Carnegie Mellon University
Research Showcase @ CMU

Computer Science Department | School of Computer Science

1-2007

Combining N-grams and Alignment in G-Protein Coupling Specificity Prediction

Betty Yee Man Cheng
Carnegie Mellon University

Jaime G. Carbonell
Carnegie Mellon University, jgc@cs.cmu.edu

Follow this and additional works at: http://repository.cmu.edu/compsci

Recommended Citation
.

This Conference Proceeding is brought to you for free and open access by the School of Computer Science at Research Showcase @ CMU. It has been accepted for inclusion in Computer Science Department by an authorized administrator of Research Showcase @ CMU. For more information, please contact research-showcase@andrew.cmu.edu.

COMBINING N-GRAMS AND ALIGNMENT IN G-PROTEIN COUPLING SPECIFICITY PREDICTION

BETTY YEE MAN CHENG[†]

Language Technologies Institute, School of Computer Science, Carnegie Mellon University, 5000 Forbes Ave., Pittsburgh, PA 15213, USA

JAIME G. CARBONELL

Language Technologies Institute, School of Computer Science, Carnegie Mellon University, 5000 Forbes Ave., Pittsburgh, PA 15213, USA

G-protein coupled receptors (GPCR) interact with G-proteins to regulate much of the cell's response to external stimuli; abnormalities in which cause numerous diseases. We developed a new method to predict the families of G-proteins with which it interacts, given its residue sequence. We combine both alignment and n-gram features. The former captures long-range interactions but assumes the linear ordering of conserved segments is preserved. The latter makes no such assumption but cannot capture long-range interactions. By combining alignment and n-gram features, and using the entire GPCR sequence (instead of intracellular regions alone, as was done by others), our method outperformed the current state-of-the-art in precision, recall and F1, attaining 0.753 in F1 and 0.796 in accuracy on the PTbase 2004 dataset. Moreover, analysis of our results shows that the majority of coupling specificity information lies in the beginning of the 2nd intracellular loop and over the length of the 3rd.

1 Introduction

G-protein coupled receptors (GPCR) are a diverse superfamily of proteins characterized by their structure of 7 transmembrane alpha helices separated by alternating intracellular and extracellular loops. They are responsible for signal transduction across the cell membrane and are the targets of 60% of all drugs[1]. Their extracellular domains are capable of recognizing a wide range of ligands such as ions, hormones and neurotransmitters. The binding of these ligands causes the receptors to change their conformation, particularly in their intracellular domains exposing sites critical for the subsequent coupling with specific G-proteins.

G-proteins consist of α-subunits bound to βγ complexes and are classified into 4 families by their α-subunits: $G_{i/o}$, $G_{q/11}$, G_s and $G_{12/13}$. G_s and $G_{i/o}$ activates and inhibits adenylyl cyclase respectively, while $G_{q/11}$ activates phospholipase C. The function of the last family, $G_{12/13}$, remains unknown. Through their coupling with G-proteins, GPCRs regulate the cell's response to external stimuli. Abnormalities in this regulation lead to numerous diseases.

To date, most of the known G-protein coupling specificity information has been obtained through experimental approaches, a survey of which can be found in [2]. An

[†] To whom correspondence should be addressed: ymcheng@cs.cmu.edu

accurate method to predict the G-protein families a given GPCR sequence can couple with is of immense value to pharmaceutical research for three reasons. First, the development of such a method can elucidate the physiological mechanisms underlying the response mediated by a GPCR in diseases. Second, the coupling specificity of a GPCR is needed to identify its activating ligands because the appropriate G-protein needs to be present in the cell while one passes potential ligands (tissue extracts or libraries of chemical compounds) over the cells and watches for the suitable response. Finally, information learned in a study on coupling specificity prediction is likely to be applicable to the more general problem of protein-protein interaction prediction as well.

