Use of Text Mining for Protein Structure Prediction and Functional Annotation in Lack of Sequence Homology

Andreas Rechtsteiner*[1], Jeremy Luinstra[2], Luis M Rocha[3], Charlie E M Strauss[2]

[1]Center of Genomics and Bioinformatics, Indiana University, Bloomington, IN 47401
[2]Bioscience Division, Los Alamos National Lab, Los Alamos, NM 87545
[3]School of Informatics, Indiana University, Bloomington, IN 47401

Email: Rechtsteiner, A.*- andreas@cgb.indiana.edu; ; Rocha, LM - rocha@indiana.edu; Strauss, CEM - cems@lanl.gov;

*Corresponding author

Abstract

Background: Linking of information from different data sources, specifically literature, becomes increasingly important to annotate the growing number of new genome sequences. For the large percentage of genes with no known sequence homologs, new, possibly integrative, methods need to be developed. Ab-initio structure prediction and comparison is a method some of us pursued previously for functional annotation of sequences with no known homologs [1]. Here we use a large set of sequences of known structure to evaluate a new method that uses keyword information from literature to improve our previously used ab-initio structure prediction method.

Results: We report two results: first, the literature and keyword similarity measure we employ here performs well in identifying functional and/or structural relationships even if there is little or no sequence homology between the compared proteins, the difficult, but frequent, so-called "twilight zone" case in annotation and structure prediction. Second, our novel method that uses literature to assist SCOP super-family prediction [2] significantly improves on our original ab-initio structure prediction algorithm.

Conclusions: We show that the literature keywords and similarity measure used here are of great value for the increasingly important field of functional annotation of new sequences with no or little sequence homology.

Background

Each newly sequenced genome is principally annotated by comparison of its sequences to previously annotated genomes. Typically 40 to 60% of a new genome can be reliably annotated in this fashion. However, this method is most successful for the genes we often care least about, placing a premium on methods that can annotate unusual or highly diverged sequences. In this twilight recognition realm, ab-initio structure prediction based annotation has proven valuable [1]. By prediction of a protein's approximate structure we can compare it's structure to proteins of known function. Because this approach is less specific than sequence based annotation, it is useful to confirm ab-initio structure prediction based annotations by other means. Here we present a novel method that uses text mining to improve and screen genome-scale structure predictions in an automated fashion assisting further human curation efforts.

We generate a body of text for test sequences from sequence-based comparison to the non-redundant UniProt sequence database by selecting the literature associated with the top BLAST hits and extracting MeSH keywords [3]. Literature is obtained similarily for the member sequences of all structural SCOP super-families [2]. The literature for each super-family is combined. We then rank the super-families by decreasing cosine keyword vector similarity for each test sequence (details of this cosine measure can be found in [5,9]). The keyword based and ab-initio structure based rankings are further combined in a single ranking with the rank-product method [6]. We find that the annotation (i.e. SCOP super-family) rankings based on this approach dramatically improve the annotation accuracy over structure based annotation alone. This is demonstrated on a large benchmark set of sequences with known structures that is carefully screened for homolog removal to simulate highly diverged sequences.

MeSH terms were shown to be useful for extracting functional information about genes or proteins previously. For example, Masys et al. [7] showed that MeSH terms associated (through publications) with two clusters of co-expressed genes were informative about the medical conditions of the gene expression samples. MacCallum et al. [8] extracted keywords for proteins from the SwissProt/UniProt database and used the cosine similarity to improve remote homolog detection over using sequence similarity alone. They evaluated their method on a set of 100 known remote homologs. Only SwissProt keywords from the exact match of the remote homolog candidates in the protein database was considered, however.

The work presented here extends work we reported previously [4,5,9]. There we showed the power of the keyword similarity method to infer functional relationships for close sequence homologs, i.e. to predict protein families that are based on sequence homology. Here we go beyond that and show that sequence similarity is not required for the keyword similarity method to detect functional (and/or structural) relationships among proteins, its potential usefulness is therefore wider than supported by our previous results.

Data and Methods

400 non-redundant test sequences with known structure were selected at random from 320 (randomly selected) SCOP super-families from all 4 main SCOP classes. We predicted 1000 model structures for each sequence using repeated runs of the Rosetta algorithm [10]. For each set of 1000 models, these were clustered for structural similarity into typically 20 clusters. The model structures of the cluster centers were compared with MAMMOTH [11] to a non-redundant set of known SCOP structures (i.e. Astral40) which resulted in a ranking of SCOP super-families based on decreasing structure similarity as reported by MAMMOTH. Details on the prediction algorithms and methods can be found in [10,11].

