberg Y, Yachdav G, Ofran Y, Schneider R and Rost B*)

Improvements to computational methods have enabled the prediction of the functional effects of single amino-acid substitutions at acceptable levels of accuracy (for some methods, the overall accuracy can be significantly greater than 70%) [91-94]. The variety of existing methods aimed at this type of prediction ranges from mathematical models relying solely on evolutionary information [95], to rule-based systems [96] and machine learning methods, all of which are based on a diverse range of sequence and structural features [93,94,97].

Predictions of nsSNPs are now sufficiently accurate that the possibility of comprehensive *in silico* mutagenesis can be explored (ie, the computational prediction of functional effects for a large number of non-native mutants). One such study has shown that computational methods can fairly accurately identify functional sites annotated by alanine scans [98]. The researchers' recent results highlight the benefits of *in silico* mimicking of experimental protocols, which reduces the need for subsequent experiments in the wet laboratory environment (Figure 2) [98,99]. Obviously, this constitutes an amazing breakthrough given the extreme difference in costs of the two mutation methodologies. The experimental mutation of each residue in a protein to alanine is often prohibitive in cost, whereas the *in silico* mutation of each residue into all of the 19 non-native amino acids is relatively easy and inexpensive. The type of *in silico* sequence annotation described here has many potential applications. For example, per-residue predictions could be used in concert with biological intuition to speed up the experimental determination of active sites (Figure 2) [99].

The ability to computationally evaluate the functional effects of the mutation of any and all residues in a protein sequence, in a fraction of the time that it would take to do so experimentally, opens new perspectives for function prediction. For example, motifs of important residues that were identified by computational mutagenesis could potentially be used to correctly transfer functional annotations from one protein to another, as has been done with various other pattern-searching algorithms

Figure 2. Mutation-based predictions of Rab5 active sites.



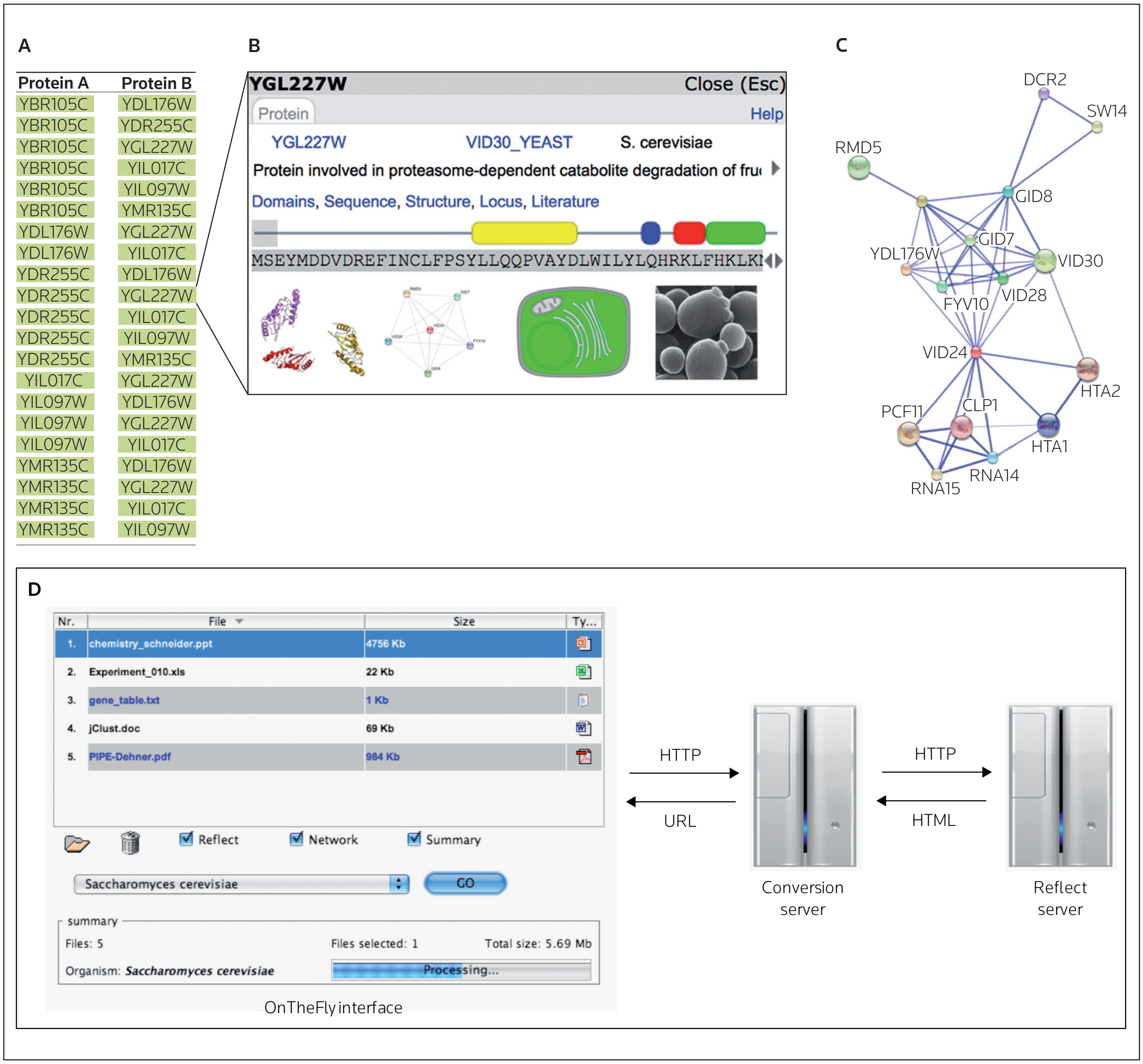
The structure of the tetramer of Ras-related protein Rab5A GTP-binding domains (PDB code: 1TU4) bound to GDP. The structure is highlighted based on SNAP predictions (from sequence information alone) of functional effects of single amino-acid substitutions. The colors indicate a range of SNAP [94] scores from 50 (red; predicted to be an active site) to -50 (blue; predicted to be a site with low or no activity). GDP is shown as a yellow wire model, the green wires are sulfate ions, and the magenta areas are cobalt ions. The clear correspondence of the predicted binding sites with the structurally likely binding grooves suggests that it is possible to predict protein active sites from sequence alone using *in silico* mutagenesis. (Molecular graphics images were produced using the UCSF Chimera package from the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco [119]).
SNAP screening for non-acceptable polymorphisms