To the best of our knowledge, there have been 4 previous studies on predicting G-protein coupling specificity from the receptor sequence, although only one of them[3] considered and validated their approach on receptors coupling to multiple families of G-proteins, as we do in our study. Each of these studies focused on the intracellular domains of the receptor, using either alignment information[3, 4], n-grams[5] or physiochemical properties of the amino acids[6]. Alignment-oriented approaches have been very popular in computational biology. They utilize biological domain knowledge via the use of amino acid similarity matrices and account for some long-range interactions but they have inherent limitations due to their assumption that contiguity of homologous segments is conserved[7]. This assumption contradicts the genetic reshuffling and recombination that occurs through evolution[8, 9], and as a result, sequence alignments become unreliable when sequences have less than 40% similarity[10] and are unusable below 20% similarity[11, 12].

Moreover, since protein-protein interactions occur in 3-d space, only the orientation of the motifs for coupling specificity need to be conserved for the interaction to occur, and not the ordering of the motifs in the linear sequence. N-grams have the potential to capture the presence and absence of coupling specificity motifs (but not their 3-d orientation) without imposing a restriction on their ordering in the primary sequence. However, since the dimension of the n-gram feature space increases exponentially with the length of the n-gram, n-grams tend to be short and do not account for long-range interactions. Hence, with the complimentary pros and cons of sequence alignment and n-grams in mind, we developed a new G-protein coupling specificity prediction method that uses both sequence alignment and n-grams.

All previous studies have used only the intracellular domains, ignoring the extracellular and transmembrane domains as those do not make physical contact with G-proteins. While it is intuitive that the majority of coupling specificity information would lie in the intracellular regions, our preliminary results showed that including the extracellular and transmembrane regions can improve the prediction accuracy because the predicted transmembrane boundary positions are not entirely accurate and there may be relayed effects in these regions from the GPCR becoming activated by the ligand, as shown in our preliminary experiments (data not shown). Thus, our approach made use of whole sequence alignment and n-grams extracted from the whole sequence. In addition,

we explored which areas within the intracellular domains contain the most discriminative coupling specificity information.

2 Methods

To address prediction of coupling to multiple families of G-proteins, we defined the coupling specificity problem as a set of 3 binary classification problems, one for each G-protein family to determine whether the given receptor couples to proteins from the family. We developed an n-gram based prediction module and an alignment based prediction module, each outputting a probability of the GPCR coupling to the particular G-protein family. Each module can be used independently to make a prediction by setting a probability threshold above which a coupling is predicted to occur. Alternatively, the modules can be combined, as we have in our hybrid prediction method, to utilize both n-gram and alignment information in making the prediction.

2.1 N-gram Based Module

The approach in our n-gram based module is analogous to the bag-of-words approach to document classification in the language technologies domain and has been successfully applied to GPCR family and subfamily classification[13]. Each GPCR sequence is represented as a vector of n-gram counts where unigrams, bigrams, trigrams and tetra-grams are extracted at each reading frame from the whole sequence. Preliminary results showed n-grams from the whole sequence yielded more accurate predictions than n-grams from the intracellular domains alone (data not shown). We used a 21-letter alphabet, 20 for the known amino acids and 1 for amino acid x, giving us a vector length of 204204.

A high dimensional feature space can confuse a classifier with irrelevant features. By using only features that are informative to the task, we can optimize our prediction accuracy while reducing the running time. Various feature selection methods have been developed for this purpose in the machine learning field, such as information gain, mutual information and chi-square. For each G-protein family C, we employed chi-square feature selection (section 2.1.1) to derive the *p* most discriminative binary features from n-gram counts in differentiating between receptors that can and those that cannot couple with proteins in C. The feature vectors were then converted to vectors containing only those *p* features and a k-nearest neighbors classifier (k-NN) was applied on them.

To predict whether a given sequence d_i couples to proteins from G-protein family C, the k-NN classifier finds the *k* "closest" sequences to d_i as defined by the normalized Euclidean distance between their feature vectors and computes their majority vote weighted by their inverse distance to d_i. The normalized Euclidean distance between 2 vectors $\vec{u}$ and $\vec{v}$ is defined as follows:

$$d(\bar{u},\bar{v}) = \left[\sum_i (\tilde{u}_i - \tilde{v}_i)^2\right]^{1/2} \text{ where } \tilde{u}_i = \frac{u_i - \min(i)}{\max(i) - \min(i)},\ \tilde{v}_i = \frac{v_i - \min(i)}{\max(i) - \min(i)} \qquad (1)$$

and min(i) and max(i) are the minimum and maximum observed values of the ith attribute. We attempted 3 weighting functions, uniform weighting, 1–distance and inverse distance, and found inverse distance to be the best (data not shown). The score of the vote is then normalized to lie between 0 and 1 to yield the probability of d_i coupling to proteins in C.

