A (not so) Quick Introduction to Protein Function Prediction

Predrag Radivojac, Indiana University

September 27, 2013

This document is primarily intended for computer scientists as a rapid introduction to the area of protein function prediction. I will try to formally introduce the protein function prediction problem and comment on why it is important and challenging. I will also discuss the Critical Assessment of Functional Annotation (CAFA), an experiment dedicated to evaluating computational tools for protein function prediction [30]. Throughout, I will provide background information and pointers on where to find data and what to pay attention to when developing new methods. My hope is that this document will help to increase the participation of computater scientists in this important computational biology problem.

Before we start, let me mention a couple of more things: (*i*) an old but still relevant introduction to molecular biology for computer scientists was written by Larry Hunter [17], and I recommend reading it; (*ii*) a good set of review papers can provide additional perspectives on protein function prediction [5, 35, 16, 40, 29, 31]; and (*iii*) if you have any questions or comments about this document, feel free to contact me. Feedback is always welcome!

1 Proteins

Proteins are biological macromolecules responsible for a wide range of activities in our cells, tissues, organs, and bodies. They constitute more than 50% of the dry weight of cells and play a central role in the structure and function of cells. Examples of important biological roles a protein may have include catalytic activity (e.g. trypsin), muscle contraction (myosin), structural support (keratins), antibacterial and antiviral defense (immunoglobulins), signaling and regulation (Fos/Jun families through DNA transcription), helping other proteins fold (e.g. the chaperone activity of GroEL in bacteria), or storage (e.g. ferritin for iron storage and release).

Proteins carry out their function in the context of environment they are in. This environment includes other macromolecules such as proteins, DNA, or RNA. Some of these molecules come from viruses that at the very least take up resources of the cellular machinery. Others, on the other hand, come from

(potentially pathogenic) microorganisms; for example, through secretion. The environment also includes small chemical compounds from food, water, or air, as well as factors such as temperature or pH level. Modulation of protein function upon many of these factors has been studied for various molecules; see [45] for a recent example.

1.1 Protein structure and dynamics

Proteins have a relatively simple chemical organization. They are linear chains (polymers) of amino acids connected by covalent (peptide) bonds [6]. Amino acids are small molecules with a common backbone (consisting of several C, H, O, and N atoms) and a side chain with up to 30 more atoms (C, H, O, N, and S). When a peptide bond is formed between two amino acids, a molecule of water is released (two hydrogen atoms and one oxigen), which is why we frequently refer to proteins as chains of *amino acid residues*.

Short polypeptide chains are called peptides, while the long ones are usually called proteins.[1] In either case, these linear chains of amino acid residues explore a large space of possible 3D conformations (shapes) and can in theory be found in any (kinetically) accessible conformation at a particular time. However, points in the protein's conformational space are not equiprobable.

I think an easy and good way to think about a protein's structure is through its probability distribution over the possible 3D conformations (biologists and physicists would probably prefer the "energy landscape" terminology). Those proteins with a single dominant (narrow) peak can be "solved" using X-ray crystallography or NMR spectroscopy and then visualized as static molecules (Figure 1). Such proteins can be found in the Protein Data Bank (PDB) [4]. Interestingly, highly dissimilar protein sequences can have highly similar dominant 3D structures (folds) [34]. In fact, there is a relatively small number of folds a protein may have. You can find more about structural classification of proteins from the SCOP [22] and CATH [25] databases.

Of course, proteins are not rigid bodies, their 3D structure changes in time. Sometimes these are only minor fluctuations around a set of time-invariant atom positions which themselves can be obtained through averaging. On the other hand, the sequences and physicochemical properties of other proteins allow large and/or irregular conformational changes. Their probability distribution over the space of conformations in physiological conditions does not have a single dominant peak. These proteins are referred to as *intrinsically disordered* proteins [12]. A large number of proteins contain both ordered and disordered regions. If a consecutive set of residues is missing in a protein's PDB structure, we typically consider such regions to be disordered, but there are several other techniques

[1] There is no clear distinction between peptides and proteins. Usually, a chain of several amino acid residues is referred to as peptide. Longer polypeptides that carry out function are usually refered to as proteins. Insulin (57 residues) and ubiquiin (76 residues) are probably the shortest human proteins; titin (34,350 residues) is the longest. On the other hand, many peptides are functional and in size close to insulin. Neuropeptides are good examples of functionally important peptides.

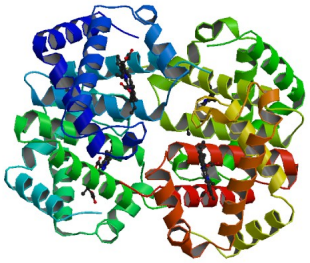

Figure 1: Structure of human hemoglobin; PDB ID 1AN3. Hemoglobin is a tetramer containing two alpha and two beta subunits (chains). Hemoglobin is involved in oxygen binding and transport. It carries oxygen from lungs to everywhere else in our bodies via red blood cells. Mutations in the sequence of this protein are associated with sickle cell disease.

for broadly characterizing a protein's conformational dynamics. Disordered proteins are listed in the DisProt database [43]. They can become ordered upon binding to other molecules. The studies of disordered proteins in the past decade have shown that the dynamics of a protein's structure plays an important role in its biological activity [41]. In fact, I think that what we used to call a *protein folding problem* to find the dominant conformation of an ordered protein, has become an incomplete and maybe even old-fashioned way of looking at things (it is still very important!).

In many situations multiple chains come together to form a complex that may be the only state in which the constitutive amino acid chains are biologically active. One such molecule, hemoglobin, is shown in Figure 1.

1.2 Ontological annotation of proteins

Many biologists and biomedical scientists are involved in elucidating what proteins do. Hundreds of thousands of papers that describe the activity of proteins in different situations have been published. However, to facilitate knowledge exchange and use of computers to process such massive data, natural language descriptions of protein function are usually ontologized. That is, a protein's activity is described using hierarchical knowledge representations such as trees or directed acyclic graphs [3]. In all such representations, nodes or terms describe particular activities, and links represent relationships between terms (e.g. is-a, part-of). Ontolgies are usually developed manually from deep understanding of the domain [32], but more recently there have been attempts to use data directly to construct ontologies or improve the existing ones [10, 1, 13].