(*Adapted with permission from Bromberg Y, Yachdav G, Ofran Y, Schneider R and Rost B © 2009 Bromberg Y, Yachdav G, Ofran Y, Schneider R and Rost B*)

(eg, reference [100]). In another case, a mutability function could be computed that would allow various predictions at each position to be weighed to determine the functional importance of each residue. Additionally, cues from life can be taken to augment the power of *in silico* mutagenesis, for example, computationally extensively evaluating the areas of a protein that surround residues associated with a particular disease phenotype. Arguably, these sequence regions are prime candidates for evaluation of their involvement in protein function.

In order for *in silico* mutagenesis to be useful in terms of defining functional sites on proteins, it must be possible to differentiate between mutations that directly disrupt function and those that affect function indirectly, by disrupting the protein scaffolds. Currently, no method is able to differentiate between such direct and indirect effects. The difference between the predictions obtained from two methods (one predicting structural changes due to a substitution and the other annotating functional changes) could conceivably be used to identify residues that directly affected protein function, and those that did so indirectly. However, in order to be complementary, both methodologies must utilize similar quantities of available information. Although several methods successfully predict the effect of single amino-acid substitutions on protein structure [92,101], no method currently does this reliably from the protein sequence alone. This severely limits the number of sequences for which predictions can be made. Another set of currently unresolved issues pertains to the dynamic aspects of how structure impacts function. Could functional hotspots be identified by simulations that reveal the energetically most expensive substitutions [102]? Significant changes in structure will almost always affect function, but are there ways to distinguish between small structural effects that change binding and those that do not? These questions will need to be addressed in the near future.

Figure 3. OnTheFly interface.



Protein A	Protein B
YBR105C	YDL176W
YBR105C	YDR255C
YBR105C	YGL227W
YBR105C	YIL017C
YBR105C	YIL097W
YBR105C	YMR135C
YDL176W	YGL227W
YDL176W	YIL017C
YDR255C	YDL176W
YDR255C	YGL227W
YDR255C	YIL017C
YDR255C	YIL097W
YDR255C	YMR135C
YIL017C	YGL227W
YIL097W	YDL176W
YIL097W	YGL227W
YIL097W	YIL017C
YMR135C	YDL176W
YMR135C	YGL227W
YMR135C	YIL017C
YMR135C	YIL097W

YGL227W Close (Esc)

Protein Help

YGL227W VID30_YEAST S. cerevisiae

Protein involved in proteasome-dependent catabolite degradation of fru

Domains, Sequence, Structure, Locus, Literature

MSEYMDDVDREFINCLFPSYLLQQPVAYDLWILYLQHRKLFHKLKI

Nr.	File	Size	Ty...
1.	chemistry_schneider.ppt	4756 Kb	
2.	Experiment_010.xls	22 Kb	
3.	gene_table.txt	1 Kb	
4.	jClust.doc	69 Kb	
5.	PIPE-Dehner.pdf	984 Kb	

☑ Reflect ☑ Network ☑ Summary

Saccharomyces cerevisiae

summary

Files: 5 Files selected: 1 Total size: 5.69 Mb

Organism: *Saccharomyces cerevisiae* Processing...

(**A**) An annotated table of protein IDs captured from a full text article in pdf format; (**B**) The popup window generated that contains information about the protein YGL227W, and (**C**) An automatically generated PPI network of associated entities for the proteins shown in (**A**). For demonstration purposes, the table was isolated from the full text article (in PDF format) and processed separately. (**D**) The architecture and functionality of the OnTheFly system. A user can drag and drop files in the OnTheFly applet. The 'GO' button sends the selected documents to the conversion server, which converts the files into HTML pages that will then be sent to the tagging server. A URL for the HTML document that was generated is returned. The organism selection drop-down list enables users to define a default species protein dictionary to be used. The 'Network' and 'Summary' options will extract the networks of associations of the recognized entities in the document(s) derived from STITCH [35], and produce a summary page listing the recognized entities.