2.1.1 Chi-square feature selection

The chi-square statistic measures the dependence between a given binary feature x and a classification category c. We chose to use chi-square in our study because it is one of the most effective feature selection methods in text classification[14] and because it has been successfully applied in GPCR subfamily classification[13].

In our task, the classification category c is the group of GPCRs which can couple to proteins in G-protein family C. Twenty binary features x are derived from each n-gram count by considering whether the n-gram has occurred at least *i* times in the sequence, where *i* = 5, 10 … 100 for unigrams and *i* = 1, 2 … 20 for all other n-grams. We computed the chi-square statistic for each feature x as the normalized square of the difference between the "expected" e(c,x) and observed o(c,x) number of objects in c with feature x. The "expected" number is the number of instances in c with feature x if x had a uniform distribution over all categories. Thus, the formula for chi-square statistic is

$$\chi^2(\mathrm{x},\mathrm{c}) = \frac{[\mathrm{e}(\mathrm{c},\mathrm{x}) - \mathrm{o}(\mathrm{c},\mathrm{x})]^2}{\mathrm{e}(\mathrm{c},\mathrm{x})} \quad \text{where} \quad \mathrm{e}(\mathrm{c},\mathrm{x}) = \#\,\mathrm{GPCRs\ in\ c} \times \frac{\#\,\mathrm{GPCRs\ having\ x}}{\#\,\mathrm{GPCRs\ in\ dataset}} \quad (2)$$

Next, for each n-gram j, we found the value i_{max} such that the binary feature x_j^* of having at least i_{max} occurrences of j has the highest chi-square statistic out of the 20 derived binary features associated with j. The n-grams were then sorted in decreasing order by the chi-square value of their respective x_j^*. The top *p* n-grams were selected where *p* is tuned from data, and each feature vector was transformed into one of length *p* where the components were the derived binary features x_j^*.

2.2 Alignment Based Module

Like our n-gram based module, our alignment based module utilized the k-NN classifier, with the bit score from Basic Local Alignment Search Tool (BLAST)[15] as its similarity metric. Given a test sequence d_i, we retrieved the top *k* training set sequences with the highest BLAST bit score. Unlike previous studies which used only the intracellular domains, we used the alignment of the whole receptor sequence, since other parts of the sequence prove to be informative as well. The probability of the GPCR sequence d_i coupling to proteins from G-protein family C was computed as the fraction of retrieved sequences which couple to proteins in C.

2.3 Hybrid Prediction Method

Our hybrid prediction method combines n-gram and alignment information in making the coupling specificity prediction by utilizing the probabilities from both the n-gram based

module and the alignment-based module. Given a GPCR sequence and a G-protein family C, the method predicts the receptor to couple to proteins in C if either the probability of the interaction occurring computed by the alignment-based module is 1 or if the probability computed by the n-gram based module is above the trained threshold.

3 Data

We compared the performance of our approach to the current state-of-the-art[3] on their own dataset derived from the *2001 Trends in Pharmaceutical Sciences (TiPS) Nomenclature Supplement*[16] with added sequences from the authors of the study. We were able to replicate the entire test set but only 81% of the training set. The test set sequences had 49.4% sequence identity on average with the most similar training set sequence. We used WEKA[17] implementation of k-NN in our n-gram based module.

In addition, we assessed our hybrid method in a ten-fold cross validation and performed feature analysis to determine the location of coupling specificity information in the GPCR sequence on a more recent dataset, the Pharmacological Targets Database (PTbase)[18]. We used only the human sequences from PTbase which includes all the unique GPCRs in the database and yields test sets having 54.7% sequence identity on average with the most similar training set sequence. While the PTbase dataset contains significantly more sequences than the dataset from [3], it is not a superset of the latter. To test robustness, we also evaluated our method on 2 subsets of PTbase dataset, removing either the 8 sequences having higher than 75% sequence identity or the 4 sequences having higher than 80% with any training set sequence. In working with PTbase, we used an implementation of k-NN designed for sparse datasets developed by Paul Bennett.

