Protein Association Discovery in Biomedical Literature

Yueyu Fu, Javed Mostafa, and Kazuhiro Seki
Laboratory of Applied Informatics Research, Indiana University, Bloomington
E-mail: {yufu, jm, kseki}@indiana.edu

Abstract

Protein association discovery can directly contribute toward developing protein pathways; hence it is a significant problem in bioinformatics. LUCAS (Library of User-Oriented Concepts for Access Services) was designed to automatically extract and determine associations among proteins from biomedical literature. Such a tool has notable potential to automate database construction in biomedicine, instead of relying on experts' analysis. This paper reports on the mechanisms for automatically generating clusters of proteins. A formal evaluation of the system, based on a subset of 2000 MEDLINE titles and abstracts, has been conducted against Swiss-Prot database in which the associations among concepts are entered by experts manually.

1. Introduction

There is a huge corpus of biomedical literature available electronically, e.g. the MEDLINE database. The complex medical concept relations in this literature are highly valuable. Unfortunately, there are few comprehensive sources of information on biomedicine that explicitly capture and record such associations.

Researchers have spent much effort developing systems to automatically mine biomedical literature [3, 5, 7, 9]. Early efforts applied Natural language processing (NLP) techniques. More recent efforts concentrate on combining NLP techniques with statistical techniques developed in IR.

In this paper, we discuss a project aimed at automated database construction in biomedicine. LUCAS was developed to automatically discover associations among proteins from biomedical literature. Various strategies, both linguistic and statistical were used in the information extraction and retrieval process.

2. Discovery algorithms

2.1. Protein discovery

Protein names are detected in two steps: protein name fragment detection and name boundary expansion of the detected fragments. In the former step, protein name fragments are detected by hand-crafted rules based on surface clues, which include Arabic numerals, Roman numerals and alphabets, and some suffixes and words peculiar to protein names (e.g., -in, -ase, and factor). As some protein names are compound nouns (e.g., parathyroid hormone-related protein), protein name boundaries of the detected fragments are expanded to recognize full protein names. Then, a protein name dictionary that does not include proteins covered in the rule set is applied to detect additional protein [6]. Finally, $tf \cdot idf$ weight [8] is computed for each unique protein found in individual documents and a list of proteins are extracted based on two user selected parameters, namely rank/document and document frequency of proteins.

2.2. Utilizing latent semantic information

To improve the performance of the protein association discovery (described in the next section), we wanted to enhance the information in the protein-doc matrix. A process known as Latent Semantic Analysis (LSA)[1, 2], to reduce the rank of the matrix, has been shown to enhance document vectors by using latent semantic structure in the vectors to help eliminate noise and deal with co-occurrence of proteins. The protein-doc matrix produced using $tf \cdot idf$ is rank reduced according to singular value decomposition. The resulting vectors help to make the implicit latent semantic information in the protein-doc matrix explicit.

2.3. Protein Association Discovery

Protein association discovery mainly consists of: an unsupervised cluster learning stage and a vector classification stage. During the learning stage, initial cluster hypotheses $[C^1, \ldots, C^k]$ are generated from a representative sample of protein vectors $[S^1, \ldots, S^N]$. Each cluster C^i is then represented by its centroid, Z^i. During the classification stage, an incoming protein V^i is classified into a particular class C^k using the learned centroids from the first stage.

A heuristic unsupervised clustering algorithm, called the Maximin-Distance algorithm [10], is used to determine the centroids. In this iterative algorithm, at

Proceedings of the 2003 Joint Conference on Digital Libraries (JCDL'03)
0-7695-1939-3/03 $17.00 © 2003 IEEE

each stage, a protein vector is selected that has the least similarity with the existing centroids. The similarity of this protein vector with the existing set of centroids, in turn, is the maximum of its similarities over all centroids. The selected point is then added as a new centroid if and only if its similarity with the existing set of centroids is less than an implementation-specified threshold parameter. This process is continued until no new centroids can be identified. During classification each new protein vector is classified to one of the centroids that has the largest similarity with the protein. We refer the reader to [4] for further details.

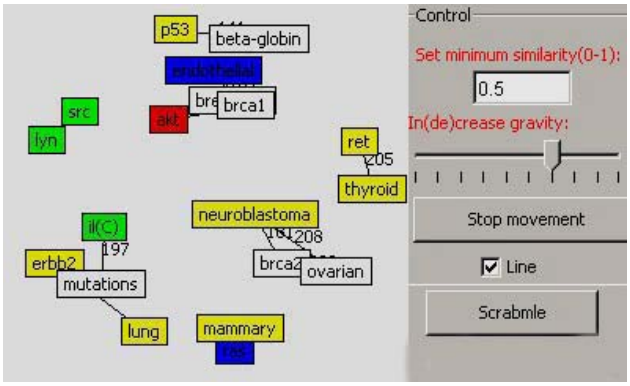

Figure 1. LUCAS interface displaying concept visualization

3. Experiments

An interactive web-based association discovery system called LUCAS was implemented (see Figure 1) to aid biomedical researchers to identify useful links among key concepts. Experiments were conducted to see how well our system could be used to discover relationships among proteins from biomedical literature. The test set was 2000 MEDLINE titles and abstracts from the GENIA corpus3 containing human annotated protein names (www-tsujii.is.s.u-tokyo.ac.jp/GENIA). 75.5% of the proteins in this corpus were detected accurately. Clusters were generated from this set using the methodology outlined above.

