Streamlining the Conserved Domain Database: A Taxonomic Approach

Praveen F. Cherukuri[1,2], Aron Marchler-Bauer[2], Lewis Y. Geer[2], and Stephen H. Bryant[2]

Keywords: protein domain, taxonomy

1 Introduction

The rapid growth of sequence databases has elevated the need for computational annotation of protein models. Not surprisingly, a variety of protein annotation resources have emerged, some examples include Pfam [1], SMART [3] and COG [5]. The Conserved Domain Database (CDD) [4] imports these and other publicly available alignment model collections, and adds curated domain definitions. The incorporation of several different source databases has created redundancy, ranging from duplication to hierarchical parent-child relationships, which may be caused by differing levels of representation in source databases. In addition, a fairly large subset of domains in CDD describes lineage-specific protein families with narrow taxonomic coverage. We describe a taxonomic-filter approach which is exclusively directed to the removal of such domain models, as we intend to offer a resource aimed at annotating “ancient” conserved domains.

2 Estimating a Domain’s Age

We define a set of nodes in the taxonomic tree of life which cover all branches represented by a significant amount of sequence data. We pick these taxonomic nodes so that the presence of a protein or domain family in more than one node indicates a certain minimum age (unless caused by horizontal gene transfer). Focusing on cellular organisms only, the final list has 66 taxonomic nodes (ex: mammalia, alphaproteobacteria, etc). The number of taxonomic nodes covered by a domain family gives a rough indication of that domain’s age.

3 Taxonomy Filter

We count the number of preferred taxonomic nodes a conserved domain detects in the NCBI NR (Non-Redundant) protein database, using pre-calculated RPS-BLAST results stored in the CDART database [2]. We recognized 1319 out of 13436 protein domains (~10%) in CDD version 1.63 to be specific to only one of the 66 taxonomic nodes. The majority of these protein domains originate from Pfam (8.5%), followed by COG (1.0%) and other databases (~0.5%). We investigate whether the presence of low-complexity regions or low sequence diversity in the domain model alignments correlates with narrow taxonomic distribution, and whether apparent narrow taxonomic distribution may be caused by bad performance of the search model.

4 Figures and tables.

Table 1: Analysis of protein domains with low taxonomic coverage

[1] Bioinformatics Program, Boston University, 44 Cummington St., Boston, MA 02215, USA. E-mail: cherukur@ncbi.nlm.nih.gov

[2] NCBI, National Institutes of Health, Bldg. 38A, 8600 Rockville Pike, Bethesda, MD 20894, USA. E-mail: bryant@ncbi.nlm.nih.gov

Source Database	Total number of protein domains	Protein Domains with Low Taxonomic Coverage	
		Number	%
Pfam	5426	1146	21.12
COG	4099	129	3.15
SMART	642	28	4.36
Cd	347	13	3.75
LOAD	53	0	0.00
KOG	2869	3	0.10
CDD [V1.63]	13436	1319	9.82

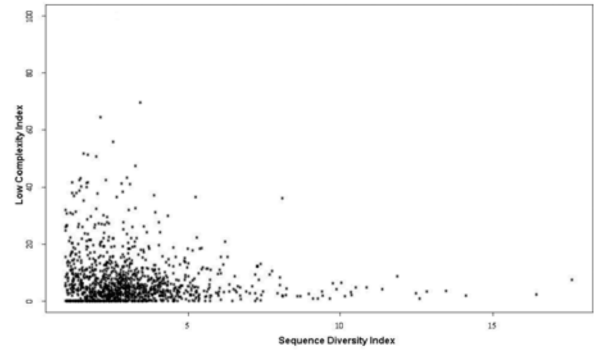

Figure 1: Sequence Diversity Index vs. Low complexity Index of protein domains with only one preferred taxonomic node hit

References

[[1] Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L., Studholme, D. J., Yeats, C. and Eddy, S. R. 2004. The pfam protein families database. Nucleic Acids Res 32:D138-41.

[2] Geer, L. Y., Domrachev, M., Lipman, D. J. and Bryant, S. H. 2002. Cdart: Protein homology by domain architecture. Genome Res 12:1619-23.

[3] Letunic, I., Goodstadt, L., Dickens, N. J., Doerks, T., Schultz, J., Mott, R., Ciccarelli, F., Copley, R. R., Ponting, C. P. and Bork, P. 2002. Recent improvements to the smart domain-based sequence annotation resource. Nucleic Acids Res 30:242-4.

[4] Marchler-Bauer, A., Anderson, J. B., DeWeese-Scott, C., Fedorova, N. D., Geer, L. Y., He, S., Hurwitz, D. I., Jackson, J. D., Jacobs, A. R., Lanczycki, C. J., Liebert, C. A., Liu, C., Madej, T., Marchler, G. H., Mazumder, R., Nikolskaya, A. N., Panchenko, A. R., Rao, B. S., Shoemaker, B. A., Simonyan, V., Song, J. S., Thiessen, P. A., Vasudevan, S., Wang, Y., Yamashita, R. A., Yin, J. J. and Bryant, S. H. 2003. Cdd: A curated entrez database of conserved domain alignments. Nucleic Acids Res 31:383-7.

[5] Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V., Krylov, D. M., Mazumder, R., Mekhedov, S. L., Nikolskaya, A. N., Rao, B. S., Smirnov, S., Sverdlov, A. V., Vasudevan, S., Wolf, Y. I., Yin, J. J. and Natale, D. A. 2003. The cog database: An updated version includes eukaryotes. BMC Bioinformatics 4:41.