A Machine Text-Inspired Machine Learning Approach for Identification of Transmembrane Helix Boundaries

Betty Yee Man Cheng[1], Jaime G. Carbonell[1], and Judith Klein-Seetharaman[1, 2]

[1] Language Technologies Institute, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA, 15213, USA
{ymcheng, jgc}@cs.cmu.edu
[2] Department of Pharmacology, University of Pittsburgh, 200 Lothrop Street, Pittsburgh, PA, 15261, USA
judithks@cs.cmu.edu

Abstract. In this paper, we adapt a statistical learning approach, inspired by automated topic segmentation techniques in speech-recognized documents to the challenging protein segmentation problem in the context of G-protein coupled receptors (GPCR). Each GPCR consists of 7 transmembrane helices separated by alternating extracellular and intracellular loops. Viewing the helices and extracellular and intracellular loops as 3 different topics, the problem of segmenting the protein amino acid sequence according to its secondary structure is analogous to the problem of topic segmentation. The method presented involves building an n-gram language model for each 'topic' and comparing their performance in predicting the current amino acid, to determine whether a boundary occurs at the current position. This presents a distinctly different approach to protein segmentation from the Markov models that have been used previously and its commendable results is evidence of the benefit of applying machine learning and language technologies to bioinformatics.

1 Introduction

Predicting the function of a protein from its amino acid sequence information alone is one of the major bottlenecks in understanding genome sequences and an important topic in bioinformatics. Mapping of protein sequence to function can be viewed as a multi-step cascaded process: the primary sequence of amino acids encodes secondary structure, tertiary or 3-dimensional structure, and finally quaternary structure, a functional unit of multiple interacting protein subunits. Proteins are divided broadly into two classes, soluble proteins and transmembrane proteins. The problem of predicting secondary structure from the primary sequence in soluble proteins has been viewed predominantly as a 3-state classification problem with the state-of-the-art performance at 76% when multiple homologous sequences are available [1]. The problem of predicting secondary structure in transmembrane proteins has been limited to predicting the transmembrane portions of helices in helical membrane proteins [2, 3]. Here, accuracy is more difficult to assess because there is a very limited number of transmembrane proteins with known 3-dimensional structure, and membrane lipids are

M.-S. Hacid et al. (Eds.): ISMIS 2005, LNAI 3488, pp. 29 – 37, 2005.

usually not included in these structures. For both soluble and transmembrane proteins, a large portion of inaccuracy comes from the boundary cases. However, in many biological applications, knowing the precise boundaries is critical. This paper addresses a subproblem of the general protein segmentation problem by limiting the context to G-protein coupled receptors, an important superfamily of helical transmembrane proteins where the order and type of secondary structures within each protein are known. However, the approach can be extended to any helical transmembrane protein. In order to address structural segmentation in proteins with high accuracy, we combine domain insights from structural biology with machine learning techniques proven for the analogous task of topic segmentation in text mining.

1.1 G Protein Coupled Receptors

G Protein Coupled Receptors (GPCRs) are transmembrane proteins that serve as sensors to the external environment. There are now more than 8000 GPCR sequences known [4], but only a single known 3-dimensional structure, namely that of rhodopsin [5, 6]. This is due to the fact that the structures of transmembrane proteins are difficult to determine by the two main techniques that give high-resolution structural information, NMR spectroscopy and x-ray crystallography. However, detailed information about the structure of individual GPCRs is urgently needed in drug design as approximately 60% of currently approved drugs target GPCR proteins [7]. The distribution of hydrophobic amino acids suggests a common secondary structure organization of alternating alpha helices and loops (Fig. 1): there are seven transmembrane helices, an (extracellular) N-terminus, three extracellular loops, a (cytoplasmic) C-terminus, and three cytoplasmic loops.

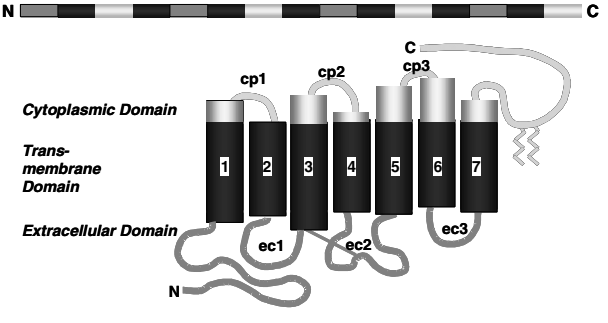

Fig. 1. Schematic of the amino acid sequence and secondary structure of a GPCR. Extracellular and cytoplasmic loops are colored dark grey and light grey respectively

Due to insufficient real training data for predicting the boundaries of transmembrane helices in GPCR, the training and testing data used in this study (except for rhodopsin) are synthetic. They are predictions based on hydrophobicity, which have been accepted by the majority of biologists as the closest estimates to the true boundaries. Because our approach does not use hydrophobicity information directly, a consensus between our predictions and the hydrophobicity predictions can be interpreted as additional evidence that the particular predicted boundary point is correct.

