BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 24 2012, pages 3203–3210
doi:10.1093/bioinformatics/bts608

Sequence analysis

Advance Access publication October 18, 2012

Assessing the relationship between conservation of function and conservation of sequence using photosynthetic proteins

Shaul Ashkenazi, Rotem Snir and Yanay Ofran*

The Goodman faculty of life sciences, Bar Ilan University, Ramat Gan 52900, Israel

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Assessing the false positive rate of function prediction methods is difficult, as it is hard to establish that a protein does not have a certain function. To determine to what extent proteins with similar sequences have a common function, we focused on photosynthesis-related proteins. A protein that comes from a non-photosynthetic organism is, undoubtedly, not involved in photosynthesis.

Results: We show that function diverges very rapidly: 70% of the close homologs of photosynthetic proteins come from non-photosynthetic organisms. Therefore, high sequence similarity, in most cases, is not tantamount to similar function. However, we found that many functionally similar proteins often share short sequence elements, which may correspond to a functional site and could reveal functional similarities more accurately than sequence similarity.

Conclusions: These results shed light on the way biological function is conserved in evolution and may help improve large-scale analysis of protein function.

Contact: yanay@ofranlab.org

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on June 4, 2012; revised on September 20, 2012; accepted on October 9, 2012

1 INTRODUCTION

Homology is believed to allow large-scale functional annotation. Since the 19th century, similarity between biological entities is defined as homology, in case that they evolved from a common ancestor, and as analogy, in case that they share a common function (Owen and Cooper, 1843). Later definitions made these terms mutually exclusive by defining analogy as convergence from two different ancestors (Fitch, 2000). Historically, these terms were introduced by naturalists and were based on anatomy and morphology. On the molecular level, however, they are used more loosely (Fitch, 2000), and homology is often used interchangeably with sequence similarity and is commonly believed to reflect similar function (Harrington *et al.*, 2007). Thus, although the vast majority of known proteins were never experimentally characterized, or even observed as RNA transcripts (Boeckmann *et al.*, 2003; Consortium, 2009), many of them have functional annotations that are based on sequence similarity to experimentally annotated proteins. These predicted functions are often used to perform functional comparisons between genomes and to analyse metagenomic data (Harrington *et al.*, 2007). One such analysis reported that by using BLAST, they were able to assign a function reliably to ≈70% of the gene products in new genomes (Harrington *et al.*, 2007). Despite the wide use of this approach, the performance of homology-based methods for function prediction was rarely assessed systematically.

Assessment of function prediction is sometimes considered 'not possible'. Assumptions regarding the relationship between sequence homology and functional analogy are difficult to assess. Many, if not most, of the proteins are multi-functional (Freitas *et al.*, 2009), making existing annotations partial. Therefore, when a predicted function does not correspond to an experimentally verified function of a protein, it is hard to determine whether it is a false prediction or a true prediction of a yet-unobserved function of that protein. Several collaborations and community-wide efforts have attempted to establish standards for assessment of computational annotation of protein function (Blaschke *et al.*, 2005; Friedberg *et al.*, 2006). The genome annotation assessment project (Reese *et al.*, 2000), an early attempt at quality assessment of genome-wide annotation, concluded that 'function predictions are the most difficult annotations to produce and to evaluate' and that owing to the incompleteness of existing annotation, 'a consistent and correct assessment of function predictions . . . was not possible'. CASP, which was originally designed to assess structure prediction methods, has widened its scope to assess also the function predictions (Lopez *et al.*, 2007; Soro and Tramontano, 2005). The assessors concluded, however, that except for the identification of ligand-binding residues, which is in essence structural, they could not find a procedure that allows quantitative assessment of performance (Lopez *et al.*, 2007). AFP/CAFA, another community-wide effort to establish a framework for assessing function prediction, was held five times since 2005 (Friedberg *et al.*, 2006; Rodrigues *et al.*, 2007). The organizers based their experiments on proteins whose functions were experimentally determined but not yet published. Assessment is based on measuring the distance between predicted and experimental annotations on the Gene Ontology (GO) graph. The benchmark sets that were established using this approach are small and do not allow for statistically significant assessment. The lack of clear definition of function is also a major drawback in assessing function prediction.

Studies that have assessed sequence similarity as a means for determining functional similarity gave contradicting results. Some concluded that sequence identity <80% (Rost, 2002)

*To whom correspondence should be addressed.

or, by other accounts, 60% (Tian and Skolnick, 2003) do not allow for reliable functional inference. Similarly, several studies have warned against blindly trusting homology as a proxy for functional identity (Gerlt and Babbitt, 2000). Studies that explored the potential for errors in genomic scale computational annotations (Bork and Koonin, 1998; Brenner, 1999; Devos and Valencia, 2001; Schnoes *et al*., 2009; Tatusov *et al*., 2003) reported error rates of between 5% and >60%.

Motifs and patterns identify similarity that is not based on homology. Another widely used tool for function assignment is sequence motifs, domains and patterns using tools such as Prosite (Sigrist *et al*., 2002), Interpro (Mulder and Apweiler, 2002), Pfam (Sonnhammer *et al*., 1997), MEX (Kunik *et al*., 2007) or MEME (Bailey *et al*., 2009). These tools typically provide a statistical measure for the significance of the specific motif, similar to BLAST's E-value. For instance, MEME's E-value, which is based on log likelihood ratio, width, sites, the background letter frequencies and the size of the training set, is an estimate of the expected number of motifs with the given log likelihood ratio (or higher), and with the same width and site count, that one would find in a similarly sized set of random sequences. Typically, these tools do not search for homology. Rather, they assume that common sequence motifs, which may reveal very distant homology or may be convergent analogy, constitute functional determinants, biologically important residues, binding sites or active sites in the protein (Aitken, 1999; Kunik *et al*., 2007; Sigrist *et al*., 2002; Wade, 2002). The precision and recall of Pfam as a means for function prediction was assessed in one study together with that of PSI-BLAST (Wass and Sternberg, 2008) as a benchmark for a GO-based function prediction method. Several other studies analysed the relationship between sequence similarity, domain conservation and function prediction (Heger *et al*., 2005; Portugaly *et al*., 2006; Schug *et al*., 2002). Some of them concluded that tools such as ProDom could reach accuracy of 87% in their functional assignments (Schug *et al*., 2002).