There are a few biomedical ontologies describing protein function [23, 3, 36]. Here we are mostly interested in those representations developed by the Gene Ontology (GO) Consortium. The Gene Ontology [3] actually consists of three in-

dependent directed acyclic graphs named Molecular Function Ontology (MFO), Biological Process Ontology (BPO), and Cellular Component Ontology (CCO). These ontologies are species independent and were originally developed with an intent to unify biology via common knowledge representation across all species. The Gene Ontology, however, does not address organismal phenotypes, that is, sets of physical and behavioral characteristis of an individual. Most ontologies involving higher levels of functional descriptors (phenotypes, diseases) are species specific, although some that are being worked on will consider large groups of species (e.g. micobes). Owing to the biomedical importance, human-centric ontologies have reached sufficient sophistication to be used for the phenotypic annotation of proteins. The Disease Ontology (DO) [39] and the Human Phenotype Ontology (HPO) [33] are good albeit constantly evolving examples.

Sometimes there is confusion when different researchers refer to the term *protein function*. Those closer to the biochemical side are typically interested in the MFO aspects of function: e.g. whether a protein is an enzyme, and if so, what kind of enzyme. For them, it is also important to understand structural and mechanistic aspects of a particular enzymatic reaction. On the other hand, researchers in the functional genomics area usually use term function to express the BPO level of activity. They are typically interested in pathways in which a protein participates and how high level cellular processes are carried out or maintained. The BPO level of functional annotation is more abstract than MFO in the sense that it describes a range of biochemical events that take place to result in a particular biolgical outcome: e.g. cell-cell signaling, locomotion, or cell death. Two proteins can have the same enzymatic activity, but because they interact with different partners, they may be involved in completely different biological processes. Furthermore, if two proteins interact, it is highly likely they are part of the same biological process; but it means little in terms of the molecular function of these molecules. Thus, when protein-protein interaction data is used to make inferences about protein function, it is important to specify what aspect of function is considered. Phenotypic characterization is even fuzzier because it describes how the changes in protein sequence or expression impact physical and behavioral characteristics of the entire organism. It is difficult to pin down what *protein function* really means; basically, it is "everything that happens to or through a protein" during its lifetime [35]. Different ontologies describe different aspects of a protein's activity.

In addition to experimentalists, a number of biocurators are also involved in annotation of protein function through interpretation of experiments from the literature and summarizing these experiments using ontolgical representation. Interestingly, there was an attempt to involve authors of the original papers to annotate their proteins directly as part of the publication process [20]. It turned out that the authors did not perform as well as biocurators, mostly because of unfamiliarity with the full range and detail of the ontologies and lack of time to learn it. Therefore, we need biocurators as much as we need experimentalists and computer scientists.[2]

[2]This is a perfect place to mention Alex Bateman's characterization, form a recent talk,

1.3 Data representation

Because there are only 20 amino acids that constitute proteins (there are some exceptions here), it is convenient to represent proteins as strings using an alphabet of 20 symbols; $\Sigma = \{$A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y$\}$. This makes proteins, similar to other biological macromolecules (DNA and RNA), ideal for the development and application of string algorithms and data structures from the computer science community. Sequence alignment, for example, is the conrnerstone of computational biology and also key to protein function prediction (one might be tempted to say that sequence alignment is a byproduct of the calulcation of edit distance between strings). Different variants involve pairwise sequence alignment, multiple sequence alignment (a known NP-complete problem [44]), alignment of a sequence or a sequence family against a database of sequences or sequence families, or genome alignment, each of which requires specific solutions.

Another powerful way to represent proteins is through graphs or networks. A node in a protein-protein interaction (PPI) network corresponds to a protein and each link represents an event that the two proteins physically bind (or interact) to carry out a biologically meaningful function. Similarly, gene regulatory networks (usually directed graphs) can be formed based on the expression patters between of the paticular groups of genes. It is important to mention that there exists a time component or a particular order in which various binding or regulatory events occur; however, current data rarely provide sufficient resolution for dynamic network modeling [28]. A single protein chain or a complex with a given structure can also be seen as a graph with amino acids (or atoms) being vertices and edges drawn based on the proximity between pairs of amino acids. A geometric representations of proteins (using the 3D coordinates of each atom in its structure) can be useful for the application of techniques from computer vision; e.g. geometric hashing [24].

We can also look at proteins as time series data. An example could be a hydropathy profile [19] or a measurement of gene expression data at several time points. Methods that can integrate a variety of evidence in principled ways are more likely to achieve good performance. However, the knowledge of biology and technology used to interrogate biological systems should not be underestimated.

Perhaps the previous discussion can be interpreted in a way that protein function prediction is simple because we only need to apply a few string, graph, or time-series algorithms and integrate them to perform statistical inference. It is actually more complex than that. A single gene sometimes can encode multiple proteins through events such as *alternative splicing* or *post-translational processing* of proteins. In alternative splicing different exons in a genomic sequence are stitched together differently at the mRNA level to produce different protein forms (called *isoforms*, perhaps incorrectly). Similarly, a protein can be processed upon translation. These events include chopping proteins into functional fragments and/or chemically modifying various amino acid residues,

that "biocuration is *bedrock* of computational function prediction". He insisted that *foundation* was too weak a term.