1.2 Related Work

A number of algorithms have been proposed to predict transmembrane domains in proteins using only amino acid sequence information, from the first Kyte-Doolittle approach based on hydrophobicity plots [8] to the more recent algorithms TMHMM [9] and PRED-TMR [10]. Most of these methods are either window-based or make use of Markov models. Window-based algorithms predict the secondary structure of the central amino acid in the window by examining the residues in the local window, using information such as frequencies of individual amino acids in each type of secondary structure, correlations among positions within the window, and evolutionary information via multiple sequence alignment of homologous sequences. Recently, improvements have also been found in considering interactions in the sequence outside the fixed window [11].

Like in most areas of computational biology, Markov models have been found to be useful in predicting the locations of transmembrane helices and are among the most successful prediction methods, including MEMSAT [12], HMMTOP [13] and TMHMM [9]. The models differ in the number of states, where each state is a Markov model on its own, representing different regions of the helices, extracellular or cytoplasmic loops.

Due to the lack of a standard dataset, the performance of the various approaches to predicting transmembrane alpha helices is controversial. Recently, a server was established that compares the performance of different methods using a single testing dataset with both soluble and transmembrane proteins. However, the training dataset is not uniform across the methods, making the results of the comparison unreliable [3, 14]. Moreover, since these methods are available only as programs pre-trained on different datasets, a fair comparison between these methods and our own is not possible.

2 Approach

In human languages, topic segmentation has many applications, particularly in speech and video where there are no document boundaries. Beeferman et al. [15] introduced a new statistical approach to segmentation in human languages based on exponential models to extract topicality and cue-word features. In essence, Beeferman and his colleagues calculated the predictive ratio of a topic model vs. a background model, and where significant changes (discontinuities) were noted, a boundary hypothesis is generated. Other features of the text string cuing boundaries were also used to enhance performance. Here, we adapted their notion of topicality features for GPCR segmentation.

2.1 Segmentation in Human Languages

Beeferman et al. [15] used the relative performance of two language models, a long-range model and a short-range model, to help predict the topic boundaries. The long-range model was trained on the entire corpus, while the short-range model was trained on only data seen since the last (predicted) boundary. This causes the short-range

model to be more specifically targeted to the current topic and as a result, it performs better than the long-range model while inside the current topic. However, at a topic boundary, the short-range model's performance would suddenly drop below the long-range model's performance because it is too specific to the last topic instead of the general corpus, and it would need to see a certain amount of data from the new topic before it can again outperform the long-range model. This was tracked using topicality measure — the log ratio of the short-range model's performance to the long-range model's performance in predicting the current word. Beeferman et al. [15] used this to detect the general position of the topic boundary, and cue-words (words that often occur near a topic boundary) to fine-tune the prediction.

2.2 GPCR Segmentation

Since the type and order of segments in the GPCR secondary structure is known (Fig. 1), we built a language model for each of the segments and compared the probability each of them assigns to the current amino acid to determine the location of the segment boundary. The reason for not building a short-range model and a long-range model as in Beeferman et al. [15] is that the average length of a protein segment is 25 amino acids — too short to train a language model. Previous segmentation experiments using mutual information [16] and Yule's association measure [17] have shown the helices to be much more similar to each other than to the extracellular and cytoplasmic loops. Similarly, the N-terminus and C-terminus have been shown to be very similar to the extracellular and cytoplasmic loops respectively. Moreover, since no two helices, extracellular or cytoplasmic segments occur consecutively, 3 segment models for helices, extracellular domains and intracellular domains are sufficient.

Each of the segment models is an interpolation of 6 basic probability models — a unigram model, a bigram model and 4 trigram models, where a 'gram' is a single amino acid. One of the trigram models, as well as the unigram and bigram models, uses the complete 20 amino acid alphabet. The other 3 trigram models make use of three reduced alphabets where a group of amino acids sharing a common physiochemical property, such as hydrophobicity, is reduced to a single alphabet letter:

1. LVIM, FY, KR, ED, AG, ST, NQ, W, C, H, P
2. LVIMFYAGCW, KREDH, STNQP, and
3. LVIMFYAGCW, KREDHSTNQP.

The reason for using reduced amino acid alphabets is because sometimes a position in a primary sequence may call for any amino acid with a certain biochemical property rather than a specific amino acid, for example, hydrophobicity in transmembrane proteins.