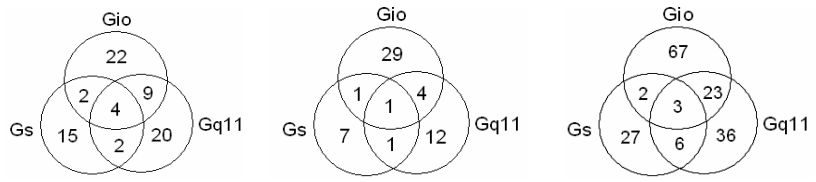

Figure 1. Distribution of training (left) and test (middle) sets from [3] and PTbase 2004 dataset (right).

4 Results and Discussion

4.1 Comparison to Current State-of-the-Art

Using the dataset from the current state-of-the-art[3] study, we assessed our n-gram and alignment based modules independently and as part of our hybrid prediction method to explore whether any advantage was gained by combining the two types of information. Since the training and test sets were not representative of each other and Cao *et al.* had optimized their parameters using the test set, we examined the performance of our method on the test set at various parameter settings (Table 1). The number of neighbors K used by each k-NN in our n-gram module was tuned on the training set constraining $K \leq 5$.

Compared to the state-of-the-art[3], our n-gram based module outperformed it in F1 and matched either its precision or recall but not both at the same time. Our alignment based module surpassed it in precision, recall and F1 all at once. Moreover, our hybrid method outperformed both modules in precision, recall and F1, demonstrating an advantage in combining n-gram and alignment information in the coupling specificity prediction task. Optimizing our parameters on the training set to avoid overfitting the prediction model, our hybrid method attained 82.4% in accuracy compared to the current state-of-the-art's reported accuracy of 72% by optimizing on the test set.

Table 1. Comparison of our n-gram based module, alignment based module and hybrid prediction method against current state-of-the-art[3]. Prob. Thres.: probability threshold in decision criterion.

Method	Prob. Thres.	Precision	Recall	F1
N-gram Module	0.26	0.514	0.889	0.651
	0.34	0.658	0.794	0.719
Alignment Module	0.50	0.630	0.921	0.748
Hybrid Method	0.66	0.698	0.952	0.805
Cao *et al.* [3]		0.577	0.889	0.700

4.2 Evaluation on Current Dataset

Having shown our hybrid method outperformed the current state-of-the-art in G-protein coupling specificity prediction, we evaluated our method on the more recent PTbase dataset[18]. Our evaluation protocol is a ten-fold cross validation, where in each trial, 8 folds were used as the training set, 1 fold as the validation set to optimize parameters for maximum F1 and 1 fold as the test set. In order to determine the number of features , we performed ten-fold cross validations with the n-gram based module at varying number of features. Figure 2 shows the average validation set F1 against the number of features. The n-gram based module attained optimal validation set F1 at 1375 features. The hybrid method using the top 1375 chi-selected features scored 0.749 precision, 0.763 recall, 0.753 F1 and 0.796 accuracy on the test set.

To test robustness, we evaluated the hybrid method on two subsets of the PTbase dataset. It attained 0.793 accuracy and 0.752 F1 on sequences less than 70% identical with any training sequence and 0.792 accuracy and 0.748 F1 on sequences less than 80%.

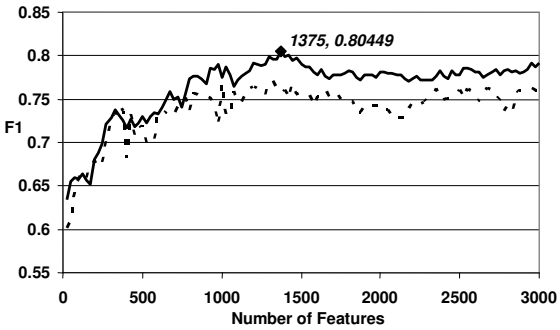

Figure 2. N-gram based module's validation set (solid) and test set (dotted) F1 versus the number of features.