PPI protein-protein interaction, **STITCH** search tools for interactions of chemicals

(*Adapted with permission from Bromberg Y, Yachdav G, Ofran Y, Schneider R and Rost B © 2009 Bromberg Y, Yachdav G, Ofran Y, Schneider R and Rost B*)

Marrying the web and the Deep Web

Worldwide web and Deep Web

One of the challenges for biomedical researchers is the retrieval and assimilation of information that is acquired from the various resources that are accessed on a daily basis. For example, the worldwide web is typically accessed through generalized search engines such as Google, or specialized information retrieval systems such as Google Scholar (*www.scholar.google.com*), PubMed [103], Scirus [104], novo|seek (*www.novoseek.com*), or Medstory (*www.medstory.com*). Alternatively, the 'Deep Web' (or hidden web) exists, which is likely to be several orders of magnitude larger than the worldwide web as indexed by Google. In the 'Deep Web' category there are data collections such as GenBank [105], EMBL (the European Molecular Biology Laboratory) [106], and Uniprot [107], as well as the hundreds of web servers that offer a wide range of in-depth information, prediction and analytical tools [106,108-112]. Community projects such as Proteopedia have also recently begun to reproduce one of the most impressive phenomena of the last decade, namely the advance of community-driven knowledge gathering as exemplified by Wikipedia [113].

Every aspect of biomedical research could be accelerated if it was possible to switch easily between the worldwide web and the Deep Web. The straightforward solution, which would link these extremely heterogeneous and largely incompatible systems, would be the semantic web. However, biomedical information is typically complex, and semantic annotation of the many resources would require substantial additional manual effort. As this effort would not immediately result in a benefit to any particular group of scientists, incentive is limited. It is therefore expected that the development of the semantic web will require the development not only of strong knowledge constructs, such as ontologies, but also of smaller linking mechanisms that will provide an incentive to populate and publish semantic information. One of these potential linking mechanisms could be via the use of augmented browsing.

Augmented browsing

Some methods, such as Whatizit [114] and iHop [115], systematically tag Medline abstracts that contain gene or protein names. However, there is some way to go before all scientific publishers consistently tag all of their content. An emerging approach, called augmented browsing, allows the tagging of all entities that relate to a specific field of interest and the exploration of the information by clicking on a tag. This on-demand tagging also ensures that the information delivered is up to date.

Augmented browsing tools are increasingly entering the biochemical sciences. For example, ChemGM [116], ConceptWeb (*conceptweblinker.wikiprofessional.org*) and the Conceptual Open Hypermedia Service (COHSE) [117] all tag entities such as genes, chemicals or diseases and link these tags to ontologies via popup windows, which then typically link to further data sources such as PubChem (*pubchem.ncbi.nlm.nih.gov*). Such popup windows are an effective means of displaying useful information on an entity, without the user needing to navigate away from the web page that was originally being viewed. Two methods of this type have recently been designed. Both Reflect (*reflect.ws*) and OnTheFly (*onthefly.embl.de*; Figure 3) have created systems that enable users to tag the genes, proteins and small molecules that appear in any web page, PDF, or Microsoft Office document within a few seconds. Clicking on a tag opens a popup window that contains a concise summary of the most significant information about the tagged entity, and with direct links to commonly used source data entries. This type of service has a strong focus on ease of use and ease of installation, and has been shown to be useful to general life scientists, not just to computational biologists.

In the future, the accuracy and usability of such augmented browsing services will be improved by enabling Wikipedia-like community-based, collaborative editing of the summary popup information. Such a vehicle could significantly reduce the difficulties of populating the semantic web and would allow a broad range of users to build an improved scientific web.

Conclusion

Computational biology has contributed to the successful development of each new drug that reaches the market today. Nevertheless, most contemporary computational methods still provide only indirect (if often crucial) information. Many breakthroughs have characterized the transition from there being almost no methods for predicting protein structure and function reliably ~ 15 years ago to the existence of a plethora of useful methods today. By many criteria, two very recent developments stand out in their relevance to drug discovery and development: (i) methods that predict interaction hotspots; and (ii) methods that can expedite the analysis of experimental data by realizing comprehensive *in silico* mutagenesis and accurately predict the effects of nsSNPs. These fields are young but have the potential to contribute significantly to the field, and both exemplify the amazing potential of computational biology to create results that accelerate progress at a lower cost than many experimental methods. The two methods also share another advantage in that their success currently appears to be limited only by the amount of data available (ie, they improve with every new experiment). Finally, a significant glimpse of the future was presented in this review, namely augmented browsing tools that will marry the worldwide web with the Deep Web, and will pioneer the advance of biomedical research into new scientific territory.

Acknowledgements

This paper was supported by the grant R01-LM07329-01 from the National Library of Medicine (NLM), and the grant

R01-GM079767 from the National Institute of General Medical Sciences (NIGMS) at the NIH. The authors would also like to thank all who deposit their experimental data into public databases, and all those who maintain these databases.

References

•• of outstanding interest
• of special interest

1. Liolios K, Mavromatis K, Tavernarakis N, Kyrpides NC: **The Genomes On Line Database (GOLD) in 2007: Status of genomic and metagenomic projects and their associated metadata.** *Nucleic Acids Res* (2008) **36**:D475-D479. *www.genomesonline.org*

2. Bairoch A, Bougueleret L, Altairac S, Amendolia V, Auchincloss A, Argoud-Puy G, Axelsen K, Baratin D, Blatter MC, Boeckmann B, Bolleman J *et al*: **The Universal Protein Resource (UniProt) 2009.** *Nucleic Acids Res* (2009) **37**:D169-D174.

3. Boeckmann B, Blatter MC, Famiglietti L, Hinz U, Lane L, Roechert B, Bairoch A: **Protein variety and functional diversity: Swiss-Prot annotation in its biological context.** *C R Biol* (2005) **328**(10-11):882-899.

4. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA *et al*: **Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* (2000) **25**(1):25-29.