Database	Link	Comment
UniProtKB	uniprot.org	Sequence, functional annotation
Pfam	pfam.sanger.ac.uk	Sequence families
PDB	pdb.org	Structures
ModBase	modbase.compbio.ucsf.edu	Predicted structures
I2D	ophid.utoronto.ca	Protein-protein interactions
GEO	ncbi.nlm.nih.gov/geo	Gene expression data
PRIDE	www.ebi.ac.uk/pride	Mass-spectrometry data

Table 1: Summary of main biological databases that can be used for function prediction. A number of other databases can also be useful. For structural classification see SCOP and CATH; for human mutation data see HGMD, UniProtKB, ClinVar, and OMIM; for human disease data see OMIM and HPO; for pharmacogenetics data see PharmGKB; for chemical compounds see ChEBI and PubChem, KEGG for pathways, etc. Protein function data can also be found in UniProt-GOA, BioCyc, and GO databases.

thereby changing their physicochemical properties and, as a consequence, the function of the protien. There are more than 400 post-translational modification events such as the attachment of a phosphate group to serine (S), threonine (T), tyrosine (Y), or histidine (H) residues in a protein (post-translational modifications effectively increase the amino acid alphabet size but can also occur in a particular order). When you take a combinatorial look at all post-transcriptional and post-translational events, they may produce a vast set of protein forms that are all encoded by the same gene and have exactly the same canonical amino acid sequence. This is one of the big challenges of protein function prediction.

1.4 Where is protein data?

Protein data is stored in a large collection of biomedical databases. Each year the journal *Nucleic Acids Research* has a special issue devoted to presenting new databases and updates of the existing databases. Other journals also accept submissions presenting new biological databases. However, most of these databases are too specialized and not relevant to this summary. A few that I think are important are highlighted in Table 1.

2 Protein Function Prediction Problem

Consider now a labeled data set $\mathcal{D}_\ell = \{(p_1, T_1), (p_2, T_2), \ldots\}$, where p_i is the i-th protein and T_i is its (potentially incomplete) experimentally-validated functional annotation from some ontology $\mathcal{O}$. Each protein p_i is represented by its sequence s_i, but also other information such as the species it came from, its 3D structure, its interactions with other molecules, etc. The species information is important as two identical proteins from different species may still be character-

ized with different functions depending upon the cellular context in that species. The functional annotation T_i corresponding to p_i is required to be a *consistent subgraph* of $\mathcal{O}$. This means that if a term (vertex) $v \in T_i$ is associated with p_i, then all ancestors of v must also be in T_i up to the root(s) of the ontology. In a slight abuse of notation, we use T_i to describe both a subgraph of $\mathcal{O}$ and the set of terms in this subgraph (a graph is defined as a set of vertices and a set of edges between pairs of these vertices). Informally, we will write $T_i \subseteq \mathcal{O}$ to describe consistent subgraphs of $\mathcal{O}$. Finally, let $\mathcal{D}$ be all other data that can be exploited in function prediction. For example, electronic medical record data can be exploited to understand disease co-morbidity and utilized in prediction of DO or HPO annotations, even though such data is not necessarily associated with any biological macromolecules.

We can now formulate the *protein function prediction (PFP) problem* as follows: given data set of labeled proteins $\mathcal{D}_\ell$, all additional data $\mathcal{D}$, and a previously unseen protein p, find a consistent subgraph $\hat{T} \subseteq \mathcal{O}$ that is most likely to be this protein's experimeal (but unknown) annotation T. In other words, we seek to predict the true annotation T as

$$\hat{T} = \arg\max_{P \subseteq \mathcal{O}} \left\{ \Pr(P|p) \right\},$$

where $\Pr(P|p)$ is the posterior distribution over all consistent subgraphs P in the ontology (note that owing to the fact that the number of consistent subgraph in the ontology is finite, $\Pr(P|p)$ is a probability mass function). In reality, computational models cannot predict T exactly. Therefore, we usually require them to output a score associated with each term or consistent subgraph in the ontology. The expectation is that a good decision threshold can be applied to balance precision and recall (see later for a formal definition) according to the needs of a particular user. An illustration of the protein function prediction problem is shown in Figure 2.

We shall distinguish this problem formulation from the one that we will refer to as the *cadidate gene prioritization (CGP) problem.* The gene prioritization problem emerged as a means of computationally guiding biological experiments to identify genes involved in disease [21]. In this framework, we have a labeled data set $\mathcal{D}_v = \{(g_1, T_1), (g_2, T_2), \ldots\}$, where g_i is the i-th gene and $v \in T_i$; that is each protein in associated with a particular *term* in the ontology $\mathcal{O}$ (we use g instead of p only to distinguish between the two problem formulations). We also have a set of genes $\mathcal{U}_{\bar{v}} = \{g_i\}_{i=1}^m$ that are not known to be associated with v and, as before, some additional data $\mathcal{D}$. The goal of the candidate gene prioritization algorithm is to determine which gene from $\mathcal{U}_{\bar{v}}$ is the most likely to be associated with term v and, therefore, should be the next in line to be experimentally studied. We seek to determine the best candidate gene as

$$\hat{g} = \arg\max_{g \in \mathcal{U}_{\bar{v}}} \left\{ \Pr(v|g) \right\},$$

where $\Pr(v|g)$ is the posterior probability that gene g is associated with term $v \in \mathcal{O}$. As we just saw, in protein function prediction we are provided with

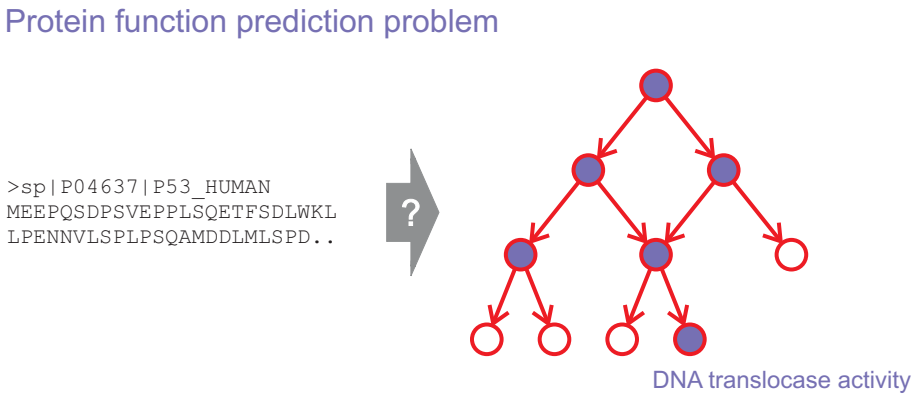

Figure 2: Illustration of the protein function prediction problem. The blue nodes in the ontology represent the predicted function of ths input protein. Note that the two leaf annotation nodes, one of which is "DNA translocase activity", uniquely identify the entire consistent subgraph.

a protein, and the goal is to predict the totality of its ontological annotation. In the gene prioritization scenario we are given a particular term v in the ontology, and the goal is to rank genes (or proteins) according to their likelihood of being associated with this particular term [21]. In the case when statistical inferene algorithms are capable of accurately estimating the posterior probabilities $\Pr(P|p)$ and $\Pr(v|g)$, the PFP and CGP problems are equivalent. However, this situation is not highly likely and, thus, the two problems are different in practice.

