

Student Research Symposium 2003

Yee Man (Betty) Cheng

Identifying Important Words in the Language of Proteins

Identification of keywords is an important task in language technologies research. For example, keyword extraction is needed in question answering and summarization. A task-oriented dialog system may recognize keywords from a lexicon specific to the task instead of entire utterances in the speech.

In the language of proteins, “keywords” are termed “motifs”, short amino acid sequences that are conserved across a family or subfamily of proteins because they are the binding sites for protein-protein interactions typical of that family or subfamily. Identifying motifs is an important task in bioinformatics because they aid in large-scale protein-protein interaction prediction and new drug design. However, like in many Asian languages, there are no word boundaries in protein sequences, making the task more difficult. Moreover, unlike human languages, we have yet to build a lexicon for the protein language.

G-protein coupled receptors (GPCR) comprise one of the largest superfamily of proteins found in the body (Gether, 2000), and are the target of approximately 60% of current drugs on the market (Muller, 2000). They are also one of the most challenging datasets in protein classification due to the extreme diversity among its members (Moriyama and Kim, 2003). The GPCR superfamily is organized hierarchically in various levels of subfamilies. Karchin et al. (2002) tested a set of classifiers of varying complexity from k-NN to SVM in GPCR classification at the superfamily and subfamily levels, and showed that the more complex classifier SVM performed better than other classifiers at the subfamily level classification. Here, we show that by choosing the right features, n-gram counts selected by chi-square, a feature selection method successful in document classification (Yang and Pedersen, 1997), the simpler Naïve Bayes classifier can outperform the SVM. In addition, the selected n-grams appear to have biological significance, since they correlate with motifs previously identified through wet-lab experiments.

References

- U. Gether. Uncovering Molecular Mechanisms Involved in Activation of G Protein-Coupled Receptors. *Endocrine Reviews*, 21(1):90-113, 2000.
- R. Karchin, K. Karplus, and D. Haussler. Classifying G-protein coupled receptors with support vector machines. *Bioinformatics*, 18(1):147-159, 2002.
- E. N. Moriyama and J. Kim. Protein Family Classification with Discriminant Function

Analysis. In Proceedings of Stadler Genetics Symposium, 2003.

G. Muller. Towards 3D structures of G protein-coupled receptors: a multidisciplinary approach. *Current Medical Chemistry*, 7(9):861-888, 2000.

Y. Yang and J. Pedersen. A Comparative Study on Feature Selection in Text Categorization. In Proceedings of 14th International Conference on Machine Learning, pages 412-420, Nashville, US, 1997.