Function prediction of ORFans is particularly difficult. Many, possibly most, known protein sequences come from metagenomic projects that produce sequences of unidentified species, and include substantial numbers of ORFans, novel protein families and remote homologs to known proteins (Ellrott *et al*., 2010). It has been suggested that metagenomic data can shed new light on many biological processes, including photosynthesis (PS) (Yutin and Beja, 2005), as many marine organisms are believed to be photosynthetic (Cuvelier *et al*., 2010), and their divergence from known terrestrial photosynthetic organisms may be large. Indeed, metagenomic data have led to the identification of novel proteins that provide insights into PS (Yutin and Beja, 2005). Identifying novel families of photosynthetic proteins that are not obvious homologs of known PS proteins is a non-trivial task. However, when it comes to ORFans, the lack of homologs precludes the use of homology-based methods altogether.

PS provides a way to assess function prediction. We focus our analysis on PS-related functions not only for the importance of identifying novel photosynthetic proteins but also because they provide a way to assess functional assignments. An assignment of photosynthetic annotation to a protein that comes from an organism that is not photosynthetic is, without doubt, an error. Thus, it is possible to unequivocally identify false assignments and assess the possibility of finding novel photosynthetic families and remote homologs to known PS-related proteins. We used GO to compile a list of functions that are unique to PS (i.e. GO terms that describe functions, which are essential to PS but are also part of other biological processes, were not included in the analysis, see Section 2). We also created a non-redundant set of proteins that were experimentally verified to be involved in PS. Finally, we identified all Pfam domains that are related to PS. These lists of proteins, GO terms and Pfam domains allowed us to assess function assignments by Pfam, BLAST and PSI-BLAST. In addition, using MEME, we generated a list of short motifs (6–9 characters) for the experimentally verified PS proteins and used these motifs to predict the functions of new proteins. We thus explore conservation of function by assessing the performance of widely used methods that rely on completely different biological rationales.

2 METHODS

2.1 Materials

2.1.1 UniProtKB UniProtKB is composed of SWISS-PROT and TrEMBL. The TrEMBL (Boeckmann *et al*., 2003) database contains most known protein sequences. Unlike SWISS-PROT (Boeckmann *et al*., 2003), it also contains fragments of proteins and proteins translated computationally from metagenomic data (Consortium, 2009).

2.1.2 Parsing UniProtKB We parsed all entries in UniProtKB and searched the DR section for Pfam domains. Parsing was done in Python.

2.1.3 PS at GO From all GO ontologies, we extracted all child terms of PS and their child terms, recursively. From this list, we removed terms that are also children of terms that do not belong to the list. That is, terms that descend from PS but also from other functions that are not PS or its descendants, were removed. The remaining terms were defined as unique to PS. We found 184 991 proteins with these terms. Many of these proteins have PS-relevant annotation in more than one of the ontologies (Hill *et al*., 2008). We excluded from the analysis all annotations that were based on predictions. The evidence codes we used are therefore: EXP, IDA, IPI, IMP, IGI, IEP, TAS and NAS. We were left with 1425 proteins. Supplementary Table S1 lists these proteins. The GO numbers that were used are listed in Supplementary Table S2. The distribution of the different codes we found is presented in Supplementary Table S11.

2.2 Redundancy reduction

We removed redundancy from the set of experimentally annotated photosynthetic proteins and from the predicted sequences using CD-HIT (Yang *et al*., 2010), which clusters all proteins by similarity over length parameters (at least 50% similarity on 90% of protein's length in our case). From each such cluster, we retained a single representative. In the case of the set of experimentally annotated photosynthetic proteins, this resulted in 1256 unique proteins.

The CD-HIT redundancy reduction was done with the following command:

cd-hit -i <1425 fasta sequences> -o <output file> -aL 0.9 -c 0.5 -n 3 -T 0

Using different cut-offs for similarity (e.g. aL 0.75 –c 0.8) did not change the overall trends (data not shown).

2.3 Pfam motifs

We searched for all the UniProtKB sequences that contain a Pfam domain. We identified 267 599 proteins from UniProtKB that are annotated as photosynthetic by Pfam. As our assessment is based on determining whether the organism itself is photosynthetic, we considered only proteins from a known organism, not environmental samples or metagenomic sequences. The identification was based on 84 Pfam families, which were annotated as PS families by Pfam and were manually checked. The families are listed in Supplementary Table S5.

Pfam-GO motifs were extracted from InterPro (Lopez *et al.*, 2007). The families are listed in Supplementary Table S4.

2.4 Creating short PS motifs

Using MEME (Bailey *et al.*, 2009; Bailey *et al.*, 2006),

(1) we extracted 400 significant 6–9 characters sequence motifs for the set MEME6..9. After filtering (see Section 2.5), we were left with 390 motifs.

(2) we extracted 300 significant 7–9 characters sequence motifs for the set MEME7..9. After filtering (see section 2.5), we were left with 293 motifs.