Let us briefly discuss the major differences between the two problem formulations. In the protein function prediction problem, we are given an unannotated protein and the space of possible outputs is the set of all consistent subgraphs in the ontology. Given the current sizes on biomedical ontologies, the space of all possible outputs (consistent subgraphs of $\mathcal{O}$) is vast. In the gene prioritization scenario, we are given a functional term in the ontology and asked to rank a relatively small set of candidate genes (at most equal to the size of genome). Due to the difference in problem formulation, the natural way to consider protein function prediction is the structured output learning framework or at the very least a multi-label classification. On the other hand, the candidate gene prioritization is an instance of a standard binary classification where one needs to find the probability that a gene is associated with a functional term. Finally, these two scenarios are very different when it comes to method evaluation. While in the protein function prediction scenario one needs to reason about graph similarity to assess accuracy of computational models, candidate gene prioritization is simple to evaluate using the well-understood metrics from binary classification.

2.1 Why is protein function prediction challenging?

There are a number of biological and computational challenges that make protein function prediction difficult. Let me enumerate a few good ones.

Biologically, protein function is determined in the context of the organism and is rarely, if ever, fully determined by any single experiment or publication. In addition, some experiments cannot be performed in some organisms for a variety of biological, budgetary, or ethical reasons. Some experiments are performed *in vitro* and may not faithfully reflect the protein's activity *in vivo*.[3] For example, if a protein is modified post-translationally, it might not carry out its function *in vitro,* because the modifying enzyme that activates the protein is missing. The available function data may also contain errors caused by misinterpretation of experiments, curation errors, experimental biases, and so on [7, 37, 38]. In short, the data provided in biological databases is (seriously) incomplete, biased, and noisy. In fact, because experimental annotations are incomplete, it is fair to question the reliability of the ground truth on which we evaluate predictors. To make things worse, current biological databases are gene-centric. That is, they list functions of all protein forms corresponding to their parent gene as one entry. At this time it is difficult to disentangle which protein form has what function.

Computationally, one can view protein function prediction as a multi-label classification problem or as an instance of structured-output learning in which the goal is to output a consistent subgraph ($\hat{T}$) of a graph ($\mathcal{O}$). Another challenge comes from the data integration perspective, where diverse biological data (sequence, structure, interactions, etc.) need to be dealt with. For example, for mice we can find a lot of gene expression data because a mouse is a reasonably reliable model organism for cancer study. We are also given a mouse PPI network (not as complete as yeast or human, but still good), or a number of mass-spectrometry data sets. On the other hand, we do not have such good volume and diversity of data for most other organisms. This makes statistical learning on one species and making inferences about another difficult. Finally, performance evaluation involves developing good similarity functions between pairs of consistent subgraphs in the ontology. It is not entirely clear how this should be done given that there are underlying relationships between terms in the ontology and differences in resolution with respect to which function is described in different branches of the ontology. The ontologies are large; there are about 40,000 terms in MFO, BPO, and CCO combined. Yet a typical protein is annotated by only a small number of terms. Even worse, the number of GO terms a protein is experimentally annotated with appears to have a scale-free-like distribution [8]. Understanding which proteins are annotated with many terms and which are annotated with only a few is very important.[4]

[3] A colleague one mine Yanay Ofran once started his comment at a coference with "We all know that what happens *in vitro* stays *in vitro*". A pretty good way to summarize it!

[4] It may be tempting to call proteins that carry out many functions "jacks of all trades", but there is really no evidence that they are 'masters of none'. It is plausible that these proteins are jacks of all trades and masters of them all. Life is not fair.

2.2 Why is protein function prediction important?

If we are to understand life at a molecular level we must understand how proteins carry out their function. This is also important for understanding the molecular mechanisms of disease, because alterations of protein function are responsible for many diseases. Here is one example: consider hemoglobin from Figure 1 and mutation of residue E into V at position 6 in its beta subunit (HBB gene in Swiss-Prot). This mutation is well-known to cause sickle cell disease, a disorder initiated by a binding event between multiple hemoglobin tetramers. This binding event, in turn, produces amyloid fibrils which disrupt the shape and function of a red blood cell, ultimately causing painful episodes and anemia. Computational function prediction can thus be used to formulate biological hypotheses and guide wet lab experiments through prioritization. It could also be used for rational drug design or to inform studies of molecular evolution. Unless new high-throughput functional assays become available (which I highly doubt!), it will not be possible to fill the gap between experimentally annotated proteins and the vast amount of sequence data already available (Figure 3).

Currently, there are about 7,000 sequenced genomes with about 21,000 in progress [26]. In addition, various efforts have catalogued about 1 million species on Earth and have estimated the existence of 10-100 million species in total. This data suggests a deepening of the gap between annotated and unannotated sequences and places computational annotation at the forefront.

Protein function prediction is also important from a statistical and computational perspective, because solutions found in computational biology may be helpful in other domains where ontological annotations are or will be assigned to objects. The data available in computational biology also provides opportunities for method development on incomplete, biased and highly heterogenious data.

2.3 Evaluation of protein function prediction algorithms

In the CAFA experiment organized in 2010-2011, we used two major types of evaluation: (*i*) protein-centric evaluation and (*ii*) term-centric evaluation [30]. The protein-centric evaluation is more related to the protein function prediction mode, whereas the term-centric evaluation is more appropriate for the gene prioritization scenario.