2.2.1 Boundary Determination

As expected from the limited context of the trigram models, the relative performance of the 3 segment models fluctuates significantly, making it difficult to pinpoint locations where one model begins to outperform another overall. To smooth out the fluctuations, we compute running averages of the log probabilities. Figure 2 shows the running averages of the log probabilities over a window size of ±2.

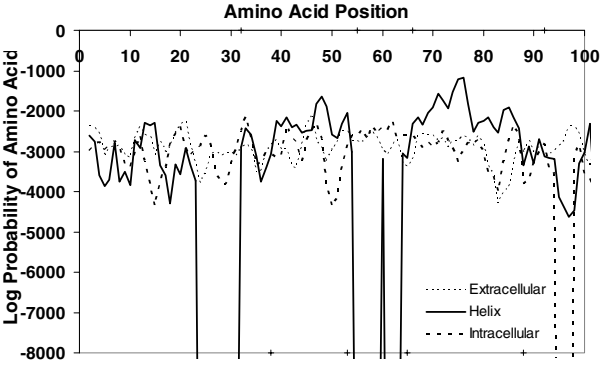

Fig. 2. Running averages of log probabilities at each position in D3DR_RAT sequence. Vertical dashed and dotted lines show the "true" and predicted boundaries respectively

While running averages minimize the fluctuations, we still do not want the system to label a position as a boundary point if the model for the next segment outperforms the current segment model only for a few positions. An example is the region in figure 2 between position 10 and 20 where the helix model performs better than the extracellular loop model temporarily before losing to the extracellular model again. Thus, we set a look-ahead interval: the model for the next segment must outperform the current segment model at the current position and at every position in the look-ahead interval for the current position to be labeled a segment boundary.

3 Evaluation

3.1 Dataset

The data set used in this study is the set of full GPCR sequences uploaded to GPCRDB [18] in September 2002. The headers of the sequence files contain the predicted segment boundaries taken as the synthetic "truth" in our training and testing data. This header information was retrievable only for a subset of these sequences, 1298 GPCRs. Ten-fold cross validation was used to evaluate our method.

3.2 Evaluation and Parameter Optimization

Two evaluation metrics were used: average offset and accuracy. Offset is the absolute value of the difference between the predicted and "true" boundary positions. An average offset was computed across all boundaries and for each of the 4 boundary types: extracellular-helix, helix-cytoplasmic, cytoplasmic-helix, and helix-extracellular. In computing accuracy, we assigned a score of 1 to a perfect match between the predicted and true boundary, a score of 0.5 for an offset of ±1, and a score of 0.25 for an offset of ±2. The scores for all the boundaries in all the proteins were averaged to produce an accuracy score.

The two parameters (running average window size and look-ahead interval) were adjusted manually to give the maximum accuracy score. One parameter was held constant, while the other parameter was adjusted to find a local maximum. Then the roles were reversed. This was repeated until the parameter values converged.

4 Results and Analysis

Table 1 describes the accuracy and offsets for all 4 types of boundaries — extracellular-helix (E-H), helix-cytoplasmic (H-C), cytoplasmic-helix (C-H), and helix-extracellular (H-E). Linear interpolation of the six probability models, after normalization to account for the differences in vocabulary size, assigns all of the interpolation weight to the trigram model with the full amino acid alphabet. We experimented with "Trained" interpolation weights (i.e. only trigram model with full amino acid alphabet), and pre-set weights to use "All" the models or only the 4 "Trigram" models. The window-size for running averages and the look-ahead interval in each case were optimized. Note there is little variance in the offset over the 4 types of boundaries.

Table 1. Evaluation results of boundary prediction. "Trained": trained interpolation weights, window-size ±2, look-ahead 5. "All": 0.1 for unigram and bigram model, 0.2 for trigram model, window-size ±5, look-ahead 4. "Trigram": 0.25 for each trigram model, window-size ±4, look-ahead 4

Weights	Accuracy	Offset				
		E-H	H-C	C-H	H-E	Avg
Trained	0.2410	35.7	35.4	34.5	36.5	35.5
All	0.2228	47.9	47.5	44.9	48.2	47.2
Trigram	0.2293	50.4	50.2	47.6	50.6	49.8

Using only the trigram model with the full amino acid alphabet shows a 5.1% improvement over using all 4 trigram models, which in turn shows a 2.9% improvement over including the unigram and bigram models. This suggests that the unigram and bigram models and reduced alphabets are not very useful in this task. However, the unigram and bi-gram models help in lessening the offset gap between predicted and true boundaries when they are more than 2 positions apart.