5. Norvell JC, Berg JM: **Update on the protein structure initiative.** *Structure* (2007) **15**(12):1519-1522.

6. Barrell D, Dimmer E, Huntley RP, Binns D, O'Donovan C, Apweiler R: **The GOA database in 2009 – An integrated Gene Ontology Annotation resource.** *Nucleic Acids Res* (2009) **37**:D396-D403.

7. Dimmer E, Berardini TZ, Barrell D, Camon E: **Methods for gene ontology annotation.** *Methods Mol Biol* (2007) **406**:495-520.

8. Lee D, Redfern O, Orengo C: **Predicting protein function from sequence and structure.** *Nat Rev Mol Cell Biol* (2007) **8**(12):995-1005.

9. Rost B, Liu J, Nair R, Wrzeszczynski KO, Ofran Y: **Automatic prediction of protein function.** *Cell Mol Life Sci* (2003) **60**(12):2637-2650.

10. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank.** *Nucleic Acids Res* (2000) **28**(1):235-242.

11. Burley SK, Joachimiak A, Montelione GT, Wilson IA: **Contributions to the NIH-NIGMS Protein Structure Initiative from the PSI production centers.** *Structure* (2008) **16**(1):5-11.

12. Nair R, Liu J, Soong TT, Acton TB, Everett JK, Kouranov A, Fiser A, Godzik A, Jaroszewski L, Orengo C, Montelione GT *et al*: **Structural genomics is the largest contributor of novel structural leverage.** *J Struct Funct Genomics* (2009) **10**(2):181-191.

13. Redfern OC, Dessailly B, Orengo CA: **Exploring the structure and function paradigm.** *Curr Opin Struct Biol* (2008) **18**(3):394-402.

14. Bertonati C, Punta M, Fischer M, Yachdav G, Forouhar F, Zhou W, Kuzin AP, Seetharaman J, Abashidze M, Ramelot TA, Kennedy MA *et al*: **Structural genomics reveals EVE as a new ASCH/PUA-related domain.** *Proteins Struct Funct Bioinform* (2008) **75**(3):760-773.

15. Watson JD, Sanderson S, Ezersky A, Savchenko A, Edwards A, Orengo C, Joachimiak A, Laskowski RA, Thornton JM: **Towards fully automated structure-based function prediction in structural genomics: A case study.** *J Mol Biol* (2007) **367**(5):1511-1522.

16. Moult J, Fidelis K, Kryshtafovych A, Rost B, Hubbard T, Tramontano A: **Critical assessment of methods of protein structure prediction – Round VII.** *Proteins Struct Funct Bioinform* (2007) **69**(Suppl 8):3-9.

17. Kryshtafovych A, Fidelis K: **Protein structure prediction and model quality assessment.** *Drug Discov Today* (2009) **14**(7-8):386-393.

18. Das R, Baker D: **Macromolecular modeling with Rosetta.** *Annu Rev Biochem* (2008) **77**:363-382.

19. Zhang Y: **Template-based modeling and free modeling by I-TASSER in CASP7.** *Proteins Struct Funct Bioinform* (2007) **69**(Suppl 8):108-117.

20. Laskowski RA, Watson JD, Thornton JM: **From protein structure to biochemical function?** *J Struct Funct Genomics* (2003) **4**(2-3):167-177.

21. Goldsmith-Fischman S, Honig B: **Structural genomics: Computational methods for structure analysis.** *Protein Sci* (2003) **12**(9):1813-1821.

22. Gherardini PF, Helmer-Citterich M: **Structure-based function prediction: Approaches and applications.** *Brief Funct Genomic Proteomic* (2008) **7**(4):291-302.

23. Punta M, Ofran Y: **The rough guide to *in silico* function prediction, or how to use sequence and structure information to predict protein function.** *PLoS Comput Biol* (2008) **4**(10):e1000160.

24. Kristensen DM, Ward RM, Lisewski AM, Erdin S, Chen BY, Fofanov VY, Kimmel M, Kavraki LE, Lichtarge O: **Prediction of enzyme function based on 3D templates of evolutionarily important amino acids.** *BMC Bioinform* (2008) **9**:17.

25. Innis CA: **siteFiNDER|3D: A web-based tool for predicting the location of functional sites in proteins.** *Nucleic Acids Res* (2007) **35**:W489-W494.

26. Brylinski M, Skolnick J: **A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation.** *Proc Natl Acad Sci USA* (2008) **105**(1):129-134.

•• *Describes a novel method for the high-accuracy prediction of protein-ligand binding sites for proteins of unknown structure. The approach uses sets of the known structures of proteins that are weakly sequence-homologous to the protein of interest to model the consensus binding site, which is then mapped onto the query. Additionally, this research finds that, in many instances, the functional annotation of template proteins (GO numbers) can be accurately transferred to the query*.

27. Gabow AP, Leach SM, Baumgartner WA, Hunter LE, Goldberg DS: **Improving protein function prediction methods with integrated literature data.** *BMC Bioinform* (2008) **9**:198.

28. Ofran Y, Mysore V, Rost B: **Prediction of DNA-binding residues from sequence.** *Bioinformatics* (2007) **23**(13):i347-i353.

•• *Can we predict protein function without relying on the experimental annotation of homologous proteins? This question is critical for more than 1 million proteins with known sequences that have no experimentally annotated homolog. This study describes one of the first methods that is able to identify DNA binding sites directly from a sequence using an integrative approach without the need for an annotated homolog. This approach may constitute a first step in the journey toward de novo function prediction that is not based on homology*.

29. Wu J, Liu H, Duan X, Ding Y, Wu H, Bai Y, Sun X: **Prediction of DNA-binding residues in proteins from amino acid sequences using a random forest model with a hybrid feature.** *Bioinformatics* (2009) **25**(1):30-35.