In the protein function prediction mode, we assume that the output of a predictor is a score for each term in the ontology. We will assume that the scores range between 0 and 1, with higher scores indicating more confident predictions. Thus, a decision threshold τ must be applied to determine the set of predicted terms $P(\tau)$.[5] Similarly, a set of experimentally determined terms will be denoted as T. To determine the quality of prediction, a similarity function must be calculated between $P(\tau)$ and T for each protein in an evaluation set.

[5]Technically, we assumed that a parent cannot have a score lower than any of its children nodes in the ontology. This can be taken care of during evaluation by assigning each parent the score that equals the maximum between its score and scores of all of its descendants.

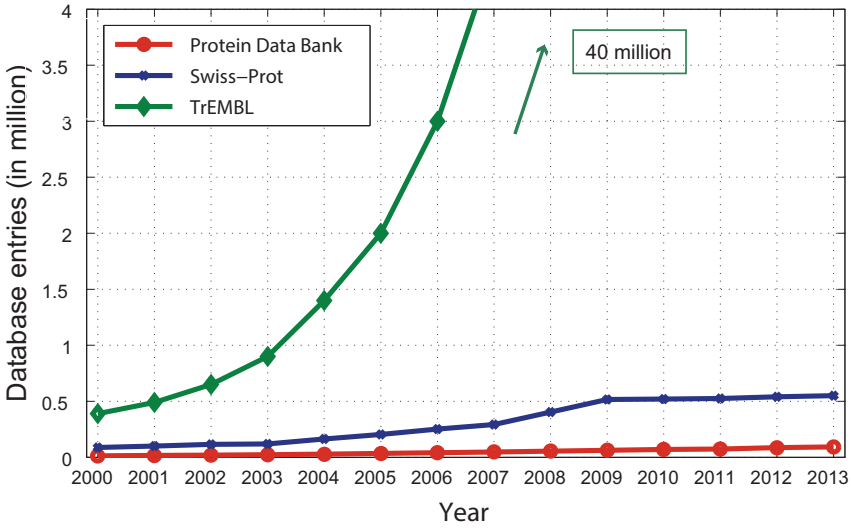

Figure 3: Growth of biological databases in the past decade or so. The Protein Data Bank plot shows the number of protein structures available over the years. Swiss-Prot lists the number of curated sequences that can be used as proxy to the growth of functional information. In August 2013 there were about 26,000 sequences annotated with at least one experimentally validated MFO term, about 40,000 sequences annotated with at least one BPO term, and another 40,000 sequences with at least one CCO term. These sets are overlapping; in total, there are only about 55,000 proteins in Swiss-Prot with at least a single experimental term in any of the three ontologies. The TrEMBL curve shows the growth of the number of uncurated protein sequences in UniProtKB.

For each protein i and threshold τ, we define its *precision* as

$$pr_i(\tau) = \frac{\sum_{v \in \mathcal{O}} I(v \in P_i(\tau) \wedge v \in T_i)}{\sum_{v \in \mathcal{O}} I(v \in P_i(\tau))}$$

and *recall* as

$$rc_i(\tau) = \frac{\sum_{v \in \mathcal{O}} I(v \in P_i(\tau) \wedge v \in T_i)}{\sum_{v \in \mathcal{O}} I(v \in T_i)},$$

where $I(\cdot)$ is an indicator function. Using a database of N proteins, we can determine the average precision from individual scores over a set of $m(\tau) \leq N$ proteins on which at least one prediction was made above threshold τ as

$$pr(\tau) = \frac{1}{m(\tau)} \cdot \sum_{i=1}^{m(\tau)} pr_i(\tau).$$

The average recall, on the other hand, is calculated as

$$rc(\tau) = \frac{1}{N} \cdot \sum_{i=1}^{N} rc_i(\tau)$$

on the entire set of N test proteins. This evaluation penalizes methods that do not make predictions on all proteins (each such protein is assigned a recall of 0 for all thresholds). While such evaluation is useful, I believe it is also important to evaluate each method using only the subset of proteins for which the method has made predictions. In other words, the number of proteins N in the previous expression can in such case be determined as $N = \max_{\tau}\{m(\tau)\}$.

When all possible thresholds are considered, a prediction model is characterized by a precision-recall curve $(pr(\tau), rc(\tau))_{\tau}$. To provide a single-score evaluation of computational models, we used the maximum F-measure over all thresholds. Specifically,

$$F_{\max} = \max_{\tau} \left\{ \frac{2 \cdot pr(\tau) \cdot rc(\tau)}{pr(\tau) + rc(\tau)} \right\}.$$

It is important to note that the F-measure, as a harmonic mean between $pr(\tau)$ and $rc(\tau)$, places equal emphasis on precision and recall. An experimental biologist, however, is more likely to be interested in predictions with high recall, because the knowledge of shallow terms in the ontology is not very informative for a detailed understanding of a protein's activity.

In the gene prioritization mode we use term-centric evaluation metrics. A Receiver Operating Characteristic (ROC) curve is particularly useful here (a precision-recall curves can be calculated as well). Consider term v and a set of N proteins on which the algorithm is evaluated. Let us define *sensitivity* (recall) and *specificity* as

$$sn_v(\tau) = \frac{\sum_{i=1}^{N} I(v \in P_i(\tau) \wedge v \in T_i)}{\sum_{i=1}^{N} I(v \in T_i)}$$

and

$$sp_v(\tau) = \frac{\sum_{i=1}^{N} I(v \notin P_i(\tau) \wedge v \notin T_i)}{\sum_{i=1}^{N} I(v \notin T_i)},$$

respectively. We can denote the ROC curve for a term v as $(sn_v(\tau), sp_v(\tau))_\tau$ and use the area under the curve (AUC) as a performance measure. Note that averaging the area under the curve over all terms may be a useful protein-centric metric, but it has severe issues related to estimating AUCs using a small number of positive examples associated with many terms as well as averaging scores over terms that are not independent. At this time I do not know of any universally accepted way to calculate and combine AUCs of individual terms to produce a good protein-centric score.