4.1 Discrepancy Between Accuracy and Offset

The accuracy in all of our results ranges from 0.22 to 0.24, suggesting an offset of ±2 positions from the synthetic boundaries. However, our measured offsets lie between 35 and 50. This is because the offset measure (in the trained interpolation weights case) has a large standard deviation of 160 and a maximum of 2813 positions. A histogram of the offsets (Fig. 3) shows a distribution with a very long tail, suggesting that large offsets between our predictions and the synthetic true boundaries are rare. After removing the 10% of proteins in our dataset with the largest offset averaged across their 14 boundaries, the average offset decreases from 36 to 11 positions. This result suggests that the large offsets are localized in a small number of proteins instead of being general for the dataset.

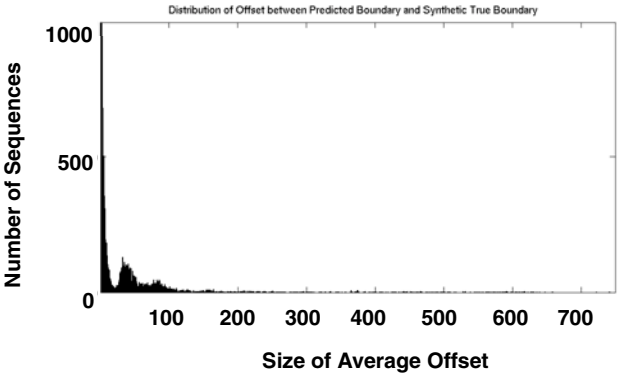

Fig. 3. Histogram of the number of sequences with the given average offset from the trained interpolated models. Note that the bars for the small offsets have been cut off at 1000 in the graph below for visibility

The distribution of offsets shows a local maximum at 36 amino acids, approximately the length of a helix plus a loop. This suggests that we may be missing the beginning of a helix and not predicting any boundaries as a result until the next helix approximately 35 positions later. To test this hypothesis, we re-evaluate our boundary predictions ignoring their order. That is, we measure the offset as the minimal absolute difference between a predicted boundary point and any synthetic true boundary point for the same sequence. The distribution of the new offsets is plotted in figure 4. The lack of a peak at position 36 confirms our hypothesis that the large offsets when evaluated in an order-specific fashion are due to missing the beginning of a helix and becoming asynchronized.

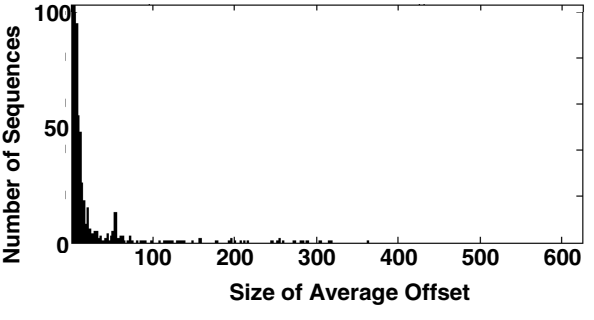

Fig. 4. Histogram of the order-independent offsets from the trained interpolated models. Bars for the small offsets have been cut off at 100 in the graph below for visibility

4.2 The Only Truth: Rhodopsin OPSD_HUMAN

As described in Section 1.1, rhodopsin is the only GPCR for which there is experimental evidence of the segment boundary positions. Below are the predictions of our

approach on rhodopsin using the trained interpolated models. The average position offset is 1.35.

Predicted:	37	61	72	97	113	130	153
Difference:	1	0	1	1	0	3	1
True:	36	61	73	98	113	133	152
Predicted:	173	201	228	250	275	283	307
Difference:	3	1	2	2	1	1	2
True:	176	202	230	252	276	284	309

5 Conclusions

In this paper, we addressed the problem of protein segmentation in the limited domain of GPCR where the order and type of secondary structure segments are known. We developed a new statistical approach to protein segmentation that is distinctly different from the fixed window and Markov model based methods currently used. Taking the different types of segments as "topics" in the protein sequence, we adapted a topic segmentation approach for human languages to this biological problem. We built a language model for each of the different segment types present in GPCRs, and by comparing their performance in predicting the current amino acid, we determine whether a segment boundary occurs at the current position. Each of the segment models is an interpolated model of a unigram, a bigram and 4 trigram language models.

