

Learning Pathogenic Proteins across Fractured and Heterogeneous Data

Eithon Cadag¹, Peter Tarczy-Hornoch, MD^{1,2,3}, Peter J. Myler, PhD^{1,4,5}
Biomedical Informatics¹, Pediatrics², Comp. Sci. & Engr.³ and Pathobiology⁴, University of
Washington, Seattle, WA; Seattle Biomedical Research Institute⁵, Seattle, WA

Abstract

In the following work, we test a generalized approach to integrating, transforming and learning data from disparate data sources for the classification of bacterial proteins involved in pathogenesis. We rely on the implicit inter-linkages between biological databases to draw relevant records, and leverage statistical learning methods to infer classification based on abundant, albeit noisy, data. Results suggest that types of public biological information have varying degrees of effectiveness in predictive data mining.

Introduction

Emerging infectious diseases have moved to the forefront of research and global health in the face of antibiotic-resistant pathogens, epidemics and the threat of bioterrorism. Our need to understand mechanisms of pathogenicity has magnified, yet the depth to which pathogenic proteins are characterized often does not go beyond basic similarity searches¹. Though triangulating data across many databases is a common practice among biologists, noise in biological repositories make it hard to reach a conclusion towards protein pathogenicity. To address this, we propose using linked, integrated data across multiple databases as a 'query space' upon which we can classify a protein. When trained against pathogenic proteins, this method facilitates the identification and annotation of pathogenic proteins while using only inexpensive and publicly-available biological data sources.

Methods

To retrieve large amounts of relevant results for a query protein, we used an existing data integration system (BioMediator²) to query a dozen different data sources. Upon seeding with a query protein of known class, data sources that take as input protein sequences return results linked to our query; these results are linked to other sources, which return more data. The query space becomes a large amount of data from a number of databases, each bit of which may be of varying relevance to our query. To make sense of the information, we map the data onto a feature space that allows us to treat each query as a labeled vector, and thus each set of queries as a matrix. This generated feature space can then be trained upon using statistical learning methods.

To evaluate this method, we trained support vector³ classifiers on 100 pathogenic proteins and 100 proteins randomly selected from GenBank. We also evaluated our approach on a specific virulence cate-

gory, pathogen-related polysaccharide proteins, by training on 50 positive and 20 negative examples. We created three feature spaces for training on these sets: 3-mer sequence windows that provide a baseline classification independent of data sources; *bag-of-words* frequency from integrated data; and GO terms that appeared in our query space more than once. Different kernels were used to train three separate classifiers for each feature space: a linear kernel; a radial basis function (RBF) kernel with high regularization cost; and an RBF whose cost and width were determined from a constrained grid search and whose features were selected using F-scores.

Results

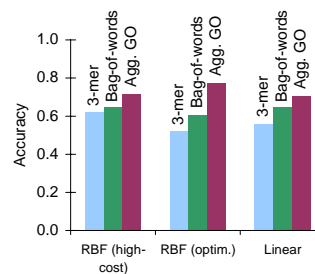


Figure: Accuracy of classification by type and kernel for general pathogenicity classification (polysaccharide results not shown).

We evaluated pathogenicity and polysaccharide prediction on 100 positive, 100 negative and 20 positive, 20 negative examples, respectively. Performance was better when specific pathogenic categories were trained, especially for bag-of-words (top accuracy, 100%). With general pathogenicity prediction, aggregated GO categories have the best performance (top accuracy, 77.5%). In both the general and specific cases integrated approaches outperformed the sequence baseline, which suggests that even a mixture of relevant and spurious data records, mined appropriately, provide a high-level predictive layer at determining pathogenicity.

Acknowledgements

This work is funded through NLM #T15 LM07442.

References

1. Raskin D., et al. Bacterial genomics and pathogen evolution. *Cell* 2006;124:703-714.
2. Donelson L., et al. The BioMediator system as a data integration tool to answer diverse biologic queries. *Proceedings of MedInfo, IMIA*, 2004.
3. Boser B., Guyon, I., Vapnik, V. A training algorithm for optimal margin classifiers. In: *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*; 1992 July 1992; Pittsburgh, PA; 1992. p. 144-152.