BIOINFORMATICS APPLICATIONS NOTE Vol. 26 no. 18 2010, pages 2361–2362
doi:10.1093/bioinformatics/btq426

Databases and ontologies

Advance Access publication August 6, 2010

RefProtDom: a protein database with improved domain boundaries and homology relationships

Mileidy W. Gonzalez[1,*] and William R. Pearson[2,*]

[1]Department of Biological Sciences, University of Maryland Baltimore County, 1000 Hilltop Circle, Baltimore, MD 21250 and [2]Department of Biochemistry and Molecular Genetics, Jordan Hall Box 800733, 1340 Jefferson Park Ave., Charlottesville, VA 22908, USA

Associate Editor: Dmitrij Frishman

ABSTRACT

Summary: RefProtDom provides a set of divergent query domains, originally selected from Pfam, and full-length proteins containing their homologous domains, with diverse architectures, for evaluating pair-wise and iterative sequence similarity searches. Pfam homology and domain boundary annotations in the target library were supplemented using local and semi-global searches, PSI-BLAST searches, and SCOP and CATH classifications.

Availability: RefProtDom is available from http://faculty.virginia.edu/wrpearson/fasta/PUBS/gonzalez09a

Contact: miledywgonzalez@gmail.com; pearson@virginia.edu

Received on January 13, 2010; revised on June 8, 2010; accepted on July 18, 2010

1 INTRODUCTION

Evaluation and improvement of protein sequence similarity searches, using algorithms such as BLAST or Smith-Waterman (SSEARCH) and more sophisticated searches such as PSI-BLAST or HMMER (Altschul *et al*., 1997; Durbin, 1998; Smith and Waterman, 1981), require query sequences and reference sets curated to accurately reflect homology relationships. Because structural similarity is preserved well beyond sequence similarity (Gibrat *et al*., 1996), protein structures are often the gold standard for annotating homology relationships. Although both structure-based homology annotations and manually annotated protein sequence relationships can very accurately record homology relationships, they do not reflect common practice in protein similarity searching, which is to characterize unknown proteins by searching large, comprehensive protein sets such as RefSeq (Pruitt *et al*., 2007) and UniProt (Consortium, 2009).

To better characterize similarity searching strategies, in particular, PSI-BLAST performance, against comprehensive protein databases, we identified a set of diverse protein domains from Pfam (Finn *et al*., 2010) v. 21 to use as queries against a set of real proteins containing those domains. Our query domain families are taxonomically broad (to provide 'harder' homology detection cases), and have long models (to better simulate full-length protein searches). Although we cannot be certain that all homologs have been found, we believe that statistically significant pair-wise alignments are annotated correctly.

*To whom correspondence should be addressed.

2 DATABASE ASSEMBLY

Evaluation datasets: from 681 initial Pfam (v. 21) families that met criteria for: (i) domain length (>200 residues); (ii) taxonomic diversity (present in two of bacteria, archaea and eukarya); (iii) family size (>100 instances); and (iv) available structure, we selected 344 query Pfam families after merging families that belonged to the same clan (Gonzalez and Pearson, 2010). In this initial set, 81 families belonged to distinct clans, while 263 families did not have an associated clan. This set was reduced to 320 non-homologous domains using information from Pfam (v. 23) (by Pfam v. 24, these domains belonged to 112 distinct clans with 168 families not in clans, for 280 non-homologous domains).

The target library was built from 234 505 full-length UniProt proteins (excluding viral sequences) containing Pfam v. 21 homologs to the original 320 Pfam families together with 1627 other domain families. Two query sets were constructed and the members of these sets evaluated further: (i) a challenging query subset (50 hard) with the lowest family coverage with BLAST; and (ii) a randomly sampled representative query set (50 sampled with replacement).

Annotation extensions: when the original Pfam v. 21 annotations were used to characterize searches with our *hard* and *sampled* queries against the target library, thousands of alignments to very similar UniProt sequences (e.g. $E() < 10^{-80}$, with >95% identity) were annotated as partial homologs or non-homologs. To correct these conservative annotations, we compared the bare domain query sequences to the target library using SSEARCH and GLSEARCH (a program that produces an alignment that is global in the query sequence but possibly local in the target or library sequence). We identified all the significantly similar sequence regions ($E() < 0.001$) with SSEARCH that were either shorter or unannotated in Pfam v. 21 and calculated the boundaries using GLSEARCH. We extended annotations on 2106 partial domains and added 24 604 domain homology annotations based on SSEARCH alignments, 13 574 of which were included in Pfam v. 24. RefProtDom describes relationships and alignment boundaries between query domains and the target library homologs according to Pfam v. 21, Pfam v. 24 and the SSEARCH/GLSEARCH alignment boundaries.

Although SSEARCH/GLSEARCH searches against the target library dramatically reduced the number of apparent false positives with very low $E()$-values, additional searches with PSI-BLAST using the queries sometimes found 'unrelated' UniProt sequences with significant ($E() < 10^{-40}$) scores. We analyzed all significant ($E() < 10^{-4}$) 'non-homologous' alignments found in the first three iterations of PSI-BLAST for the 100 queries (94 distinct families). Non-homologous alignments to regions with no annotated Pfam domains were used as queries in reciprocal PSI-BLAST searches for three iterations. Reciprocal searches that recovered at least 25% of a domain family were annotated as homologous, yielding 375 additional homology annotations across 33/94 families. Structures of significant 'non-homologs' that mapped to unrelated Pfam families were examined in SCOP and CATH; if they shared the same SCOP fold or CATH topology they were annotated as homologs. For example, Pfam annotates a

PF00346 domain for residues 295–537 on the Q8ZMJ0_SALTY sequence. RefProtDom also annotates PF00374 in the same region (390–535) because both domains share the same SCOP fold and superfamily classifications (e.18.1). This structural annotation is also reported by SCOOP (Bateman and Finn, 2007). We found structural evidence to add 37 additional clans; Pfam v. 24 matches 14 of those 37 structural clans. These additional clans would reduce the number of non-homologous domains in our query set from 94 to 90, but the families were not combined, the cryptic homology was simply annotated. Structure classifications yielded 2124 additional homology annotations across 16/94 queries.