These two sets of motifs were used to test how longer and shorter motifs fair in predicting function.

2.5 Motif elimination

To avoid statistical biases, we did not use any motif that had more than three repetitions of the same amino acid. e.g. 'HHHHHH[HQ][PR]' is unacceptable, whereas 'HHHTHH[HQ][PR]' is valid. Furthermore, in specific motifs, which had parentheses with the problematic extra repeated amino acid, we decided to modify the expression in the parentheses and keep the motif, as can be seen in Supplementary Table S10.

The MEME search was done with the following commands:

meme <1256 fasta sequences> -protein -mod zoops -nmotifs 100 -minsites 2 -maxsites 300 -minw 6 -maxw 6 -maxsize 650000

meme <1256 fasta sequences> -protein -mod zoops -nmotifs 100 -minsites 2 -maxsites 300 -minw 7 -maxw 7 -maxsize 650000

meme <1256 fasta sequences> -protein -mod zoops -nmotifs 100 -minsites 2 -maxsites 300 -minw 8 -maxw 8 -maxsize 650000

meme <1256 fasta sequences> -protein -mod zoops -nmotifs 100 -minsites 2 -maxsites 300 -minw 9 -maxw 9 -maxsize 650000

The resulting motifs sets are listed in Supplementary Tables S6, S7, S8 and S9. Note that long motifs may have a highly significant E-value and yet be of very little use, as they are not likely to be found in other proteins that are not a close homolog to the one from which the motif was derived.

2.6 Predictions based on MEME

The sequence motifs created by MEME were used to retrieve all the proteins that contain any of them.

2.7 HSSP

HSSP score between proteins was calculated based on the HSSP curve (Schneider et al., 1997). A protein was predicted to be photosynthetic if it had an HSSP score of $\geq$10 with one of the experimentally annotated PS proteins in the initial set of experimentally annotated proteins.

2.8 Assessing motifs

To assess the precision of the predictions, we checked whether a predicted PS protein comes from a species that belongs to taxonomy ranks that perform PS. The taxonomy ranks are listed in Supplementary Table S3. This was done to get the most optimistic assessment of the prediction. All other proteins with traceable source (e.g. proteins from vertebrates) were marked as false positives. Precision, which measures what fraction of the proteins that are predicted to be PS related are, indeed, PS related, is defined as:

$$precision = \frac{tp}{tp + fp} \quad (1)$$

where tp is the number of true positive predictions, namely the number of proteins predicted to be photosynthetic that originated from a photosynthetic organisms, and fp is the number of proteins predicted as photosynthetic that originated from a non-photosynthetic organism. Precision is typically complemented by another measure, recall, defined as:

$$recall = \frac{tp}{tp + fn} \quad (2)$$

where fn is the number of false negative predictions.

Recall assesses how well the method identifies the desired instances in the dataset. It is impossible to get an exact assessment of the recall in this case, as there is no benchmark set that includes all proteins that are involved in PS. As a proxy for recall, we recorded the number of hits, namely the number of proteins predicted by each method to be PS related. Combined with the precision, this measure provides an assessment of the performance of a prediction method.

2.9 Assessing sequence similarity as a prediction method

We performed BLAST and PSI-BLAST searches against the UniProtKB using our initial dataset of 1256 experimentally annotated PS proteins as queries. From the initial dataset, we looked for all BLAST results with $0 \leq$ E-value ≤ 1. PSI-BLAST parameters were as follows: Database used—UniProtKB; Number of iterations—3; PSI-BLAST Threshold for next iteration—0.001. In the end of the third iteration, we used all proteins with E-value $0 \leq$ E-value ≤ 1 as our dataset.

3 RESULTS

3.1 Photosynthetic-related proteins

Of 15 082 690 entries in UniProtKB, we extracted 184 991 with PS-related GO annotations. Of these, only 1425 proteins have GO evidence codes that are based on experiments. Clustering these proteins with CD-HIT (Yang et al., 2010) yielded 1256 sequence-unique experimentally annotated proteins. We used these proteins to identify new putative photosynthetic proteins based on sequence similarity, using BLAST and PSI-BLAST. It has been shown that using HSSP (Schneider et al., 1997) to rescore homology based also on the length of the alignment and not only on the sequence similarity can improve homology-based predictions (Rost, 2002). Therefore, we also assessed the performance of HSSP scores. We also extracted short sequence motifs from the initial set of sequence unique experimentally annotated proteins and used them to annotate other proteins as putatively PS related. In addition, we found in Pfam all the domains that are marked in Pfam as related to functions that are exclusively photosynthetic and used these domains to identify other proteins that are putatively PS related. Finally, we used the InterPro mapping of Pfam to GO terms (Lopez et al., 2007), which is a manual assignment of Pfam domains to specific functions in a similar manner to predict function of new proteins.

3.2 Short sequence motifs

Four hundred short sequence motifs were extracted by MEME from our non-redundant dataset of 1256 experimentally annotated PS proteins. The motifs are listed in Supplementary Tables S6, S7, S8 and S9. After filtering (see Section 2), we were left with 390 motifs that were used in two separated sets: MEME6..9, comprising all the motifs, and MEME7..9 (293 motifs), comprising only the longer motifs of length 7, 8 or 9.