The results from our approach is promising, with an accuracy of 0.241 on a scale where 0.25 is an offset of ±2 positions from the synthetic boundaries predicted by hydrophobicity profiles. When the gap between the predicted boundary and the synthetic "true" boundary is 3 or more amino acids wide, the gap tends to be much larger than 3. This is because our approach relies on knowledge of the segment order and a 'missed' boundary can cause the system's perception of the protein to be misaligned, leading it to compare the wrong models to detect the upcoming boundaries. This occurred with a small number of GPCRs which have an N-terminus that is several orders of magnitude longer than the average length of that segment. For such proteins, we plan to use HMM in the future to predict multiple possibilities for the first segment boundary and then apply our approach to predict the upcoming boundaries given the first boundary. The resulting sets of 14 boundaries can then be evaluated to determine the most likely one. Furthermore, the addition of "cue-words" — n-grams frequently found close to segment boundaries — and long-range contact information should help to reduce the offset of ±2.

Acknowledgements

This research was supported by National Science Foundation Large Information Technology Research grant NSF 0225656.

References

1. Rost, B., *Review: protein secondary structure prediction continues to rise.* J Struct Biol, 2001. **134**(2-3): p. 204-218.
2. Chen, C.P., A. Kernytsky, and B. Rost, *Transmembrane helix predictions revisited.* Protein Science, 2002. **11**(12): p. 2774-2791.
3. Chen, C.P. and B. Rost, *State-of-the-art in membrane protein prediction.* Applied Bioinformatics, 2002. **1**(1): p. 21-35.
4. Bateman, A., et al., *The Pfam protein families database.* Nucleic Acids Res, 2002. **30**(1): p. 276-80.
5. Okada, T., et al., *Functional role of internal water molecules in rhodopsin revealed by X-ray crystallography.* Proc Natl Acad Sci U S A, 2002. **99**(9): p. 5982-5987.
6. Palczewski, K., *Crystal structure of rhodopsin: implication for vision and beyond. Mechanisms of activation.* Scientific World Journal, 2002. **2**(1 Suppl 2): p. 106-107.
7. Muller, G., *Towards 3D structures of G protein-coupled receptors: a multidisciplinary approach.* Current Medical Chemistry, 2000. **7**(9): p. 861-888.
8. Kyte, J. and R.F. Doolittle, *A simple method for displaying the hydropathic character of a protein.* J Mol Biol, 1982. **157**(1): p. 105-32.
9. Sonnhammer, E.L., G. von Heijne, and A. Krogh, *A hidden Markov model for predicting transmembrane helices in protein sequences.* Proc Int Conf Intell Syst Mol Biol, 1998. **6**: p. 175-182.
10. Pasquier, C., et al., *A novel method for predicting transmembrane segments in proteins based on a statistical analysis of the SwissProt database: the PRED-TMR algorithm.* Protein Engineering, 1999. **12**(5): p. 381-385.
11. Schmidler, S.C., J.S. Liu, and D.L. Brutlag, *Bayesian segmentation of protein secondary structure.* Journal of Computational Biology, 2000. **7**(1-2): p. 233-248.
12. Jones, D.T., W.R. Taylor, and J.M. Thornton, *A model recognition approach to the prediction of all-helical membrane protein structure and topology.* Biochemistry, 1994. **33**(10): p. 3038-3049.
13. Tusnady, G.E. and I. Simon, *Principles governing amino acid composition of integral membrane proteins: application to topology prediction.* J Mol Biol, 1998. **283**(2): p. 489-506.
14. Kernytsky, A. and B. Rost, *Static benchmarking of membrane helix predictions.* Nucleic Acids Research, 2003. **31**(13): p. 3642-3644.
15. Beeferman, D., A. Berger, and J. Lafferty, *Statistical Models for Text Segmentation.* Machine Learning, Special Issue on Natural Language Learning, 1999. **34**(1-3): p. 177-210.
16. Weisser, D. and J. Klein-Seetharaman, *Identification of Fundamental Building Blocks in Protein Sequences Using Statistical Association Measures*. 2004: ACM SIG Proceedings. p. in press.
17. Ganapathiraju, M., et al. *Yule values tables from protein datasets of different categories: emphasis on membrane proteins.* in *Biological Language Conference*. 2003. Pittsburgh, PA, USA.
18. Horn, F., et al., *GPCRDB: an information system for G-protein coupled receptors.* Nucleic Acids Research, 1998. **26**(1): p. 275-279.