

## Student Research Symposium 2004

---

Yee Man (Betty) Cheng

### Language Technologist's Approach to Understanding G-Protein-GPCR Interaction

String alignments and n-grams are commonly used in language technology applications, such as machine translation, information retrieval, speech recognition and synthesis. In machine translation, alignment can yield high accuracy if the source and target languages have similar word order. However, if the two languages have very different word order, getting a correct alignment can be difficult and an n-gram based MT system may perform better. Likewise, a correct alignment of protein sequences can yield high accuracy in prediction problems. But segments or "words" in the protein sequence can shuffle in their linear order while preserving their orientation in 3D space and therefore the protein's function or "meaning" as well.

The superfamily of proteins in this study, G-protein coupled receptors (GPCR), are important in pharmacological research as they are the target of approximately 60% of current drugs on the market (Muller, 2000). Coupling with G-proteins, these receptors regulate much of the cell's reactions to external stimuli. Abnormalities in this regulation can lead to cancer, Alzheimer's, Parkinson's and other diseases. Identification of the type of G-proteins that can bind to a particular GPCR can provide information on the causes and symptoms of the disease the receptor is involved in.

Previous studies on predicting the family of G-proteins that can couple to a given GPCR sequence have focused on the intracellular domains of the receptor sequence, either using alignment-based features (Cao et al., 2003; Qian et al., 2003), n-gram features (Moller et al., 2001) or physiochemical properties of the amino acids (Henriksson, 2003). From the roles of alignments and n-grams in MT and their analogy to the protein language, we have chosen to combine alignment and n-gram information in a hybrid prediction method using a k-nearest neighbours (k-NN) classifier on sequence alignment similarity and a k-NN classifier on Euclidean distance of n-gram counts. Our method outperforms the current state-of-the-art in precision, recall and F1. Systematic experiments with our prediction method were able to validate biologists' hypothesis that most of the coupling specificity information resides in the 2nd and 3rd intracellular loops of the receptor, while providing evidence for a new hypothesis that the information is more localized to the beginning of the 2nd intracellular loop.

Cao, J., R. Panetta, et al. (2003). "A naive Bayes model to predict coupling between seven transmembrane domain receptors and G-proteins." *Bioinformatics* 19(2): 234-40.

Henriksson, A. (2003). Prediction of G-protein Coupling of GPCRs - A Chemometric Approach. Engineering Biology. Linkoping, Linkoping University: 79.

Moller, S., J. Vilo, et al. (2001). "Prediction of the coupling specificity of G protein coupled receptors to their G proteins." Bioinformatics 17 Suppl 1: S174-81.

Muller, G. (2000). "Towards 3D structures of G protein-coupled receptors: a multidisciplinary approach." Curr Med Chem 7(9): 861-88.

Qian, B., O. S. Soyer, et al. (2003). "Depicting a protein's two faces: GPCR classification by phylogenetic tree-based HMMs." FEBS Lett 554(1-2): 95-9.