3.3 Assessing precision for all methods

Figure 1 shows the precision and the number of hits for all the approaches. A functional assignment was defined as false if the query protein comes from a genus that is not known to include any photosynthetic species (see Section 2). Thus, the real precision is probably lower, as this assessment disregards false function assignments for proteins within photosynthetic genera. Proteins in photosynthetic organisms constitute close to 15% of UniProtKB proteins, and therefore random assignment of function is expected to have precision of around 0.15. All methods have precision that is substantially higher than random. BLAST, PSI-BLAST and Pfam are within a similar range (between 0.27 for PSI-BLAST and Pfam and 0.29 for BLAST). The two sets of MEME yield different results: MEME6..9 reaches precision of 0.3, similar to the other methods, whereas functional assignments that are based on the longer motifs of MEME7..9 reach substantially more accurate precision of 0.43. The variability in the number of hits was greater than that of the precision, with PSI-BLAST (against UniProtKB, see Section 2) identifying almost 10 hits for every hit of Pfam and >30 hits for every hit of MEME. Interestingly, the manually curated domains of Pfam-GO yielded exceptional precision of 0.74; however, the hits were meager. The reported results are after redundancy reduction with CD-HIT (Yang et al., 2010) (see Section 2).

These results are, by and large, consistent with the known trade-off between precision and recall, by which more accurate methods have lower discovery rates. The extreme case is that of the manually curated set of Pfam-GO. It is clear that these carefully selected domains are very tightly associated with PS. However, these domains can hardly identify any new proteins. We found that virtually all the proteins that are identified by Pfam-GO could be identified by BLAST or PSI-BLAST (data not shown). To assess how well each of these approaches identifies novel families, we checked the overlap between the non-redundant set of Pfam and all the other redundant sets, and the non-redundant set of MEME 6..9 and all the other redundant sets. The reason we performed the comparison this way is that we wish to determine how many of the unique sequences identified by Pfam and MEME, respectively, overlap with any protein identified by the other methods. HSSP, by definition, is contained within BLAST results; hence, its potential advantage is only in identifying better hits within BLAST hits, and not in identifying novel families. Similarly, BLAST hits are almost contained within PSI–BLAST hits (depending on PSI–BLAST

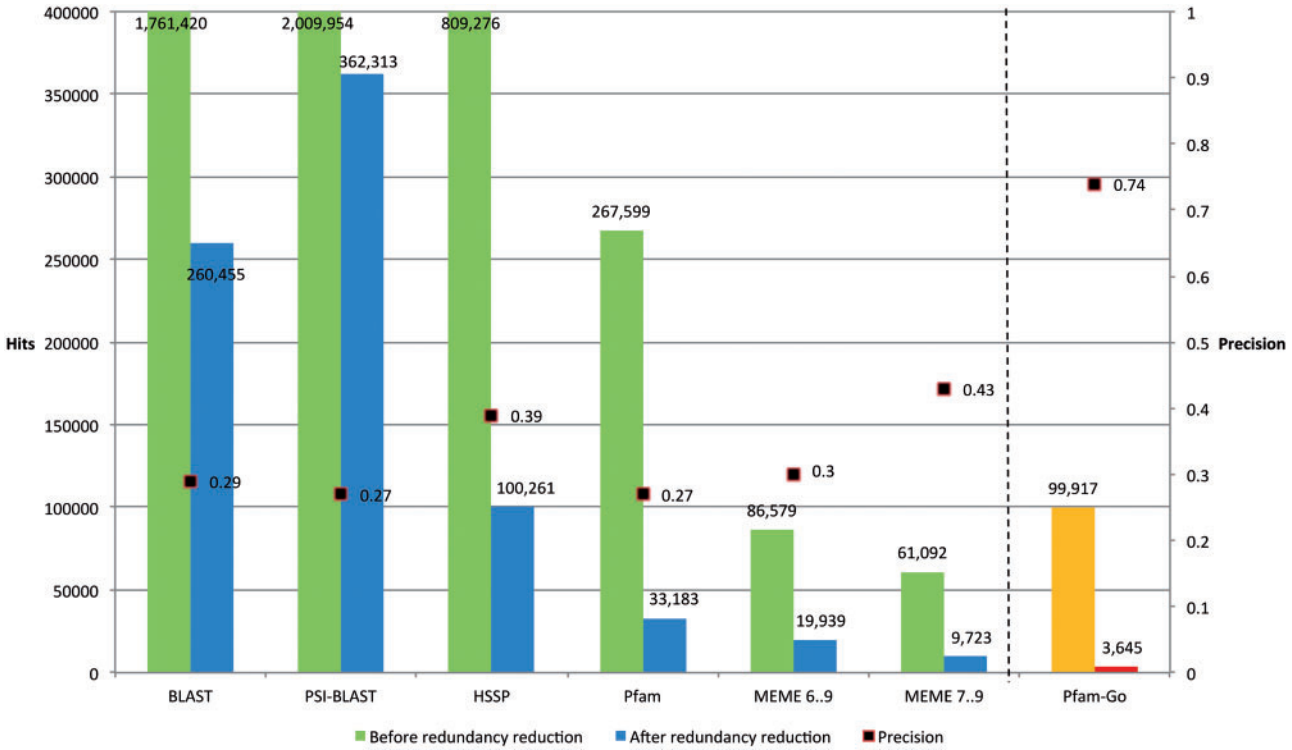

Fig. 1. Sequence relatedness and functional divergence according to different methods for functional inference. The precision (squares) was calculated based on the number of predicted proteins that come from photosynthetic organisms and the number of predicted proteins that come from organisms that are not photosynthetic. The number of hits (bars) is the number of sequence unique proteins identified by each method. Green (and blue) bars represent the number of hits before redundancy reduction (and after redundancy reduction). The yellow and red bars represent the results for the manually curated motifs in Pfam-GO. Precision was calculated for the non-redundant results. Redundancy was removed using CD-HIT