We used BLAST to find the top matches for the 400 test sequences in the UniProt [12] protein database. We proceeded similarly for the non-redundant (Astral40) member sequences of all 1280 SCOP super-families (ver. 1.63), except that we removed SCOP sequences if they had a BLAST e-value smaller than 5 to one of the test sequences and both were members of the same SCOP super-family. The goal of this filter was to test for the difficult twilight cases expanded on above, where we do not have close sequence homology between a sequence we want to annotate and the member sequences of the correct super-family. Literature references for the BLAST hits are obtained from UniProt. MeSH keywords[1] for these references were extracted from PubMed/MEDLINE [3]. After frequency based keyword filtering and weighting, the cosine similarity measure between the keyword vectors of the test sequences and each of the pooled SCOP super-family vectors were calculated and used to rank the super-families by decreasing similarity. Further details can be found in [4,5,9].

To combine the independent structure prediction and keyword similarity based super-family rankings we used the non-parametric rank-product method [6]. For a given test sequence i and super-family j we have the structure based ranking $s_{i,j}$ and the keyword based ranking $k_{i,j}$. We obtain a rank product score for sequences i and super-families j by calculating $rp_{i,j} = s_{i,j} * k_{i,j}$ for all i and j. We then rank for each test

[1] Other sets of keywords are evaluate currently.

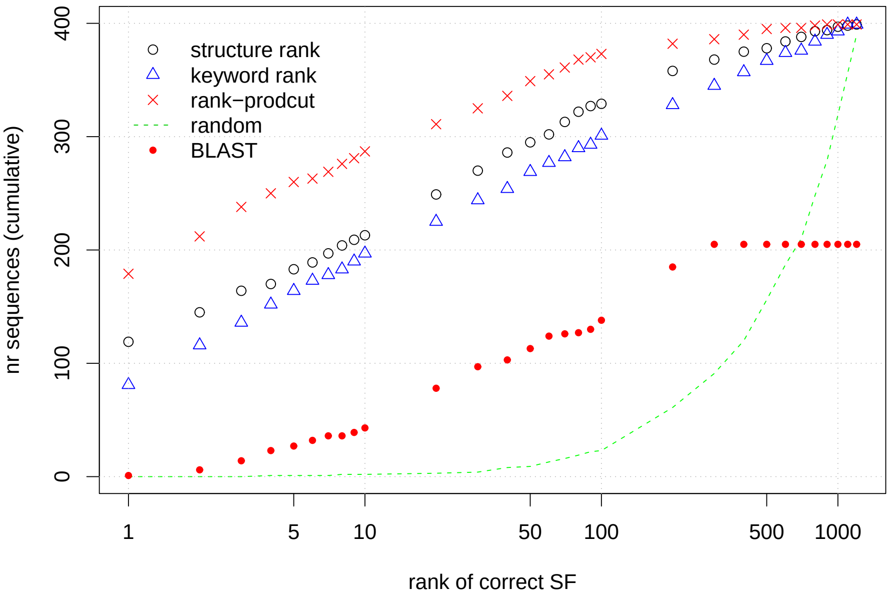

Figure 1: Ranking of correct SCOP super-families for 400 test sequences by the following methods: (i) BLAST e-value, (ii) keyword similarity, (iii) ab-initio structure prediction method, (iv) by combining the keyword and structure based rankings with the rank-product method, and (v), for comparison, picking and ranking super-families randomly. Our novel combined method performs significantly better than either the original structure prediction or keyword based prediction methods alone. The keyword method performs well even though the literature comes from sequences with little (BLAST) detectable sequence homology.

sequence i the SCOP super-families j by increasing $rp_{i,j}$ and obtain a new ranking of super-families based on the structure and the keyword ranking.

Results and Discussion

Figure 1 shows for our above explained 3 super-family prediction methods the number of sequences (y-axis) that had their correct super-family ranked in the position indicated by the x-axis or higher (i.e. better)[2]. The structure based ranking performs slightly better than the keyword similarity based method by itself. The combined rank-product method improves prediction considerably over both methods individually. For example, a typical set of structure predictions provided to a human curator are the top 10 predictions made by Rosetta and Mammoth. Among the top 10 predictions of the structure based method the correct super-families for 210 of the 400 sequences can be found. For the combined method, the correct super-families for 285 sequences can be found, an improvement of 35%.

