Stream Wang
Biochemistry 218
Final Project

Prediction of Tight Junction Localization Signal(s) in the Mammalian Systems

I. The Localization Question:

Cellular compartmentalization is one of the major characteristics of the higher organisms. By dividing their cells into different sub-compartments, the mammals exert spatial control and separation of their cellular proteins. Improper localization of essential proteins can have consequences ranging from mild diseases to lethality. Therefore, the ability for a cell to target its proteins to the correct place at the correct time is extremely essential for both cellular functions and survival.

The first insight into the importance of protein localization came from the classic studies on yeast secretory mutants. In these mutants, a variety of the proteins responsible fore regulating the secretory pathway were mutated, causing mis-localization of cellular proteins. (Kaiser, 1990) As a result, these mutants demonstrated an abnormal accumulation of transport vesicles in sub-cellular organelles and for some, a decrease in cell viability. Over the past decade, much effort has been spent on identification of localization signals that are responsible for targeting proteins to a particular sub-compartment of a cell, based on the primary amino acid sequence of a protein. One of the representative programs in the field is the pSort II program developed by Dr. Kenta Nakai's research group. This program enable us to predict the presence of a variety of localization signal for a number of organelles such as ER, Golgi, mitochondria, or nucleus.

Another important characteristic of higher animal cells is the formation of intercellular junctions. Although many localization prediction programs are currently available, no program is yet capable of predicting protein localization to the tight junction. In this paper, I propose to identify and predict tight junction targeting signals

by applying similar algorithms used previously by a number of programs to determine other types of localization signals. I will start this paper by presenting a brief review on the biology of tight junction, followed by short summary discussions on the methods currently available for identifying and characterizing sub-organelle localization signals. I will then suggest method for predicting tight junction localization signals based on similar prediction programs described. And finally, I will address the possibility of using these algorithms to predict, on a whole genome basis, the proteins that are involved in trafficking to the tight junction.

II. The Biology of Tight Junction:

Epithelial cells associate with one another and form a lateral sheet lining of the small intestine. Tight junctions (TJ) are specialized plasma membrane microdomains that form continuous branching strands around each epithelial cell, separating the apical side from the basolateral side. TJ serves two important cellular functions: (1) It serves as a physical barrier that ensures no intermixing of the proteins between the apical (facing the luman) and the basolateral (facing the blood) membranes. (2) It regulates diffusion of molecules and ions across the paracellular route. Overall, proper formation of the tight junction ensures the development and maintenance of epithelial polarity. (Zahraoui, 2000)

Entrez search resulted in one hundred and ninety-two mammalian proteins that have been demonstrated to localize to the tight junction. All of these proteins appear to have intrinsic ability to localize to the tight junction, instead of indirectly targeted to tight junction by its interaction with tight junction residence molecules. Assuming that the tight junction localization signal is encoded in these primary amino acid sequences, I will use these proteins as training set in the following exercise for developing localization algorithms. In the immediate paragraphs, I have included a short description of these proteins and their possible biological functions.

<u>Rab Family of Proteins:</u>

The Rab proteins are small (20-40kD) G protein binding molecules that have been implied as the master switch in regulating intracellular trafficking. Many Rabs in the

mammalian system have been shown to localize to primarily one sub-cellular compartment. Specifically, the carboxyl terminal tail of the Rab proteins seems to be essential in specifying the target site, as suggested by chimeric experiments. Two Rabs, Rab8 and Rab13, have been demonstrated to localize to the tight junction.

Occludin Family of Proteins:

Occludins are integral membrane proteins which were found to concentrate within the tight junction fibrils. It is thought to play a role in the formation and regulation of the tight junction paracellular permeability barrier.

Claudin Family of Proteins:

Claudins are Integral membrane proteins that are components of the tight junction strands.

ZO Family of Proteins:

ZO Proteins belong to the Maguk family of cell junction proteins. It is thought to function as the molecular scaffold brining together many proteins at the tight junction.

Miscellaneous Proteins:

Other proteins that resulted form the search included the followings: paracellin, VAP-33, mint3, cingulin, MAGI-1, and symplekin.