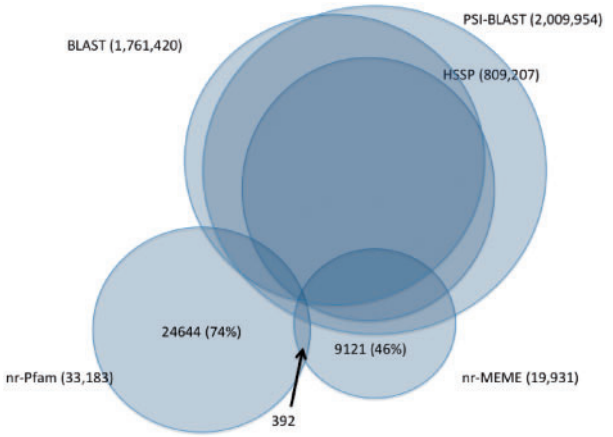

Fig. 2. Overlap between the novel proteins identified by each method. Although HSSP, BLAST and PSI-BLAST overlap very significantly by definition, to assess the unique protein families identified by MEME and Pfam, we compared the non-redundant (nr) set of hits for each of them to the full set of the other methods. The numbers in parentheses next to the circles are the number of hits for each method. The numbers inside the circles (with percentage in parentheses) are the numbers of proteins that were not identified by other methods. For example, the non-redundant set of proteins identified by Pfam as photosynthetic included 33 183 proteins. In all, 24 644 of these proteins, or 74% of them, were not identified by BLAST or PSI-BLAST. The overlap between the unique hits of MEME and Pfam was very low (only 392 proteins)

parameters). Figure 2 shows the results of this comparison. In all, 74% and 46% of the unique sequences identified by Pfam and MEME, respectively, are not detected by any of the other methods.

A possible way to further improve the precision of sequence motifs or domains is by trusting only proteins in which multiple motifs occur. Using MEME6..9 and MEME7..9, we found 805 different combinations of short motifs that occur in 5652 proteins, and 460 different combinations of short motifs that occur in 3922 proteins, respectively. Using Pfam and Pfam-GO, we found 22 different combinations of Pfam domains that occur in 403 proteins, and nine different combinations of Pfam domains that occur in 164 proteins, respectively.

As shown in Figure 3, this increased the precision of MEME6..9 and MEME7..9 to 0.55 and 0.6, respectively, and the precision of Pfam to 0.45. Clearly, this comes on the account of the number of hits. The manual set of domains identified by Pfam-GO reaches precision of 0.98, but the number of hits drops to 164 proteins—an order of magnitude below the size of the initial set of experimentally annotated proteins, which suggests that in this setting Pfam-GO is unable to find novel PS-related proteins. The number of hits for Pfam was also low (403). This is not surprising, given that Pfam motifs, by design, tend to be much longer than MEME motifs.

3.4 Smaller E-value does not always imply more similar function

Several studies have searched for a level of sequence similarity that can assure identical function (Rost, 2002). To test whether closer homologs are less functionally divergent, we checked the precision and number of hits for different ranges of E-values. As shown in Figure 4, the effect of increasing E-value on precision is hardly noticeable for most E-values. There is a significant increase in precision only for very small E-values. Thus, at least for PS proteins, it seems that more significant E-values may not increase precision in most cases. The precision of the un-binned sets (Fig. 1) is slightly lower than the precision of the binned data. This effect is due to the redundancy in the un-binned sets. A protein may be homologous to several proteins from our initial set and thus appear in several bins, each time different E-values. This is more likely to happen to a real photosynthetic protein (i.e. a PS protein is more likely than a non-PS protein to be similar to several PS proteins). When we group all the BLAST results together without binning different E-values. This protein will appear only once. When we bin them according to their E-value, it may appear in several different bins. Therefore, redundancy reduction of hits for the binned sets, which was done for each bin separately, has a slightly higher chance of retaining positive examples than redundancy reduction for the entire set.

3.5 Errors in short motifs-based functional annotation of PS proteins

Reviewing the false positive predictions, namely proteins that were wrongly identified as PS related, revealed some common errors. Although some proteins, such as Q27472_CAEEL (found by PSI-BLAST) from *Caenorhabditis elegans* or Q0PWT1_DIACI (found by Pfam) from Asian citrus psyllid (an insect), come from organisms that are undoubtedly not photosynthetic, other cases are more subtle. Some of these annotations can be explained by similarity of functions: proteins, which are involved in the electron transport chain, ATP transformers or ubiquinon related functions, which are part of the PS process, may be homologous to proteins in similar processes in non-photosynthetic organisms and thus might be mistakenly associated with PS. Some proteins may have an ancestral source among PS organisms. Although they evolved to different functions, the motif may have remained intact. A4IE64_LEIIN is an example for a protein that includes several strongly predictive PS-related motifs, but belongs to *Leishmania infantum* (a non-photosynthetic eukaryote). Previous studies have shown that trypanosomatid parasites, including Leishmania, possessed a plastid at some point in their evolutionary history, and that their extant genomes contain 'plant-like' genes encoding homologs of proteins found in either chloroplasts or in the cytosol of plants and algae (Hannaert et al., 2003; Weber and Fischer, 2007). Another source of false positive is marine viruses. Q6H943_9CAUD is one of many proteins that were predicted by Pfam to be PS2 complex proteins but come from viral genomes. A recent study (Sharon et al., 2009) revealed the presence of PS proteins in marine viruses. Obviously, false positive may arise also from short motifs that randomly occur in some sequences.