Also shown are results if super-families are ranked based on sequence similarity, by increasing BLAST e-value. Sequence homologs to the test sequences had been removed thoroughly, as BLAST was not able to rank the correct super-family at the top for a single test sequence[3]. This result shows very clearly the

[2] Note that the number of sequences on the y-axis is cumulative, e.g. the number of sequences shown for rank 5 includes all sequences that had their super-family ranked at position 5 or higher.

[3] The leveling off of BLAST rankings after about 200 sequences is due to the e-value cutoff of 10,000 (the database size was

strength of the keyword similarity method, as it is able to correctly indicate functional (and/or structural) similarity between proteins with no detectable sequence homology.

Conclusions

We show that literature keyword similarity measures can infer functional and structural relationships among proteins even if there is no, or very little, sequence homology among the respective protein sequences, an ability searched for by the community to make predictions in the difficult twilight zone. We were able to show further that our combined method of predicting structural super-families with ab-initio structure prediction performs significantly better than either method individually. Current and future work focuses on different keyword sets, like abstract keywords, and obtaining confidence measures for the keyword based predictions.

In conclusion, our results are encouraging to further pursue text mining for Bioinformatics for the challenging tasks at hand and to search for ways to link different methods and sources of information.

Acknowledgements

Judith Cohn set up the internal Protein Function Inference Group (PFIG) database at LANL which was very valuable to the tasks at hand. She helped further in obtaining data from the database and provided many useful comments. Lars Malmström assisted greatly in obtaining the ab-initio structure predictions for the test sequences. The CGB at Indiana University as well as Los Alamos National Lab, specifically the PFIG grant, funded the work by the authors.

References

1. Bonneau R, Strauss C, Rohl C, Chivian D, Bradley P, Malmstrom L, Robertson T: **De novo prediction of three-dimensional structures for major protein families**. *Journal of Molecular Biology* 2002, **322**:65–78.
2. Murzin AG, Brenner SE, Hubbard T, Chothia C: **SCOP: a structural classification of proteins database for the investigation of sequences and structures.** *J. Mol. Biol.* 1995, **247**:536–540.
3. National Library of Medicine: **PubMed**. *http://www.ncbi.nlm.nih.gov/entrez/* 2006.
4. Rechtsteiner A, Rocha L, Strauss C: **Clustering of Protein Families in literature keyword space.** In *Currents in Computational Molecular Biology (RECOMB 2005)*, Boston, MA 2005.
5. Maguitman AG, Rechtsteiner A, Verspoor K, Strauss CE, Rocha LM: **Large-Scale Testing of Bibliome Informatics Using Pfam Protein Families**. In *Pacific Symposium on Biocomputing*, *Volume 11* 2006:76–87.
6. Breitling R, Armengaud P, Amtmann A, Herzyk P: **Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments.** *FEBS Lett.* 2004, **573**:83–92.
7. Masys D, Welsh J, Lynn Fink J, Gribskov M, Klacansky I, Corbeil J: **Use of keyword hierarchies to interpret gene expression patterns.** *Bioinformatics* 2001, **17**(4):319–26.
8. MacCallum R, Kelley L, Sternberg M: **SAWTED: structure assignment with text description–enhanced detection of remote homologues with automated SWISS-PROT annotation comp arisons**. *Bioinformatics* 2000, **16**(2):125–129.
9. Rechtsteiner A: **Multivariate Analysis Of Gene Expression Data And Functional Information: Automated Methods For Functional Genomics**. *PhD thesis*, Portland State University 2005.
10. Rohl C, Strauss C, Misura K, Baker D: **Protein structure prediction using Rosetta**. *Numerical Computer Methods* 2004, **383**:66+.
11. Ortiz A, Strauss C, Olmea O: **MAMMOTH (Matching molecular models obtained from theory): An automated method for model comparison**. *Protein Science* 2002, **11**:2606–2621.
12. SIB/EBI: **UniProt/Swiss-Prot**. *http://www.ebi.ac.uk/swissprot/* 2004.

10,000 sequences)