The metrics presented above were used in CAFA 1. They are reasonable but have their own drawbacks; for example, all terms in the ontology are treated equally to calculate $F_{\max}$, but these terms are neither independent nor equally important. There are alternative ways to evaluate function prediction. I will only mention hierarchical precision and recall by Verspoor et al. [42] and, shamelessly, misinformation and remaining uncertainty by Clark and myself [9]. Misinformation and remaining uncertainty are information-theoretic analogs of precision and recall. The minimum semantic distance is calculated from the misinformation and remaining uncertainty in a way similar to calculating $F_{\max}$ from precision and recall.

2.4 Protein function at a residue level

Most of the discussion here has been about function of a particular protein as a whole molecule. Unfortunately, this does not provide a complete picture about the machanistic aspects of its activity, the residues that are involved in particular function, or whether changes to these residues can alter the activity of this protein (and if so, how?). For example, let us consider a "DNA-binding activity" as our function of interest (it indeed is a term in MFO). With such terms, it becomes important to understand which particular residues are DNA-binding and how this function is impacted upon sequence variation of these or neighboring residues.

We can assign ontological terms to individual residues in a similar way we assign ontological terms to the entire molecules. These are not necessarily the same ontological terms, but in the many instances they are. In addition to DNA-binding, functional residues of interest would be catalytic residues, post-translationally modified sites, ligand-binding residues, protein-protein interaction residues, hot spots, metal-binding residues, etc. There is a large body of literature covering computational prediction of functional residues; for some review papers see [15, 14, 46]. These approaches also suggest that the prediction

of functional residues can be incorporated into the prediction of protein function at the whole-molecule level [18, 27].

3 The CAFA challenge

The Critical Assessment of Functional Annotation (CAFA) is a challenge dedicated to comprehensive evaluation of computational protein function prediction methods in an unbiased manner. The idea of the challenge is simple. The organizers (and I am one of them) provide a large set of unannotated or incompletely annotated proteins to the community and ask the groups to predict functional annotation for these proteins. All predictions must be uploaded at the CAFA web site before the submission deadline, which is a period of several months. After the submission deadline, we enter the phase of accumulation of experimentally validated annotations in UniProtKB. After about 6 months, or more during later re-assessments, the methods are evaluated on the proteins that accumulated new experimentally validated terms during the target accumulation phase. We as organizers have no control over the set of proteins that will be used for benchmarking. Biocurators in UniProtKB, who map experimental annotations to ontological terms, have no control over the predictions from the community. We think that the experiment is fair albeit less than ideal (for several reasons).

CAFA is one of many critical assessments in biomedical sciences [11]. Its distinguishing feature, however, is that the CAFA experiment is simple. We do not provide a data set that the predictors must use to train their data. We allow anyone to participate and use any data they may find useful. Our goal is just to provide a fair evaluation of the methods and gain insight into how well we are all doing, what works the best, and what ideas have stalled.

More details about CAFA can be found at http::/biofunctionprediction.org.

3.1 The main outcomes of CAFA 1 (2010-2011)

The first CAFA experiment was conducted between September 2010 and December 2011 [30], with a preliminary evaluation in July of 2011 during the ISMB conference in Vienna (AFP-SIG).[6] Computational prediction was evaluated in MFO and BPO categories of functional annotation on a total of 866 proteins from 11 species.

The major conclusions of the experiment can be summarized as follows: (*i*) the algorithms for protein function prediction developed in the 2000s clearly outperform traditional function transfer using a straightforward application of BLAST [2]; (*ii*) the performance of methods in the MFO category was assessed as useful to guide biological experiments; however, the performance in the BPO

[6] AFP-SIG stands for the Automated Function Prediction Special Interest Group meeting associated with the International Conference on Intelligent Systems for Molecular Biology (ISMB). ISMB is the flagship conference of the International Society for Computational Biology (ISCB).

category was below our expectations, particularly for eukaryotic species; (*iii*) for most algorithms, but excluding BLAST, the sequence similarity between a target and the most similar protein with experimentally annotated function is a relatively poor indicator of how good the overall prediction will be; this is because algorithms are capable of combining information from multiple sequence hits as well as integrating data other than sequence alignments; (*iv*) evaluation of protein function prediction seems to be equally challenging as the prediction itself. Due to data biases and incomplete experimental annotations there is no single metric that can be definitively used to assess function prediction; (*v*) the methods that performed best typically exploited multiple types of data, but there are exceptions; and (*vi*) although relatively good methods already exist for some prediction tasks, there is lack of standalone tools that are updated and maintained on a regular basis to be practically useful in guiding experimental biology.

The CAFA challenge was described in the main *Nature Methods* article [30], with a number of useful results presented in the Supplement. In addition, a special issue of the AFP-SIG has been published in *BMC Bioinformatics*, Volume 14, Supplement 3.

3.2 What is new in CAFA 2 (2013-2014)

We are introducing several novel aspects to CAFA 2 that will enable us to expand the challenge to new ontologies but also provide more detailed evaluation of protein function prediction tools. In terms of new ontologies, we are expanding CAFA 1 by adding the CCO and HPO annotations. While this may seem trivial, there is a difference between methods developed for MFO, BPO and those used for predicting the cellular compartment or human phenotypes. In addition to new ontologies, we would like to provide more detailed evaluation in two aspects: (*i*) unlike in CAFA 1 where each target was not allowed to have any experimental annotation in any of the ontologies prior to the submission deadline, in CAFA 2 targets that are partially annotated will also be considered. This will allow us to see, for example, whether knowledge of a protein's BPO annotation can be used to better infer its MFO, CCO or HPO (for human proteins) annotation. It will also allow us to better evaluate predictions on proteins that were partially annotated in the same ontology for which the evaluation is performed. For example, a protein may be annotated with a "zinc binding" term, and this could be useful information for infering other MFO terms for that protein; (*ii*) to better evaluate progress in function prediction, it is important to distinguish (as much as one can) to what extent is the performance improvement (if there is any) in CAFA 2 due to better/larger data used for model training compared to the influence of more powerful algorithms. To attempt to reason about this, we asked each group that participated in CAFA 1 to re-train their methods on the data from 2013, instead of 2010 (in addition to providing predictions of their new methods). As organizers, we will do this with the CAFA 1 baseline methods which will facilitate at least some understanding of where the improvement is coming from. In general, the predictions from all CAFA experiments will be

stored for future re-assessments and over times will provide us with a good data to track the progress of the field.