3 SUMMARY

Iterative similarity searches are usually performed against full-length proteins with complex domain architectures. Evaluating similarity-searching methods against benchmarking sets with incomplete or missing annotations can introduce dramatic statistical inaccuracies. RefProtDom's greatest strength is its use of a taxonomically diverse set of full-length, multi-domain, proteins in the target library. Searches against RefProtDom resemble searches against SwissProt, UniProt or RefSeq (though those databases are much larger). Moreover, the query sequences are evolutionarily independent; based on structural comparisons, 90 query domains are non-homologous. Thus, RefProtDom can simulate searches against comprehensive sequence databases while evaluating success on challenging homologies.

Pfam is a powerful resource for identifying homology relationships and domain boundaries, but strategies that use a single hidden Markov model (HMM) to identify every homolog will be challenged by distant sequences at the detection horizon for the model. For many families, Pfam has addressed this problem by grouping families into clans. But, sometimes homologs are missed; sequences that share strong similarity across the length of a domain to an annotated homolog are surely homologous, even if they do not produce a significant score against the HMM model.

The RefProtDom query and target libraries seek to reduce the number of un-annotated homologies with statistically significant similarities, and to more accurately estimate homologous domain boundaries. Although our curation may have missed some homologs, we are confident in the homologies we annotate. Homology is annotated for domains that share significant pair-wise similarity, show significant family coverage after three PSI-BLAST iterations, or when they share structures. Domain boundaries were revised based on significant local or global similarity. By combining single domain queries with full-length, multi-domain proteins, RefProtDom can highlight alignment errors and evaluate improvements in alignment accuracy.

Accurate boundary annotation has been largely overlooked in pair-wise sequence comparison, because incorrect alignment boundaries rarely detract from the identification of homologous proteins. Pfam's annotations are now generated with HMMER3, which only performs local alignments (Finn *et al*., 2010). Therefore, one might expect that future Pfam annotations may have an even harder time at identifying complete domains, and thus, should continue to benefit from the extension curation provided by RefProtDom. Nonetheless, HMMER3 compensates with increased sensitivity as a result of better statistics. In fact, 59% of the domain extensions and 55% of the missed homologs added to Pfam v. 21 using our protocol were incorporated in Pfam v. 24. Thus, Pfam v. 24, using HMMER3, has independently addressed many of the missing annotations, validating our approach. However, we believe that the problems inherent in using a single model for diverse protein family searches will always miss homologs and domain boundaries that can be found with individual domains across the family's phylogenetic tree. We plan to continue to update the homology relationships and boundary assignments in RefProtDom.

For iterative sequence comparison methods, alignment accuracy is crucial; inaccurate alignments can cause non-homologous domains to be included in the profiles and decrease their specificity in subsequent iterations. Using RefProtDom's annotations, we identified a previously unrecognized alignment overextension error in PSI-BLAST responsible for the corruption of its PSSMs and its poor specificity (Gonzalez and Pearson, 2010). Additional evaluations with RefProtDom revealed that while JACKHMMER (HMMER3's iterative implementation) is susceptible to the same error, it overextends more slowly and, thus, shows better performance than unmodified PSI-BLAST (M.W.G. and W.R.P., manuscript in preparation).

Domains are the basic units of protein function and evolution; thus, improved homology detection requires improved domain alignment accuracy. Large-scale automatic annotation of gene function is limited by local alignments' incomplete motif matches and fuzzy domain boundaries (Kann *et al*., 2007). Establishing homology is central to a wide array of bioinformatics methodologies; improved domain alignments can improve 3D protein structural predictions that use homology modeling, and also clarify how protein domain networks interact to generate disease phenotypes. RefProtDom provides a comprehensive set of full-length UniProt proteins that can be used to evaluate domain alignment accuracy.

Funding: National Library of Medicine, grant LM04969.

Conflict of Interest: none declared.

REFERENCES

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

Bateman,A. and Finn,R.D. (2007) SCOOP: a simple method for identification of novel protein superfamily relationships. *Bioinformatics*, **23**, 809–814.

Durbin,R. *et al.* (1998) *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge University Press, UK.

Finn,R.D. *et al.* (2010) The Pfam protein families database. *Nucleic Acids Res.*, **38**, D211–D222.

Gibrat,J.F. *et al*. (1996) Surprising similarities in structure comparison. *Curr. Opin. Struct. Biol.*, **6**, 377–385.

Gonzalez,M.W. and Pearson,W.R. (2010) Homologous Over-extension: a challenge for iterative similarity searches. *Nucleic Acids Res.*, **38**, 2177–2189.

Kann,M.G. *et al.* (2007) The identification of complete domains within protein sequences using accurate E-values for semi-global alignment. *Nucleic Acids Res.*, **35**, 4678–4685.

Pruitt,K.D. *et al.* (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **35**, D61–D65.

Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.

UniPort Consortium (2009) The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res.*, **38**, D142–D148.