III. Currently Available Methods for Localization Prediction Programs

Simple Consensus Sequences Based Method:

One type of localization signal is specified purely on the basis of its primary amino acid sequence. In this type of situation, an alignment among the training sequences is not necessary. A sliding window of a defined size is selected and each the windows of one sequence is used to match against similar windows in the rest of the training sequences. In this method, no gap is allowed and each amino acid is considered independent of its neighbors. A scoring matrix is generated from the training set and the best consensus

sequence is identified. The best-studied example of this type of targeting sequence is the ER retention signal, which consists of the amino acid bases of KDEL. This signal is used in the pSORT II program for identifying ER residence proteins.

Weight Matrix Based Methods:

The weight matrix based program was initially invented by von Heijne in 1986 and undergone considerable improvement over the past decade. The weight matrix, also known as the Position-specific scoring matrix (PSSM), compares vertical variations across pre-alignment family of sequences and no gaps are allowed in this method. Values are assigned to each of the twenty amino acids, and the matrix values are converted to log odds scores then to the ratio of the log score. High values of the matching log odds scores along the sequences will be selected and identified as similar regions. Some examples of programs based on this method include SPScan (GCG10, 1999) and SigCleave (EMBOSS, 1999).

Hidden Markov Model Method:

The Hidden Markov Model is a probabilistic model which considers all possible combination of matches, mismatches, and gaps among the input sequences. Although the most common use of this model to make multiple sequence alignments, it has also been used successfully to make sequence profiles as well as for pattern recognition. The standard Hidden Markov Model is build by Markov chains that consists of standard state, insertion state, and deletion state. The training sequences were used as input and the transitional probability between the states were calculated. The output model build by the HMM program dependent largely on the inputs, and differs from training set to training set. Some example of the programs based on the Hidden Markov Model method includes SignalP V2.0.b2-HMM (CBS, 2001) and the mitochondria localization signal prediction program developed by Fujiwara et al. (Fujiwara, 1997).

Neural Network Method:

Neural Network is an artificial intelligence based algorithm that trains a program by simulating how our brain works. The input of the network consisted of units that are

composed of symbolic data, and relationship between these input are used to train the program to adjust the weights connecting these units. Typically, an input sequence window is transformed into a symbolic input layer. This layer is then translated into an output layer, often via a middleman hidden layer, to a predicted outcome. Some example of the programs based on the Neural Network method includes the SignalP program V1.1 (Nielsen, 1997) and the TargetP program (Emanuelsson, 2000).

Overall, consent of the current reviews seem to suggest that both the HMM and the neural network program function much better than the consensus and weight matrix methods for predicting localization signals. Between the HMM and the Neural Network methods, the success rates are comparable and are largely dependent on the types of localization signal in question as well as input training sequences. One of the drawbacks of the neural networks is that it is generally more difficult to understand and interpret how and why it makes its predictions.

IV. Generation of Tight Junction Localization Signal Prediction Programs

I have identified, from the current database, one hundred and ninety-two mammalian proteins which have the intrinsic ability to localize to the tight junction. If a localization signal does exist for these tight junction associated proteins, it is likely encoded in its primary amino acid sequence. In this paper, I suggest to use these one hundred and ninety-two sequences as input data to train programs that will be able to identify and predict tight junction localization signals. I suggest to divide the sequences into twelve training set, each set composed of one hundred and seventy-six sequences as training data and the remaining sixteen sequences as test subjects for evaluating the success rate of this program.

First of all, if the tight junction localization signal is based purely on the presence of a certain amino acid sequence, a consensus sequence based method can be used to identify and locate identical or nearly identical, motif-like localization signals. I have tried this approach by using block program but was unable to identify a clear consensus

sequence among the input data. This strongly suggests that tight junction localization signal is not simply a consensus sequence among all the known training sequences.