30. Lippi M, Passerini A, Punta M, Rost B, Frasconi P: **MetalDetector: A web server for predicting metal-binding sites and disulfide bridges in proteins from sequence.** *Bioinformatics* (2008) **24**(18):2094-2095.

31. Babor M, Gerzon S, Raveh B, Sobolev V, Edelman M: **Prediction of transition metal-binding sites from apo protein structures.** *Proteins Struct Funct Bioinform* (2008) **70**(1):208-217.

32. Lobley AE, Nugent T, Orengo CA, Jones DT: **FFPred: An integrated feature-based function prediction server for vertebrate proteomes.** *Nucleic Acids Res* (2008) **36**:W297-W302.

• *Describes the publicly accessible server implementation of a method for annotating sequences of unknown function with proper GO terms. The method uses sequence-derived protein features, and therefore does not require much input information. However, GO is based on machine learning, and is not limited to annotation transfer through sequence homology. Importantly, FFPred is fast, accurate (each annotation is assigned a score representative of its reliability), and may be applied to any sequence*.

33. Hawkins T, Chitale M, Luban S, Kihara D: **PFP: Automated prediction of gene ontology functional annotations with confidence scores using protein sequence data.** *Proteins Struct Funct Bioinform* (2008) **74**(3):566-582.

34. Joshi T, Zhang C, Lin GN, Song Z, Xu D: **GeneFAS: A tool for prediction of gene function using multiple sources of data.** *Methods Mol Biol* (2008) **439**:369-386.

35. Kuhn M, von Mering C, Campillos M, Jensen LJ, Bork P: **STITCH: Interaction networks of chemicals and proteins.** *Nucleic Acids Res* (2008) **36**:D684-D688. *stitch.embl.de*

• Describes a database of interactions between proteins and small-molecule ligands. The (large) interaction data set is extracted from standardized sources of experimental data (eg, PDB) and from free text using text-mining techniques. The data are presented in a selection of descriptive views (including the reliability scores for each interaction). This ability to simply browse through the visual representations of extracted interactions significantly contributes to the overall understanding of involved pathways.

36. Radivojac P, Peng K, Clark WT, Peters BJ, Mohan A, Boyle SM, Mooney SD: **An integrated approach to inferring gene-disease associations in humans.** *Proteins Struct Funct Bioinform* (2008) **72**(3):1030-1037.

37. Muilu J, Peltonen L, Litton JE: **The federated database – A basis for biobank-based post-genome studies, integrating phenome and genome data from 600,000 twin pairs in Europe.** *Eur J Hum Genet* (2007) **15**(7):718-723.

38. Aerts S, Lambrechts D, Maity S, Van Loo P, Coessens B, De Smet F, Tranchevent LC, De Moor B, Marynen P, Hassan B, Carmeliet P *et al*: **Gene prioritization through genomic data fusion.** *Nat Biotechnol* (2006) **24**(5):537-544.

39. Segal E, Friedman N, Kaminski N, Regev A, Koller D: **From signatures to models: Understanding cancer using microarrays.** *Nat Genet* (2005) **37**:S38-S45.

40. Boutte CC, Srinivasan BS, Flannick JA, Novak AF, Martens AT, Batzoglou S, Viollier PH, Crosson S: **Genetic and computational identification of a conserved bacterial metabolic module.** *PLoS Genet* (2008) **4**(12):e1000310.

41. Brilli M, Fani R, Lio P: **Current trends in the bioinformatic sequence analysis of metabolic pathways in prokaryotes.** *Brief Bioinform* (2008) **9**(1):34-45.

42. Raes J, Bork P: **Molecular eco-systems biology: Towards an understanding of community function.** *Nat Rev Microbiol* (2008) **6**(9):693-699.

43. Laskowski RA, Thornton JM: **Understanding the molecular machinery of genetics through 3D structures.** *Nat Rev Genet* (2008) **9**(2):141-151.

44. Tosatto SC, Toppo S: **Large-scale prediction of protein structure and function from sequence.** *Curr Pharm Des* (2006) **12**(17):2067-2086.

45. Shulman-Peleg A, Shatsky M, Nussinov R, Wolfson HJ: **MultiBind and MAPPIS: Webservers for multiple alignment of protein 3D-binding sites and their interactions.** *Nucleic Acids Res* (2008) **36**:W260-W264.

46. Neuvirth H, Raz R, Schreiber G: **ProMate: A structure based prediction program to identify the location of protein-protein binding sites.** *J Mol Biol* (2004) **338**(1):181-199.

•• Introduces a tool for the accurate prediction of interface residues based on the 3D structure of an unbound chain. The method was trained taking into account the complex structure of interfaces and accounts for cavities that contribute to the stability of the complexes.

47. Najmanovich R, Kurbatova N, Thornton J: **Detection of 3D atomic similarities and their use in the discrimination of small molecule protein-binding sites.** *Bioinformatics* (2008) **24**(16):i105-i111.

48. de Vries SJ, Bonvin AM: **How proteins get in touch: Interface prediction in the study of biomolecular complexes.** *Curr Protein Peptide Sci* (2008) **9**(4):394-406.

•• Provides a comprehensive analysis of the challenges of protein interaction site prediction that also includes a survey of existing methods and a comparison of their performance.

49. Fischer JD, Mayer CE, Söding J: **Prediction of protein functional residues from sequence by probability density estimation.** *Bioinformatics* (2008) **24**(5):613-620.