5 Biological analysis

One advantage of our prediction method is its simplicity and modularity which allows us to make biological interpretations of the data and our predictions. Of particular interest is the location of coupling specificity information in the GPCR sequence. Since the GPCR makes physical contact with the G-protein via its intracellular domains, it is likely the intracellular domains would contain the majority of the coupling specificity information. Our goal is to determine which intracellular domains and which areas within them contain the most information.

5.1 Domain Combination Analysis

To determine which intracellular domains together provide the most information in predicting coupling specificity, we compared the performance of our alignment based module in a ten-fold cross validation on each of the 15 possible combinations of the 4 intracellular domains. In each trial, we reserved 1 fold as test set and 1 fold as validation set to optimize parameters for maximum F1 while training on the other 8 folds. From Table 4, the 2nd and 3rd intracellular domains together yielded the best F1. Using a single intracellular domain, the 2nd domain generated the highest F1, followed by the 1st, 3rd and 4th domains in order. However, the 1st and 2nd domains together yielded a lower F1 than the 2nd domain with either the 3rd or 4th. This suggested that the coupling specificity information in the 1st intracellular domain overlaps largely with the information in the 2nd.

Table 4. Performance of alignment based module on different intracellular domain combinations in ten-fold cross validation on PTbase dataset. IC: intracellular domains

IC	Precision	Recall	F1	Accuracy
1	0.782	0.703	0.739	0.796
2	0.820	0.799	0.808	0.845
3	0.661	0.721	0.682	0.730
4	0.632	0.755	0.670	0.694
1, 2	0.820	0.805	0.811	0.847
1, 3	0.799	0.765	0.780	0.825
1, 4	0.780	0.755	0.765	0.807
2, 3	0.837	0.825	0.828	0.861
2, 4	0.828	0.816	0.821	0.853
3, 4	0.773	0.807	0.788	0.821
1, 2, 3	0.822	0.814	0.816	0.850
1, 2, 4	0.807	0.809	0.807	0.843
1, 3, 4	0.792	0.807	0.797	0.832
2, 3, 4	0.839	0.820	0.828	0.861
1, 2, 3, 4	0.824	0.813	0.817	0.853

5.2 Motif Location Analysis

We examined the portions of intracellular domains that contain the majority of coupling specificity information by finding the maximally discriminative (maximally predictive) n-grams. For each G-protein family, we applied chi-square feature selection to identify discriminative binary features derived from n-gram counts which can differentiate between receptors that couple to the G-protein family and those that do not.

A feature is considered highly discriminative if its presence in receptors that couple to the G-protein family is much more prevalent than its presence in receptors that do not, or vice versa. For the majority of the top 100 selected features, we observed that the presence of the selected features for $G_{i/o}$ is indicative of not coupling to $G_{i/o}$ while the presence of the selected features for $G_{q/11}$ and G_s is indicative of coupling to $G_{q/11}$ and G_s

respectively (data not shown). Thus, we examined the location of the selected features for $G_{i/o}$ in receptors that do not couple to $G_{i/o}$ and the location of the selected features for $G_{q/11}$ and G_s in receptors that couple to $G_{q/11}$ and G_s.