To train for tight junction localization signals, I will not be applying the weight matrix method for the following reasons: (1) The Position-specific scoring matrix method is based on the assumption that the input sequences are somewhat related sequences. However, from the initial biological analysis of the sequences in Part II, this is unlikely the case since the training set composed of diverse sub-families of proteins as well as outliers. (2) As stated from the end of part III, review papers in the field have suggested that the weight matrix seems to be a less reliable method in general when used in predicting localization signals to sub-compartments. One of these papers described experiments performed to evaluate the different localization programs and found that PSSM based programs are less capable of recognizing false positive sequences embedded in the test set, as compared to the HMM or Neural Network based programs. (Menne, 2000).

Although there are differences in success rate between a HMM based method and a neural network based method in identifying localization signals, both methods are quite good and it is difficult to know *a prior*, which method is the better one for predicting tight junction localization signals. The training sets suggested earlier in this part of the paper is large enough to be used for developing both of these types of programs. To use a HMM based method, a program similar to the motif-based hidden Markov Model will be developed. (Grundy, 1997) In this training program, one can use either a sliding window of sequences as inputs, or apply an alternative program (for example, the EM algorithm) to locate ungapped sequences that are similar and present in the majority of the sequences to use as inputs. Either way, these sequences will then be used as the standard states, and the transition probability between it and the insertion or deletion state will be calculated. Given enough input sequences, a model can then be build that will represent the profiles of sequences which is responsible for targeting a protein to the tight junction. Similar sliding window input sequences or ungapped sequences can also be used as queries to build a neural network based program. The prediction efficiency can then be compared between a HMM based program and a Neural Network based program

to decide which of the two makes a better prediction program for tight junction localization signals.

This localization signal could exist in the form of one or more small blocks of consensus sequence(s), or as a global property of a certain part of the protein. Alternatively, it can also exist in a form of types of post-translational modification to the protein, which will be more difficult to decypher. If the tight junction localization signal is a sequence based, rather than a structure based signal, the amino acid sequences from the training sets can be used directly as inputs. However, if the success rate of the program based on primary amino acid sequence does not provide us with a satisfactory prediction rate, an improvement can potentially be made by entering the training sequences in a structural context. One possibility is by converting the primary amino acid sequence into a string that reflects its neighboring environment before using it in the training set. One such possibility is to use an Amino Acid indexing method such as the one developed by Bannai et al (Bannai, 2002). After converting the training sequences to the appropriate index, one can then feed these sequences into our HMM or Neural Network based programs.

Last but not least, once a program has been generated which can be used to identify the signature of a tight junction localization signal, one can then generate a profile for the localization signal and mines the genomes for yet unidentified family of proteins that can potentially be localized to the tight junction. Experiments such as immunofluorescence studies can then be performed in order to determine if these newly identified molecules do indeed localize to the tight junction. If proven true, these sequences can then be used in the future training set and by this iterative effort, we can hopefully advance our knowledge on how a protein is localized to the tight junction.

V. REFERENCES:

Bannai H., Tamada Y., Maruyama O., Nakai K., and Miyano S. Extensive feature detection of N-terminal protein sorting signals. Bioinformatics 18: 298-305.

Emanuelsson O., Nielsen H., Soren B., and von Heijne, G. 2000. Predicting subcellular localization of proteins based on their N-terminal amino acid sequences. J. Mol. Biol 300: 1005-1016.

Fujiwara Y., Asogawa M., and Nakai K. Prediction of mitochondrial targeting signals using hidden markov models. Genome Informatics Workshop. 53-60.

Grundy W., Bailey T.L., Elkan C.P., and Baker M.E. Meta-MEME: Motif-based hidden Markov models of protein families. CABIOS 13: 397-406.

Kaiser C.A., and Schekman R. Distinct sets of SEC genes govern transport vesicle formation and fusion early in the secretory pathway. Cell 61:723-733

Menne K.M.L., Hermjakob H., Apweiler R. A comparison of signal sequence prediction methods using a test set of signal peptides. Bioinformatics 16: 741-742.

Nakai K., and Kanehisa M. A knowledge base for predicting protein localization sites in eukaryotic cells. Genomics 14: 897-911.

Nielsen H., Engelbrecht J., Brunak, S., and Von Heijne G. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Protein Engineering. 10: 1-6.

Zahraoui A., Louvard D., and Galli T. Tight junction, a platform for trafficking and signaling protein complexes. J. Cell. Biol 151: F31-F36.