Finally, let me comment on data anonymity policy we have. Each group can remain anonymous and even withdraw from the experiment after the prediction submission deadline (and before the first presentation of results). We think this is an important policy, because some methods are hastily developed or may even have bugs. Groups that try new ideas but did not have time for internal rigorous accuracy assessment should not be panelized in terms of their reputation if the ideas did not work or there were problems in their code. However, we do not think that the performance accuracy of published methods should remain anonymized. In addition, all methods ranked in the top 10 based on the major assessment criteria will be de-anonymized (a small price to pay if you are in the top 10).

References

[1] G. Alterovitz, M. Xiang, D. P. Hill, J. Lomax, J. Liu, M. Cherkassky, J. Dreyfuss, C. Mungall, M. A. Harris, M. E. Dolan, J. A. Blake, and M. F. Ramoni. Ontology engineering. *Nat Biotechnol*, 28(2):128–130, 2010.

[2] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, 25(17):3389–3402, 1997.

[3] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, 25(1):25–29, 2000.

[4] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The Protein Data Bank. *Nucleic Acids Res*, 28(1):235–242, 2000.

[5] P. Bork, T. Dandekar, Y. Diaz-Lazcoz, F. Eisenhaber, M. Huynen, and Y. Yuan. Predicting function: from genes to genomes and back. *J Mol Biol*, 283:707–725, 1998.

[6] C. Branden and J. Tooze. *Introduction to protein structure*. Garland Publishing, Inc., New York, NY, USA, 1999.

[7] S. E. Brenner. Errors in genome annotation. *Trends Genet*, 15(4):132–133, 1999.

[8] W. T. Clark and P. Radivojac. Analysis of protein function and its prediction from amino acid sequence. *Proteins*, 79(7):2086–2096, 2011.

[9] W. T. Clark and P. Radivojac. Information-theoretic evaluation of predicted ontological annotations. *Bioinformatics*, 29(13):i53–i61, 2013.

[10] J. C. Costello, D. Schrider, J. Gehlhausen, and M. Dalkilic. Data-driven ontologies. *Pac Symp Biocomput*, pages 15–26, 2009.

[11] J. C. Costello and G. Stolovitzky. Seeking the wisdom of crowds through challenge-based competitions in biomedical research. *Clin Pharmacol Ther*, 93(5):396–398, 2013.

[12] A. K. Dunker, J. D. Lawson, C. J. Brown, R. M. Williams, P. Romero, J. S. Oh, C. J. Oldfield, A. M. Campen, C. M. Ratliff, K. W. Hipps, J. Ausio, M. S. Nissen, R. Reeves, C. Kang, C. R. Kissinger, R. W. Bailey, M. D. Griswold, W. Chiu, E. C. Garner, and Z. Obradovic. Intrinsically disordered protein. *J Mol Graph Model*, 19(1):26–59, 2001.

[13] J. Dutkowski, M. Kramer, M. A. Surma, R. Balakrishnan, J. M. Cherry, N. J. Krogan, and T. Ideker. A gene ontology inferred from molecular networks. *Nat Biotechnol*, 31(1):38–45, 2013.

[14] B. Eisenhaber and F. Eisenhaber. Prediction of posttranslational modification of proteins from their amino acid sequence. *Methods Mol Biol*, pages 365–384, 2010.

[15] I. Ezkurdia, L. Bartoli, P. Fariselli, R. Casadio, A. Valencia, and M. L. Tress. Progress and challenges in predicting protein-protein interaction sites. *Brief Bioinform*, 10(3):233–246, 2009.

[16] I. Friedberg. Automated protein function prediction–the genomic challenge. *Brief Bioinform*, 7(3):225–242, 2006.

[17] L. Hunter. Molecular biology for computer scientists. In L. Hunter, editor, *Artificial Intelligence for Molecular Biology*, pages 1–46. AAAI Press, 1993.

[18] L. J. Jensen, R. Gupta, N. Blom, D. Devos, J. Tamames, C. Kesmir, H. Nielsen, H. H. Staerfeldt, K. Rapacki, C. Workman, C. A. Andersen, S. Knudsen, A. Krogh, A. Valencia, and S. Brunak. Prediction of human protein function from post-translational modifications and localization features. *J Mol Biol*, 319(5):1257–1265, 2002.

[19] J. Kyte and R. F. Doolittle. A simple method for displaying the hydropathic character of a protein. *J Mol Biol*, 157:105–132, 1982.

[20] F. Leitner, A. Chatr-aryamontri, S. A. Mardis, A. Ceol, M. Krallinger, L. Licata, L. Hirschman, G. Cesareni, and A. Valencia. The FEBS Letters/BioCreative II.5 experiment: making biological information accessible. *Nat Biotechnol*, 28(9):897–899, 2010.

[21] Y. Moreau and L. C. Tranchevent. Computational tools for prioritizing candidate genes: boosting disease gene discovery. *Nat Rev Genet*, 13(8):523–536, 2012.

[22] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*, 247(4):536–540, 1995.

[23] NC-IUBMB. *Enzyme nomenclature*. Academic Press, New York, NY, USA, 1992.

[24] R. Nussinov and H. J. Wolfson. Efficient detection of three-dimensional structural motifs in biological macromolecules by computer vision techniques. *Proc Natl Acad Sci U S A*, 88(23):10495–10499, 1991.

[25] C. A. Orengo, A. D. Michie, S. Jones, D. T. Jones, M. B. Swindells, and J. M. Thornton. CATH–a hierarchic classification of protein domain structures. *Structure*, 5(8):1093–1108, 1997.