3.6 Homology should not be automatically considered a proxy for functional similarity

BLAST, PSI-BLAST, Pfam and short motifs generated by MEME represent different approaches for the detection of

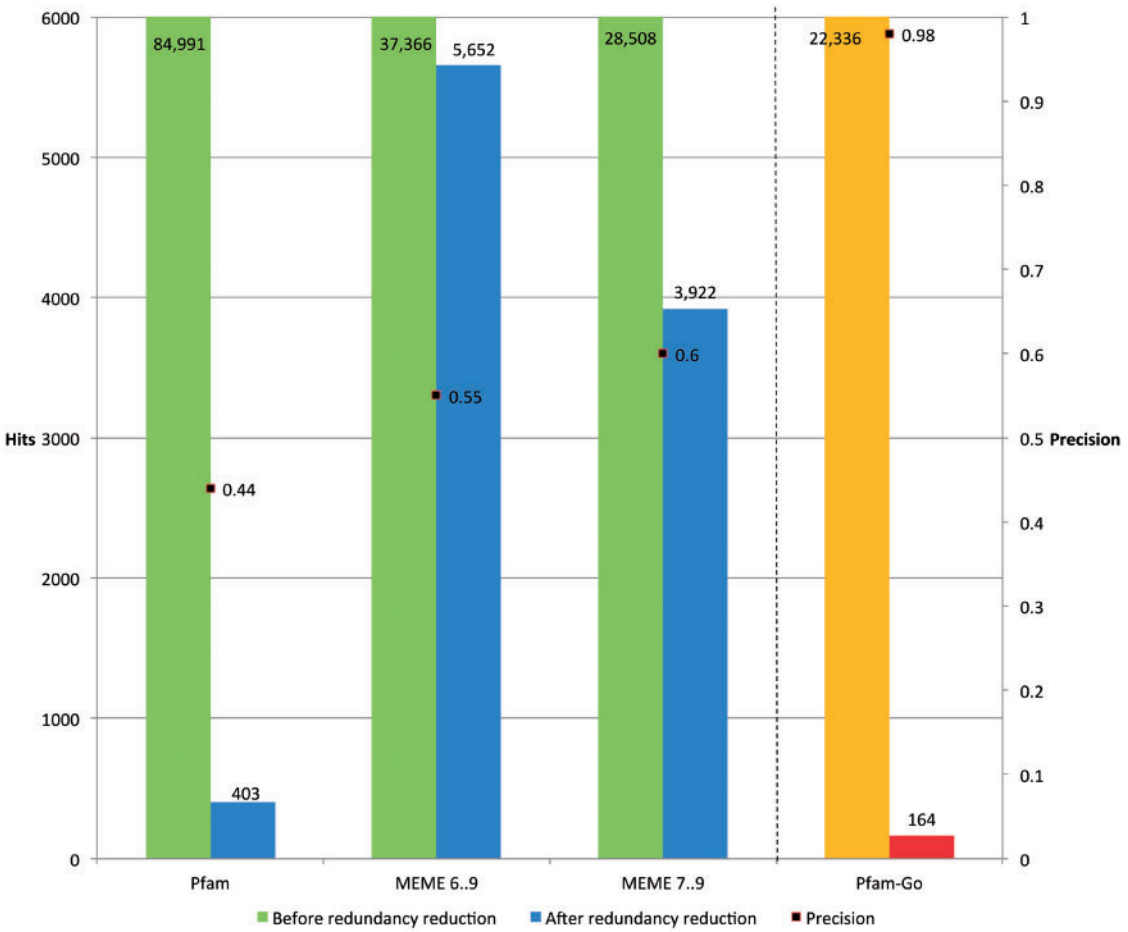

Fig. 3. Performance of multiple domains and multiple short motifs. Trusting only the functional assignments that are based on at least two motifs (or domains) of MEME (or Pfam) in each protein resulted in better overall precision at the cost of fewer hits. The manually curated domains of Pfam-GO reach striking precision of 0.98; however, they hardly discovered any proteins, and MEME7..9 reached precision of 0.6 with 3922 hits. The yellow and red bars represent the results for the manually curated motifs in Pfam-GO

homology and remote homology based on sequence similarity. Although there are more direct ways to assess homology (e.g. through phylogentic analysis), these methods gained their popularity, among other reasons, because of the implicit assumption that they are quick to identify functionally related sequences. Our results, however, show that these expectations may not be correct, particularly when it comes to the annotation of metagenomic data. Despite the sophistication of Pfam, its precision is not better than that of the universally applicable PSI-BLAST, BLAST or automatically detected short motifs, and it finds much less proteins than BALST or PSI-BLAST. However, most of the protein families Pfam identifies could not be detected by the other methods. BLAST itself provides at least 70% false functional assignments, which improves only for extremely high E-vlaues (Fig. 4). Our results indicate that using the manually curated motifs of Pfam-GO leads to the most reliable predictions of function. However, the extremely low number of hits indicates that this approach hardly finds novel proteins and identifies mostly the proteins that were used to extract the motifs. Short motifs identified by MEME, which may be poor indicators of homology can reflect function more accurately than homology-based methods. However, this comes at the cost of lower number of hits. As shown in Figure 2, our results reflect the complementarity of the different approaches: BLAST, Pfam and short motifs identify novel proteins that could not be detected by the other methods. Of the true positives of MEME7..9, 20% are not discovered by any other method. Of the PFAM true positives, 76% are not discovered by any other method. It is possible to improve precision on the account of recall for short motifs by requiring more than one motif in each protein. It is possible to achieve similar improvement by increasing BLAST E-value. Thus, at least when it comes to the ability to discover new PS proteins in metagenomic data, the combination of short motif-based approaches with homology-based ones seems a promising path.