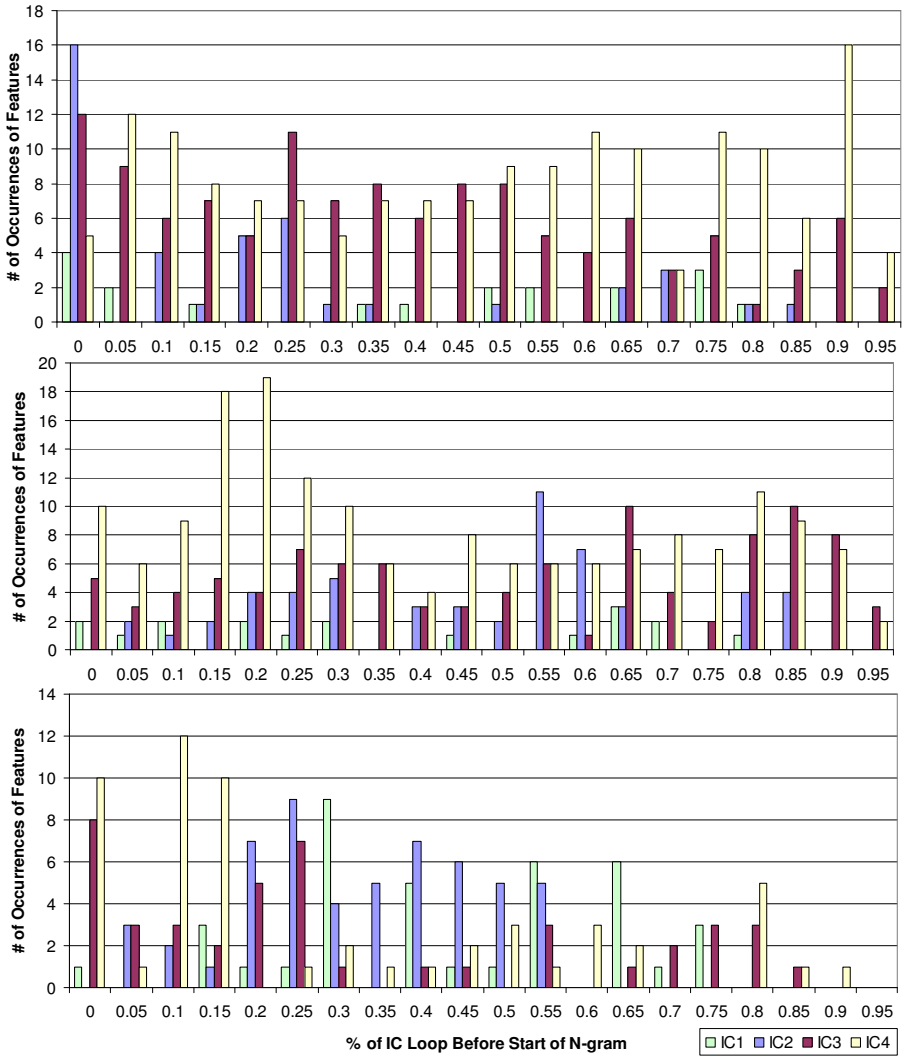

Figure 3. Histograms of the locations of the top 50 features selected by chi-square to distinguish GPCRs that couple to $G_{i/o}$ (top), $G_{q/11}$ (middle) and G_s (bottom) proteins, in receptors not coupling to $G_{i/o}$ (top) and receptors coupling to $G_{q/11}$ (middle) and G_s (bottom), respectively.

Figures 3 shows histograms of where in the intracellular domains the top 50 n-grams start for $G_{i/o}$, $G_{q/11}$ and G_s respectively. For instance, if the intracellular domain is 25 amino acids long and the n-gram on which the selected feature is based starts at the 6th

amino acid, then 5% of the domain lies before the n-gram and this contributes a count towards the '0.05' column. The selected n-grams for all 3 G-protein families were concentrated in the 2nd, 3rd and 4th intracellular domains. In the 2nd intracellular domain, the selected n-grams for $G_{q/11}$ were spread across the entire length of the domain, while those for $G_{i/o}$ and G_s were concentrated in the first 30% and 60% of the domain respectively. The selected n-grams in the 3rd intracellular domain were spread across the entire length of the domain for all 3 families. The concentration of discriminative n-grams in the beginning portion of the 2nd intracellular loop and over the entire length of the 3rd intracellular loop correlates with the physical positioning of the G-protein with the GPCR in 3-dimensional space when they couple.

6 Conclusions

G-protein coupled receptors are involved in numerous diseases due to their role with G-proteins in regulating the cell's response to external stimuli. Studies on the interaction between GPCR and G-proteins can lead to insights and potential drug targets for these diseases. In this paper, we developed a new method to predict the families of G-proteins a GPCR can interact with given the receptor sequence, that outperformed the current state-of-the-art[3]. Analyzing the features used by this method, we found the coupling specificity information to be concentrated in the beginning of the 2nd intracellular loop in the GPCR and over the entire length of the 3rd loop.

