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

Mileidy W. Gonzalez1,* and William R. Pearson2,∗

1Department of Biological Sciences, University of Maryland Baltimore County, 1000 Hilltop Circle, Baltimore, MD 21250 and 2Department of Biochemistry and Molecular Genetics, Jordan Hall Box 800733, 1540 Jefferson Park Ave., Charlottesville, VA 22906, USA

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 pairwise 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/rasta/PURS/gonzalez20a

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

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

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 homology 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) 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 bare 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 original Pfam annotation 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 homology 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 families 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 homology library 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 3394 families.

© The Author(s) 2010. Published by Oxford University Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.5), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
PF00346 domain for residues 295–537 on the Q8ZM0, SALTY sequence. RefProtDom also annotates PF00374 in the same region (390–535) because both domains share the same SCOP fold and superfamily classifications (e18.1). This structural annotation is also reported by SCOP (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 1694 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 homologs.

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 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