• Reports on an interesting approach to improving the prediction of functional sites in a protein. Instead of using conservation alone, it relies on identifying positions in a sequence that can, because of their localization or other features, actually be involved in protein-ligand interactions. Although the reported accuracy and coverage can still be improved upon, the method performs significantly better than many existing approaches.

50. Chien TY, Chang DT, Chen CY, Weng YZ, Hsu CM: **E1DS: Catalytic site prediction based on 1D signatures of concurrent conservation.** *Nucleic Acids Res* (2008) **36**:W291-W296.

• Describes a novel approach to identifying 3D enzyme catalytic sites from sequence motifs. The resulting database of motifs can be scanned through for identification of functional sites in newly discovered proteins.

51. Komurov K, White M: **Revealing static and dynamic modular architecture of the eukaryotic protein interaction network.** *Mol Syst Biol* (2007) **3**:110.

52. Humphris EL, Kortemme T: **Prediction of protein-protein interface sequence diversity using flexible backbone computational protein design.** *Structure* (2008) **16**(12): 1777-1788.

53. Huang B, Schroeder M: **Using protein binding site prediction to improve protein docking.** *Gene* (2008) **422**(1-2):14-21.

54. Laskowski RA, Luscombe NM, Swindells MB, Thornton JM: **Protein clefts in molecular recognition and function.** *Protein Sci* (1996) **5**(12):2438-2452.

55. Pettit FK, Bare E, Tsai A, Bowie JU: **HotPatch: A statistical approach to finding biologically relevant features on protein surfaces.** *J Mol Biol* (2007) **369**(3):863-879.

56. Bahadur RP, Chakrabarti P, Rodier F, Janin J: **A dissection of specific and non-specific protein-protein interfaces.** *J Mol Biol* (2004) **336**(4):943-955.

57. Jones S, Thornton JM: **Principles of protein-protein interactions.** *Proc Natl Acad Sci USA* (1996) **93**(1):13-20.

58. Fernandez-Recio J, Totrov M, Abagyan R: **Identification of protein-protein interaction sites from docking energy landscapes.** *J Mol Biol* (2004) **335**(3):843-865.

59. Ritchie DW: **Recent progress and future directions in protein-protein docking.** *Curr Protein Peptide Sci* (2008) **9**(1):1-15.

60. Ofran Y, Rost B: **Analysing six types of protein-protein interfaces.** *J Mol Biol* (2003) **325**(2):377-387.

61. Kini RM, Evans HJ: **Prediction of potential protein-protein interaction sites from amino acid sequence. Identification of a fibrin polymerization site.** *FEBS Lett* (1996) **385**(1-2):81-86.

62. Jones S, Thornton JM: **Prediction of protein-protein interaction sites using patch analysis.** *J Mol Biol* (1997) **272**(1):133-143.

•• Reports one of the first methods for structure-based interaction site prediction, and introduces some of the fundamental insights in the field.

63. Bradford JR, Needham CJ, Bulpitt AJ, Westhead DR: **Insights into protein-protein interfaces using a Bayesian network prediction method.** *J Mol Biol* (2006) **362**(2):365-386.

64. Armon A, Graur D, Ben-Tal N: **ConSurf: An algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information.** *J Mol Biol* (2001) **307**(1):447-463.

•• Reports on a versatile tool for the identification of functional residues based on evolutionary conservation and structural analysis.

65. Ofran Y, Rost B: **ISIS: Interaction Sites Identified from Sequence.** *Bioinformatics* (2007) **23**(2):e13-e16.

•• Introduces the first integrative tool for sequence-based protein interaction site prediction. The tool incorporates physicochemical factors, predicted structural features and evolutionary conservation.

66. Res I, Mihalek I, Lichtarge O: **An evolution based classifier for prediction of protein interfaces without using protein structures.** *Bioinformatics* (2005) **21**(10):2496-2501.

67. Kufareva I, Budagyan L, Raush E, Totrov M, Abagyan R: **PIER: Protein interface recognition for structural proteomics.** *Proteins Struct Funct Bioinform* (2007) **67**(2):400-417.

68. Darnell SJ, LeGault L, Mitchell JC: **KFC Server: Interactive forecasting of protein interaction hot spots.** *Nucleic Acids Res* (2008) **36**:W265-W269.

69. Murga LF, Ondrechen MJ, Ringe D: **Prediction of interaction sites from apo 3D structures when the holo conformation is different.** *Proteins Struct Funct Bioinform* (2008) **72**(3):980-992.

70. Guo Y, Yu L, Wen Z, Li M: **Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences.** *Nucleic Acids Res* (2008) **36**(9):3025-3030.

71. Keskin O, Tuncbag N, Gursoy A: **Characterization and prediction of protein interfaces to infer protein-protein interaction networks.** *Curr Pharm Biotechnol* (2008) **9**(2):67-76.

72. Pazos F, Valencia A: ***In silico*** **two-hybrid system for the selection of physically interacting protein pairs.** *Proteins Struct Funct Bioinform* (2002) **47**(2):219-227.

73. Ofran Y, Punta M, Schneider R, Rost B: **Beyond annotation transfer by homology: Novel protein-function prediction methods to assist drug discovery.** *Drug Discov Today* (2005) **10**(21): 1475-1482.

74. Wells JA, McClendon CL: **Reaching for high-hanging fruit in drug discovery at protein-protein interfaces.** *Nature* (2007) **450**(7172):1001-1009.