It is not clear whether sharing a common sequence motif (or a combination thereof) is more likely to be the result of remote homology or the result of convergence. This question has far-reaching implications for the understanding of the evolution of function. Addressing it may require the development of new tools for the analysis of function and a thorough review of the phylogeny of each sequence. We can speculate that some

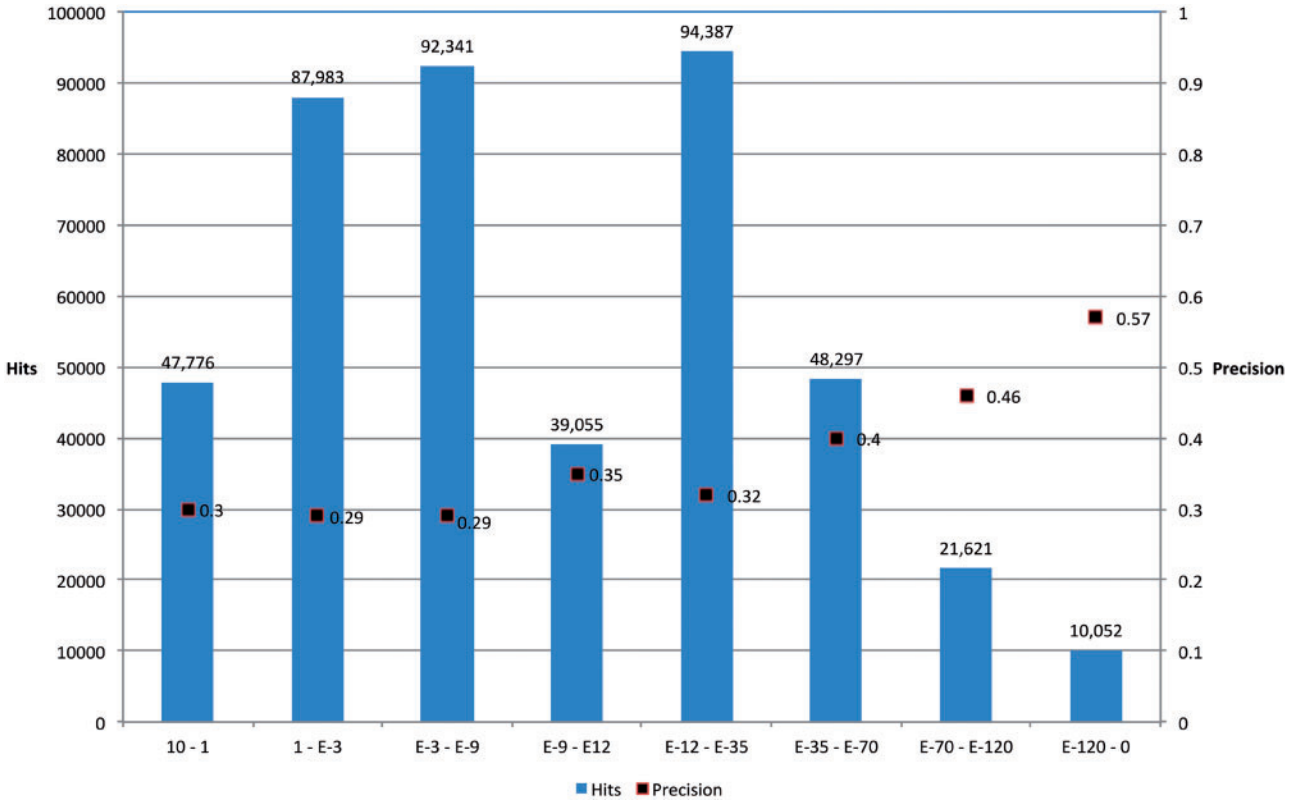

Fig. 4. Precision and number of hits for different BLAST E-values. To assess to what extent more significant sequence similarities imply greater functional similarity, we checked the precision and recall (after redundancy reduction) of different intervals of BLAST E-values, using all 1256 experimentally annotated PS-related proteins in our set as queries

functions are more likely than others to be predictable-based sequence motifs. For example, in the GO classifications, molecular function may be more predicable from motifs than biological processes. However, for each function this should be assessed separately.

4 CONCLUSIONS

Photosynthetic proteins provide an opportunity to assess the false positive rate of function prediction methods. At least for photosynthetic proteins, false positive rate of methods that are based on overall sequence similarity is ~70%. Short motifs-based approaches have false positive rate of 57%, but this improved precision comes with lower discovery rate. Beyond the evolutionary insights that may be drawn from these findings, they indicate that combining these approaches will improve large-scale function prediction

Funding: This work was supported in part by Microsoft Research Connections.

Conflict of Interest: none declared.

REFERENCES

Aitken,A. (1999) Protein consensus sequence motifs. *Mol. Biotechnol.*, **12**, 241–253.

Bailey,T.L. *et al.* (2009) Meme suite: tools for motif discovery and searching. *Nucleic Acids Res.*, **37**, W202–W208.

Bailey,T.L. *et al.* (2006) Meme: discovering and analyzing dna and protein sequence motifs. *Nucleic Acids Res.*, **34**, W369–W373.

Blaschke,C. *et al.* (2005) Evaluation of biocreative assessment of task 2. *BMC Bioinformatics*, **6** (**Suppl. 1**), S16.

Boeckmann,B. *et al.* (2003) The swiss-prot protein knowledgebase and its supplement trembl in 2003. *Nucleic Acids Res.*, **31**, 365–370.