Our method differs from previous G-protein coupling specificity prediction methods in two major ways. First, previous studies focused only on the intracellular domains of the receptor as those are the regions having direct contact with the interacting G-protein. We found evidence of coupling specificity information in the non-intracellular domains of the receptor sequence in our preliminary studies and developed our method to utilize information from the whole receptor sequence instead of the intracellular regions alone.

Second, features derived from n-grams and sequence alignments are commonly used in many prediction problems in bioinformatics. Previous coupling specificity studies have all used either n-grams or alignment information but not both. Yet, the two types of features have complimentary strength: alignment can capture long-range interaction information but is unreliable below 40% sequence similarity due to its assumption that the linear ordering of conserved segments is preserved, while n-grams makes no such assumption but cannot capture long-range interactions. By combining the information in both types of features, our prediction method outperformed the current state-of-the-art[3] with only 81.3% of the training data and attained 0.753 F1 and 0.796 accuracy on the PTbase 2004 dataset[18]. Moreover, our method suffered less than 0.005 drop in accuracy and F1 when sequences sharing more than 75% sequence identity were removed. This demonstrates the potential in combining multiple representations of sequence or other information in prediction problems.

Acknowledgments

The authors would like to thank Dr. Paul N. Bennett for the use of his efficient k nearest neighbors program for sparse feature vectors, and Dr. Judith Klein-Seetharaman for sharing her GPCR expertise. This material is based upon work supported by the National Science Foundation under grant no. 0225656.

References

1. G. Muller, Towards 3D structures of G protein-coupled receptors: a multidisciplinary approach. Curr Med Chem, 2000. **7**(9): p. 861-88.
2. J. Wess, Molecular basis of receptor/G-protein-coupling selectivity. Pharmacol Ther, 1998. **80**(3): p. 231-64.
3. J. Cao, et al., A naive Bayes model to predict coupling between seven transmembrane domain receptors and G-proteins. Bioinformatics, 2003. **19**(2): p. 234-40.
4. B. Qian, et al., Depicting a protein's two faces: GPCR classification by phylogenetic tree-based HMMs. FEBS Lett, 2003. **554**(1-2): p. 95-9.
5. S. Moller, J. Vilo, and M.D. Croning, Prediction of the coupling specificity of G protein coupled receptors to their G proteins. Bioinformatics, 2001. **17 Suppl 1**: p. S174-81.
6. A. Henriksson, Prediction of G-protein Coupling of GPCRs - A Chemometric Approach, in Engineering Biology. 2003, Linkoping University: Linkoping. p. 79.
7. S. Vinga and J. Almeida, *Alignment-free sequence comparison-a review.* Bioinformatics, 2003. **19**(4): p. 513-23.
8. M. Lynch, *Intron evolution as a population-genetic process.* Proc Natl Acad Sci U S A, 2002. **99**(9): p. 6118-23.
9. Y.X. Zhang, et al., Genome shuffling leads to rapid phenotypic improvement in bacteria. Nature, 2002. **415**(6872): p. 644-6.
10. C.H. Wu, et al., *Protein family classification and functional annotation.* Comput Biol Chem, 2003. **27**(1): p. 37-47.
11. W.R. Pearson, *Effective protein sequence comparison.* Methods Enzymol, 1996. **266**: p. 227-58.
12. W.R. Pearson, Empirical statistical estimates for sequence similarity searches. J Mol Biol, 1998. **276**(1): p. 71-84.
13. B.Y. Cheng, J.G. Carbonell, and J. Klein-Seetharaman, *Protein Classification based on Text Document Classification Techniques.* Proteins: Structure, Function and Bioinformatics, 2005. **58**(4): p. 955-70.
14. Y. Yang and J. Pedersen. A comparative study on feature selection in text categorization. in 14th International Conference on Machine Learning. 1997. Nashville, US: Morgan Kaufmann Publishers, San Francisco, US.
15. S.F. Altschul, et al., *Basic local alignment search tool.* J Mol Biol, 1990. **215**(3): p. 403-10.
16. S. Alexander, et al., *TiPS Receptor and Ion Channel Nomenclature Supplement.* Trends in Pharmacological Sciences, 2001.
17. I.H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools with Java Implementations. 2000, San Francisco: Morgan Kaufmann.
18. PTbase, *PTbase*. 2004, BioMedNet.