75. Vassilev LT: **Small-molecule antagonists of p53-MDM2 binding: Research tools and potential therapeutics.** *Cell Cycle* (2004) **3**(4):419-421.

76. Koblish HK, Zhao S, Franks CF, Donatelli RR, Tominovich RM, LaFrance LV, Leonard KA, Gushue JM, Parks DJ, Calvo RR, Milkiewicz KL *et al*: **Benzodiazepinedione inhibitors of the Hdm2:p53 complex suppress human tumor cell proliferation *in vitro* and sensitize tumors to doxorubicin *in vivo*.** *Mol Cancer Ther* (2006) **5**(1):160-169.

77. Rudolph J: **Inhibiting transient protein-protein interactions: Lessons from the Cdc25 protein tyrosine phosphatases.** *Nat Rev Cancer* (2007) **7**(3):202-211.

78. Bogan AA, Thorn KS: **Anatomy of hot spots in protein interfaces.** *J Mol Biol* (1998) **280**(1):1-9.

79. Wells JA: **Systematic mutational analyses of protein-protein interfaces.** *Methods Enzymol* (1991) **202**:390-411.

80. Ofran Y, Rost B: **Protein-protein interaction hotspots carved into sequences.** *PLoS Comput Biol* (2007) **3**(7):e119.

81. Thorn KS, Bogan AA: **ASEdb: A database of alanine mutations and their effects on the free energy of binding in protein interactions.** *Bioinformatics* (2001) **17**(3):284-285.

82. Guney E, Tuncbag N, Keskin O, Gursoy A: **HotSprint: Database of computational hot spots in protein interfaces.** *Nucleic Acids Res* (2008) **36**:D662-D666.

83. Buckle AM, Schreiber G, Fersht AR: **Protein-protein recognition: Crystal structural analysis of a barnase-barstar complex at 2.0-A resolution.** *Biochemistry* (1994) **33**(30):8878-8889.

84. Boas FE, Harbury PB: **Design of protein-ligand binding based on the molecular-mechanics energy model.** *J Mol Biol* (2008) **380**(2):415-424.

85. Huszar D, Lynch CA, Fairchild-Huntress V, Dunmore JH, Fang Q, Berkemeier LR, Gu W, Kesterson RA, Boston BA, Cone RD, Smith FJ *et al*: **Targeted disruption of the melanocortin-4 receptor results in obesity in mice.** *Cell* (1997) **88**(1):131-141.

86. Yang Y, Chen M, Lai Y, Gantz I, Georgeson KE, Harmon CM: **Molecular determinants of human melanocortin-4 receptor responsible for antagonist SHU9119 selective activity.** *J Biol Chem* (2002) **277**(23):20328-20335.

87. Markiewicz P, Kleina LG, Cruz C, Ehret S, Miller JH: **Genetic studies of the lac repressor. XIV. Analysis of 4000 altered *Escherichia coli* lac repressors reveals essential and non-essential residues, as well as 'spacers' which do not require a specific sequence.** *J Mol Biol* (1994) **240**(5):421-433.

88. Rennell D, Bouvier SE, Hardy LW, Poteete AR: **Systematic mutation of bacteriophage T4 lysozyme.** *J Mol Biol* (1991) **222**(1):67-88.

89. Gondo Y: **Trends in large-scale mouse mutagenesis: From genetics to functional genomics.** *Nat Rev Genet* (2008) **9**(10):803-810.

90. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, Hirschhorn JN: **Genome-wide association studies for complex traits: Consensus, uncertainty and challenges.** *Nat Rev Genet* (2008) **9**(5):356-369.

91. Ng PC, Henikoff S: **Predicting the effects of amino acid substitutions on protein function.** *Annu Rev Genomics Hum Genet* (2006) **7**:61-80.

92. Cheng TM, Lu YE, Vendruscolo M, Lio' P, Blundell TL: **Prediction by graph theoretic measures of structural effects in proteins arising from non-synonymous single nucleotide polymorphisms.** *PLoS Comput Biol* (2008) **4**(7):e1000135.

•• *Describes a novel method for annotating structurally disruptive nsSNPs, and compares its predictions of structural disruptions to predictions of functional disruption from other methods and to experimental disease-association studies. The prediction method described is interesting and relatively accurate, but it is not particularly informative in its output and is only applicable to proteins of known structure. More importantly, this study reports that the method performs as well in annotating disease-associated mutations as do other methods designed specifically for that purpose. This finding suggests that many of the disease-associated nsSNPs are structurally disruptive*.

93. Barenboim M, Masso M, Vaisman, II, Jamison DC: **Statistical geometry based prediction of nonsynonymous SNP functional effects using random forest and neuro-fuzzy classifiers.** *Proteins Struct Funct Bioinform* (2008) **71**(4):1930-1939.

94. Bromberg Y, Yachdav G, Rost B: **SNAP predicts effect of mutations on protein function.** *Bioinformatics* (2008) **24**(20):2397-2398.

•• *Describes a publicly accessible implementation of the SNAP server – a neural-network-based method aimed at predicting functional effects of single amino-acid substitutions from sequence. Adding to the accuracy of SNAP predictions is the ability of the server to generate in silico mutagenesis results for selected residues*.

95. Lee TC, Lee AS, Li KB: **Incorporating the amino acid properties to predict the significance of missense mutations.** *Amino Acids* (2008) **35**(3):615-626.

96. Ramensky V, Bork P, Sunyaev S: **Human non-synonymous SNPs: Server and survey.** *Nucleic Acids Res* (2002) **30**(17):3894-3900.