The particular rank to select when LSA is used to improve information discovery depends on the domain and the corpus (i.e., it is an empirical problem). Hence, we conducted a series of experiments to examine the impact of varying the LSA rank on cluster overlap. Two proteins are said to overlap if they co-occur in an entry (record) in the Swiss-Prot protein database. The average largest overlap in all the clusters returned from the Maximin clustering was computed. However, since varying parameters in our experiments also varies the number and sizes of clusters, it is not enough to measure only the average size of the overlap. A very large cluster could produce a large overlap, but also contain many proteins that are not related. To account for this, the key result we were interested in is the average ratio of the largest overlap to cluster size across all clusters for a given test. A ratio of 1.0 would mean that all of the proteins in each cluster are related according to the protein database. The best overlap ratio as shown in Figure 2 was 0.82 when rank equaled 5. Another larger experiment, which contained 247 protein names, produced a similar trend of the curve in the larger experiment. The best result was 0.90 when rank equaled 10. As seen in the plot with increasing rank more "noise' is reintroduced in the protein-doc matrix and hence overlap drops.

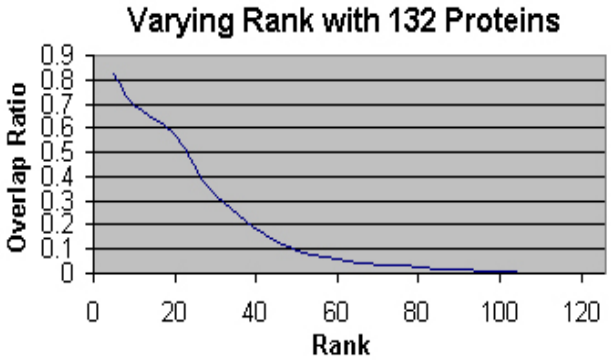

Figure 2. Impact of varying LSA rank on associations

4. Conclusions

In this paper, we investigated the effectiveness of the implemented algorithms in identifying protein association. A general finding was that the implemented algorithms are stable, robust, and are capable of providing useful results. A more specific finding was that LSA can yield successful results for extracting relevant associations among proteins with appropriate selection of parameter value.

5. Acknowledgement

This research was partially supported by a NSF ITR grant #9817572.

6. References

[1] Berry, M., Dumais, S., and Letsche, T. "Computational methods for intelligent information access". *In Proceedings of Supercomputing '95*, San Diego, CA, 1995.

[2] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. "Indexing by latent semantic analysis". *Journal of the American Society for Information Science*, 1990, 41:391–407.

[3] Eriksson, G., Franzen, K., Olsson, F., Asker, L., and Liden, P. "Exploiting syntax when detecting protein names in text". *In Proceedings of Workshop on Natural Language Processing in Biomedical Applications*, 2002.

[4] Fu, Y., Bauer, T., Mostafa, J., Palakal, M., & Mukhopadhyay, S. "Concept extraction and association from cancer literature". *In Proceedings of the fourth international workshop on Web information and data management of ACM CIKM*, 2002, 100-103.

[5] Fukuda, K., Tsunoda, T., Tamura, A., and Takagi, T. "Toward information extraction: identifying protein names from biological papers". *In Pac Symp Biocomput*, 1998, 707–18.

[6] Kazuhiro, S., and Mostafa, J. "An approach to protein name extraction using heuristics and a dictionary". Laboratory of Applied Information Research Tech Report 2003-2. Indiana University, Bloomington, IN, USA, 2003.

[7] Rindflesch, T., Rajan, J., and Hunter, L. "Extracting molecular binding relationships from biomedical text". *In Proceedings of the 6th Applied Natural Language Processing Conference*, 2000, 188–195.

[8] Salton, G. *Introduction to modern information retrieval*. McGraw-Hill, New York, 1983.

[9] Thomas, J., Milward, D., Ouzounis, C., Pulman, S., and Carroll, M. "Automatic extraction of protein interactions from scientific abstracts". *In Pac Symp Biocomput*, 2000, 541–52.

[10] Tou, J. T. and Gonzalez, R. C. *Pattern Recognition Principles*. Addision-Wcslcy, 1974.