

# Clustering of Protein Families in Literature Space

Andreas Rechtsteiner<sup>1</sup>, Luis M Rocha<sup>2</sup>, Charlie E Strauss<sup>3</sup>

**Keywords:** knowledge and information mining, vector space model, MeSH, protein families

## 1 Motivation

Automated mining of biological information from literature and databases is a subject of increasing interest. Different studies have applied different techniques to mostly very specific problem domains [4, 2, 3]. What has mostly been missing in the field are large-scale studies that allow for quantitative validation and a gold standard defining an effective basis for method comparison. Here we perform such a large scale study for the vector space model of Information Retrieval (IR) [1] and the Medical Subject Heading (MeSH) vocabulary [5] used by the National Library of Medicine to index publications in the PubMed/MEDLINE database. We specifically test how well we can predict the Pfam protein sequence family of a protein purely based on the MeSH terms of the publications referenced for that protein in the protein sequence database UniProt/SwissProt [6]. The Pfam family classification presents a suitable test case for information mining algorithms as its many (thousands) of families are based on the physical property of protein sequence and are largely congruent with functional classes, whose identification is often at the center of information mining in Bioinformatics.

## 2 Data and Methods

26,411 publication references in the UniProt/SwissProt protein database were identified for 15,217 proteins from 1611 Pfam families. The family sizes range between 3 and 20 protein members. The MeSH terms that are used to index the 26,411 publications were retrieved from PubMed/MEDLINE. Similar to the vector space model in IR which represents documents in keyword space, we represent proteins in MeSH space (based on the publications). We applied a weighting of the MeSH terms, a linear scaling of the MeSH space dimensions, which we call 'Inverse Pfam Frequency' (for its relatedness to Inverse Document Frequency (IDF) in IR). An unsupervised (k-nearest neighbor related) classification algorithm was used to rank for a given protein the Pfam families based on the number of protein members they have in a given cosine neighborhood in MeSH space. The top ranked Pfam family has most members in the neighborhood of the given protein, the second ranked the second most etc.

## 3 Results and Discussion

For 47% of proteins, the first ranked Pfam family was the correct Pfam family and for 70% of the proteins (an additional 23%) the correct Pfam family was within the first 5 ranked families. Almost 80% of proteins have their correct Pfam family ranked within the first 10 Pfam families. With a uniform family size distribution, we would have expected 1 in 1611 proteins, or 0.06%, to be predicted correctly by chance. Our prediction rate is more than 700 times higher. The fact that we use an unsupervised classification algorithm insures that this success rate is not achieved by 'over-learning'.

The fact that many proteins have their Pfam family ranked within the first few predicted families, if not the first, suggests that mispredictions occur to related families. Human curation confirmed that

---

<sup>1</sup>Bioscience Division, Los Alamos National Lab, Los Alamos NM, 87545. E-mail: andreas@lanl.gov

<sup>2</sup>School of Informatics and Cognitive Science Program, Indiana University, 1900 East Tenth Street, Bloomington IN 47406. E-mail: rocha@indiana.edu

<sup>3</sup>Bioscience Division, Los Alamos National Lab, Los Alamos NM, 87545. E-mail: cems@lanl.gov

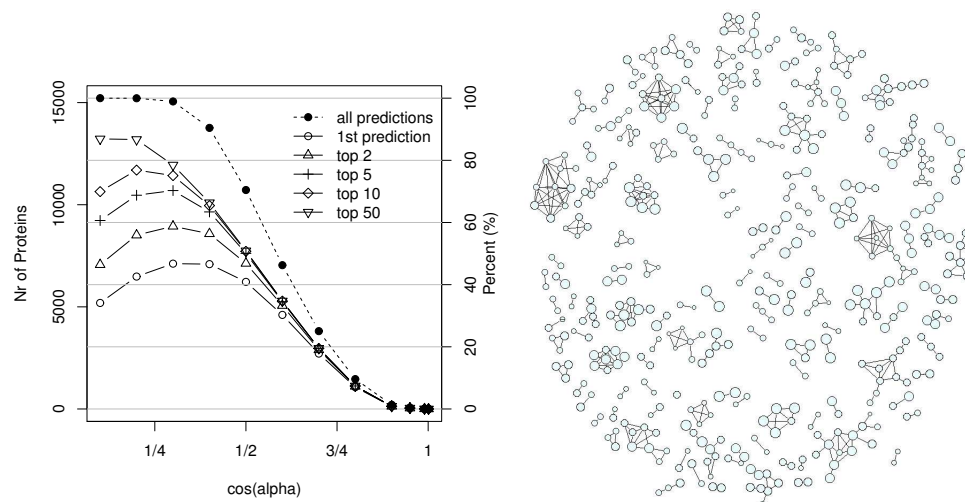


Figure 1: On the left, prediction success for the 15,217 proteins from 1611 Pfam families. At  $\cos(\alpha) = 0.3$ , 47% of the proteins have their Pfam family predicted correctly by the top ranked family. For 70% of the proteins the correct Pfam family is ranked among the first 5 families. On the right is a graph showing that most mispredictions occur between relatively small cliques of Pfam families. Family members of such cliques tend to be related.

indeed most mispredictions occur within cliques of related families. Examples of such cliques are families with related enzymatic functions or families that are subunits of larger protein complexes. We found further that very few and very specific MeSH terms are highly associated with a Pfam family, indicating high consistency in the indexing process of NLM for MEDLINE.

Our study showed that the indexing vocabulary MeSH and the indexing process employed by the National Library of Medicine does capture a significant amount of information about the sequence family of a protein and its biological function. Further studies will use these results as a baseline to which different approaches, e.g. vocabularies and algorithms, can be compared to.

## References

- [1] Ricardo Baeza-Yates, Berthier Ribiero-Neto, and Berthier Ribeiro-Neto. *Modern Information Retrieval*. Pearson Education, 1999.
- [2] T.K. Jensen, A. Laegreid, J. Komorowski, and E. Hovig. A literature network of human genes for high-throughput analysis of gene expression. *Nat. Genet.*, 28(1):21–28, 2001.
- [3] RM MacCallum, LA Kelley, and MJE Sternberg. SAWTED: structure assignment with text description-enhanced detection of remote homologues with automated SWISS-PROT annotation comparisons. *Bioinformatics*, 16(2):125–129, Feb 2000.
- [4] D.R. Masys, J.B. Welsh, J. Lynn Fink, M. Gribskov, I. Klacansky, and J. Corbeil. Use of keyword hierarchies to interpret gene expression patterns. *Bioinformatics*, 17(4):319–26, 2001.
- [5] National Library of Medicine. MeSH Fact Sheet. <http://www.nlm.nih.gov/pubs/factsheets/mesh.html>, 2004.
- [6] SIB/EBI. UniProt/Swiss-Prot. <http://www.ebi.ac.uk/swissprot/>, 2004.