[26] I. Pagani, K. Liolios, J. Jansson, I. M. Chen, T. Smirnova, B. Nosrat, V. M. Markowitz, and N. C. Kyrpides. The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res*, 40(Database issue):D571–579, 2012.

[27] D. Pal and D. Eisenberg. Inference of protein function from protein structure. *Structure*, 13(1):121–130, 2005.

[28] T. M. Przytycka, M. Singh, and D. K. Slonim. Toward the dynamic interactome: it's about time. *Brief Bioinform*, 11(1):15–29, 2010.

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

[30] P. Radivojac, W. T. Clark, T. R. Oron, A. M. Schnoes, T. Wittkop, A. Sokolov, K. Graim, C. Funk, K. Verspoor, A. Ben-Hur, G. Pandey, J. M. Yunes, A. S. Talwalkar, S. Repo, M. L. Souza, D. Piovesan, R. Casadio, Z. Wang, J. Cheng, H. Fang, J. Gough, P. Koskinen, P. Toronen, J. Nokso-Koivisto, L. Holm, D. Cozzetto, D. W. Buchan, K. Bryson, D. T. Jones, B. Limaye, H. Inamdar, A. Datta, S. K. Manjari, R. Joshi, M. Chitale, D. Kihara, A. M. Lisewski, S. Erdin, E. Venner, O. Lichtarge, R. Rentzsch, H. Yang, A. E. Romero, P. Bhat, A. Paccanaro, T. Hamp, R. Kassner, S. Seemayer, E. Vicedo, C. Schaefer, D. Achten, F. Auer, A. Boehm, T. Braun, M. Hecht, M. Heron, P. Honigschmid, T. A. Hopf, S. Kaufmann, M. Kiening, D. Krompass, C. Landerer, Y. Mahlich, M. Roos, J. Bjorne, T. Salakoski, A. Wong, H. Shatkay, F. Gatzmann, I. Sommer, M. N. Wass, M. J. Sternberg, N. Skunca, F. Supek, M. Bosnjak, P. Panov, S. Dzeroski, T. Smuc, Y. A. Kourmpetis, A. D. van Dijk, C. J. ter Braak, Y. Zhou, Q. Gong, X. Dong, W. Tian, M. Falda, P. Fontana, E. Lavezzo, B. Di Camillo, S. Toppo, L. Lan, N. Djuric, Y. Guo, S. Vucetic, A. Bairoch, M. Linial, P. C. Babbitt, S. E. Brenner, C. Orengo, B. Rost, S. D. Mooney, and I. Friedberg. A large-scale evaluation of computational protein function prediction. *Nat Methods*, 10(3):221–227, 2013.

[31] R. Rentzsch and C. A. Orengo. Protein function prediction–the power of multiplicity. *Trends Biotechnol*, 27(4):210–219, 2009.

[32] P. N. Robinson and S. Bauer. *Introduction to bio-ontologies*. CRC Press, Boca Raton, FL, USA, 2011.

[33] P. N. Robinson and S. Mundlos. The human phenotype ontology. *Clin Genet*, 77(6):525–534, 2010.

[34] B. Rost. Twilight zone of protein sequence alignments. *Protein Eng*, 12(2):85–94, 1999.

[35] B. Rost, J. Liu, R. Nair, K. O. Wrzeszczynski, and Y. Ofran. Automatic prediction of protein function. *Cell Mol Life Sci*, 60(12):2637–2650, 2003.

[36] A. Ruepp, A. Zollner, D. Maier, K. Albermann, J. Hani, M. Mokrejs, I. Tetko, U. Guldener, G. Mannhaupt, M. Munsterkotter, and H. W. Mewes. The funcat, a functional annotation scheme for systematic classification of proteins from whole genomes. *Nucleic Acids Res*, 32(18):5539–5545, 2004.

[37] A. M. Schnoes, S. D. Brown, I. Dodevski, and P. C. Babbitt. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS Comput Biol*, 5(12):e1000605, 2009.

[38] A. M. Schnoes, D. C. Ream, A. W. Thorman, P. C. Babbitt, and I. Friedberg. Biases in the experimental annotations of protein function and their effect on our understanding of protein function space. *PLoS Comput Biol*, 9(5):e1003063, 2013.

[39] L. M. Schriml, C. Arze, S. Nadendla, Y. W. Chang, M. Mazaitis, V. Felix, G. Feng, and W. A. Kibbe. Disease Ontology: a backbone for disease semantic integration. *Nucleic Acids Res*, 40(Database issue):D940–946, 2012.

[40] R. Sharan, I. Ulitsky, and R. Shamir. Network-based prediction of protein function. *Mol Syst Biol*, 3:88, 2007.

[41] P. Tompa. Intrinsically disordered proteins: a 10-year recap. *Trends Biochem Sci*, 37(12):509–516, 2012.

[42] K. Verspoor, J. Cohn, S. Mniszewski, and C. Joslyn. A categorization approach to automated ontological function annotation. *Protein Sci*, 15(6):1544–1549, 2006.

[43] S. Vucetic, Z. Obradovic, V. Vacic, P. Radivojac, K. Peng, L. M. Iakoucheva, M. S. Cortese, J. D. Lawson, C. J. Brown, J. G. Sikes, C. D. Newton, and A. K. Dunker. DisProt: a database of protein disorder. *Bioinformatics*, 21(1):137–140, 2005.

[44] L. Wang and T. Jiang. On the complexity of multiple sequence alignment. *J Comput Biol*, 1(4):337–348, 1994.

[45] A. S. Wibowo, M. Singh, K. M. Reeder, J. J. Carter, A. R. Kovach, W. Meng, M. Ratnam, F. Zhang, and C. E. Dann. Structures of human folate receptors reveal biological trafficking states and diversity in folate and antifolate recognition. *Proc Natl Acad Sci U S A*, 110(38):15180–15188, 2013.

[46] F. Xin and P. Radivojac. Computational methods for identification of functional residues in protein structures. *Curr Protein Pept Sci*, 12(6):456–469, 2011.