97. Ju W, Shan J, Yan C, Cheng HD: **Discrimination of disease-related non-synonymous single nucleotide polymorphisms using multi-scale RBF kernel fuzzy support vector machine.** *Pattern Recog Letts* (2009) **30**(4):391-396.

98. Bromberg Y, Rost B: **Comprehensive *in silico* mutagenesis highlights functionally important residues in proteins.** *Bioinformatics* (2008) **24**(16):i207-i212.

•• *Describes the use of computational analysis of single amino-acid substitutions to mimic experiments aimed at identifying binding hotspots, and also suggests a novel approach of using the in silico facility with large-scale mutagenesis to highlight functionally important residues. The paper highlights an important idea in function prediction: while it cannot be expected that experimental mutagenesis studies directly translate into computational studies, the latter can be successfully applied to identifying functionally important sites*.

99. Bromberg Y, Overton J, Vaisse C, Leibel RL, Rost B: ***In silico*** **mutagenesis: A case study of the melanocortin 4 receptor.** *FASEB J* (2009): *In Press*.

•• *Describes a novel approach to identifying functionally important sites in proteins using in silico mutagenesis. This is the first paper of its kind, and addresses the feasibility of in silico mutagenesis studies in detail. The suggested approach is evaluated in application to a paralog and ortholog of the human melanocortin-4 receptor. The method accurately identifies sequence residues that are important to the function of each of the described proteins, and, possibly even more importantly, accurately distinguishes between residues that are responsible for the molecules' differing functions.*

100. Shu N, Zhou T, Hovmoller S: **Prediction of zinc-binding sites in proteins from sequence.** *Bioinformatics* (2008) **24**(6):775-782.

101. Zoete V, Meuwly M: **Importance of individual side chains for the stability of a protein fold: Computational alanine scanning of the insulin monomer.** *J Comput Chem* (2006) **27**(15): 1843-1857.

102. Guerois R, Nielsen JE, Serrano L: **Predicting changes in the stability of proteins and protein complexes: A study of more than 1000 mutations.** *J Mol Biol* (2002) **320**(2):369-387.

103. Sayers EW, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S, Feolo M *et al*: **Database resources of the National Center for Biotechnology Information.** *Nucleic Acids Res* (2009) **37**:D5-D15.

104. Seeber F: **Patent searches as a complement to literature searches in the life sciences – A 'how-to' tutorial.** *Nat Protoc* (2007) **2**(10):2418-2428.

105. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW: **GenBank.** *Nucleic Acids Res* (2009) **37**:D26-D31.

106. Galperin MY, Cochrane GR: **Nucleic Acids Research annual database issue and the NAR online molecular biology database collection in 2009.** *Nucleic Acids Res* (2009) **37**:D1-D4.

107. Mulder NJ, Kersey P, Pruess M, Apweiler R: ***In silico* characterization of proteins: UniProt, InterPro and Integr8.** *Mol Biotechnol* (2008) **38**(2):165-177.

108. Rost B, Yachdav G, Liu J: **The PredictProtein server.** *Nucleic Acids Res* (2004) **32**:W321-W326.

109. Altman RB, Bergman CM, Blake J, Blaschke C, Cohen A, Gannon F, Grivell L, Hahn U, Hersh W, Hirschman L, Jensen LJ *et al*: **Text mining for biology – The way forward: Opinions from leading scientists.** *Genome Biol* (2008) **9**(Suppl 2):S7.

110. Nair R, Rost B: **Predicting protein subcellular localization using intelligent systems.** *Methods Mol Biol* (2008) **484**:435-463.

111. Bigelow H, Rost B: **Online tools for predicting integral membrane proteins.** *Methods Mol Biol* (2009) **528**:3-23.

112. Pruitt KD, Tatusova T, Klimke W, Maglott DR: **NCBI Reference Sequences: Current status, policy and new initiatives.** *Nucleic Acids Res* (2009) **37**:D32-D36.

113. Hodis E, Prilusky J, Martz E, Silman I, Moult J, Sussman JL: **Proteopedia – A scientific 'wiki' bridging the rift between three-dimensional structure and function of biomacromolecules.** *Genome Biol* (2008) **9**(8):R121.

•• *Describes one of the most impressive community efforts to facilitate the understanding of protein structure (and its links to function). Proteopedia is a wiki-like collection of user-generated content including various views of the same structures tagged with comprehensive text descriptions*.

114. Rebholz-Schuhmann D, Arregui M, Gaudan S, Kirsch H, Jimeno A: **Text processing through Web services: Calling Whatizit.** *Bioinformatics* (2008) **24**(2):296-298.

115. Fernandez JM, Hoffmann R, Valencia A: **iHOP web services.** *Nucleic Acids Res* (2007) **35**(Web Server issue):W21-W26.

116. Willighagen EL, O'Boyle NM, Gopalakrishnan H, Jiao D, Guha R, Steinbeck C, Wild DJ: **Userscripts for the life sciences.** *BMC Bioinformatics* (2007) **8**:487.

117. Bechhofer SK, Stevens RD, Lord PW: **Ontology driven dynamic linking of biology resources.** *Pac Symp Biocomput* (2005):79-90.

118. Zhu G, Zhai P, Liu J, Terzyan S, Li G, Zhang XC: **Structural basis of Rab5-Rabaptin5 interaction in endocytosis.** *Nat Struct Mol Biol* (2004) **11**(10):975-983.

119. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE: **UCSF Chimera – A visualization system for exploratory research and analysis.** *J Comput Chem* (2004) **25**(13):1605-1612.