Bork,P. and Koonin,E.V. (1998) Predicting functions from protein sequences–where are the bottlenecks? *Nat. Genet.*, **18**, 313–318.

Brenner,S.E. (1999) Errors in genome annotation. *Trends Genet.*, **15**, 132–133.

Consortium,T.U. (2009) The universal protein resource (uniprot) 2009. *Nucleic Acids Res.*, **37**, D169–D174.

Cuvelier,M.L. *et al.* (2010) Targeted metagenomics and ecology of globally important uncultured eukaryotic phytoplankton. *Proc. Natl Acad. Sci. USA*, **107**, 14679–14684.

Devos,D. and Valencia,A. (2001) Intrinsic errors in genome annotation. *Trends Genet.*, **17**, 429–431.

Ellrott,K. *et al.* (2010) Expansion of the protein repertoire in newly explored environments: human gut microbiome specific protein families. *PLoS Comput. Biol.*, **6**, E1000798.

Fitch,W.M. (2000) Homology a personal view on some of the problems. *Trends Genet.*, **16**, 227–231.

Freitas,R.N. *et al.* (2009) A HMGCR polymorphism is associated with relations between blood pressure and urinary sodium and potassium ratio in the Epic-Norfolk study. *J. Am. Soc. Hypertens.*, **3**, 238–244.

Friedberg,I. *et al.* (2006) New avenues in protein function prediction. *Protein Sci.*, **15**, 1527–1529.

Gerlt,J.A. and Babbitt,P.C. (2000) Can sequence determine function? *Genome Biol.*, **1**, REVIEWS0005.

Hannaert,V. *et al.* (2003) Plant-like traits associated with metabolism of trypanosoma parasites. *Proc Natl Acad. Sci. USA*, **100**, 1067–1071.

Harrington,E.D. *et al.* (2007) Quantitative assessment of protein function prediction from metagenomics shotgun sequences. *Proc. Natl Acad. Sci. USA*, **104**, 13913–13918.
Heger,A. *et al.* (2005) ADDA: a domain database with global coverage of the protein universe. *Nucleic Acids Res.*, **33**, D188–D191.
Hill,D.P. *et al.* (2008) Gene ontology annotations: what they mean and where they come from. *BMC Bioinformatics*, **9** (**Suppl. 5**), S2.
Kunik,V. *et al.* (2007) Functional representation of enzymes by specific peptides. *PLoS Comput. Biol.*, **3**, E167.
Lopez,G. *et al.* (2007) Assessment of predictions submitted for the casp7 function prediction category. *Proteins*, **69** (**Suppl. 8**), 165–174.
Mulder,N.J. and Apweiler,R. (2002) Tools and resources for identifying protein families, domains and motifs. *Genome Biol.*, **3**, REVIEWS2001.
Owen,R. and Cooper,W.W. (1843) *Lectures on the Comparative Anatomy and Physiology of the Invertebrate Animals: Delivered at the Royal College of Surgeons, in 1843*. Longman, Brown, Green, and Longmans, London.
Portugaly,E. *et al.* (2006) Everest: automatic identification and classification of protein domains in all protein sequences. *BMC Bioinformatics*, **7**, 277.
Reese,M.G. *et al.* (2000) Genome annotation assessment in drosophila melanogaster. *Genome Res.*, **10**, 483–501.
Rodrigues,A.P. *et al.* (2007) The 2006 automated function prediction meeting. *BMC Bioinformatics*, **8** (**Suppl. 4**), S1–S4.
Rost,B. (2002) Enzyme function less conserved than anticipated. *J. Mol. Biol.*, **318**, 595–608.
Schneider,R. *et al.* (1997) The HSSP database of protein structure-sequence alignments. *Nucleic Acids Res.*, **25**, 226–230.
Schnoes,A.M. *et al.* (2009) Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS Comput. Biol.*, **5**, E1000605.
Schug,J. *et al.* (2002) predicting gene ontology functions from prodom and cdd protein domains. *Genome Res.*, **12**, 648–655.
Sharon,I. *et al.* (2009) Photosystem I gene cassettes are present in marine virus genomes. *Nature*, **461**, 258–262.
Sigrist,C.J. *et al.* (2002) Prosite: a documented database using patterns and profiles as motif descriptors. *Brief. Bioinform.*, **3**, 265–274.
Sonnhammer,E.L. *et al.* (1997) Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins*, **28**, 405–420.
Soro,S. and Tramontano,A. (2005) The prediction of protein function at CASP6. *Proteins*, **61** (**Suppl. 7**), 201–213.
Tatusov,R.L. *et al.* (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**, 41.
Tian,W. and Skolnick,J. (2003) How well is enzyme function conserved as a function of pairwise sequence identity? *J. Mol. Biol.*, **333**, 863–882.
Wade,R.H. (2002) Sequence landmark patterns identify and characterize protein families. *Structure*, **10**, 1329–1336.
Wass,M.N. and Sternberg,M.J. (2008) Confunc–functional annotation in the twilight zone. *Bioinformatics*, **24**, 798–806.
Weber,A.P. and Fischer,K. (2007) Making the connections–the crucial role of metabolite transporters at the interface between chloroplast and cytosol. *FEBS Lett.*, **581**, 2215–2222.
Yang,F. *et al.* (2010) Using affinity propagation combined post-processing to cluster protein sequences. *Protein Pept. Lett.*, **17**, 681–689.
Yutin,N. and Beja,O. (2005) Putative novel photosynthetic reaction centre organizations in marine aerobic anoxygenic photosynthetic bacteria: insights from metagenomics and environmental genomics. *Environ. Microbiol.*, **7**, 2027–2033.