

Neural Nets, SVM, EM and Related Techniques for Sequence Analysis

Matthew Pendleton

MICaB 8006 - Fall semester

2006.11.07

The current trend in the field of structural and functional protein analysis is one of disjunction in that there is a wide chasm of facility between the methods for elucidating protein sequence and those methods employed to find either the tertiary structure of a protein, the functional mechanisms of a protein and finally how the former derives mainly from the latter. As such, there is a great need for automatic protein structure prediction systems(Armano).

As far back as 50 years ago it was demonstrated that in some cases the bare protein sequence information is both a necessary and sufficient determinant for the structure and functionality of a peptide, and this paradigm has held true to this day for all but a miniscule minority of special cases. The conclusion that can be drawn from this observation is that although science has yet to obtain direct structural functional data on many proteins already characterized by sequence data, such sequence data is somehow a sufficiently rich source of unique correlations that structure and function can be derived directly from it (MIC8006, lec. 2). This observation invokes the crux of this situation from the viewpoint of computational biology; how does one locate and extract the significant sequence correlations from a vast repository of raw sequence so as to be able to infer from them predictions of a structural or functional nature, particularly when those correlations are between many distant residues and are frequently only vaguely conserved? The fact that general strategies and their associated algorithms for pattern prediction are useful in a wide variety of fields has been a particular boon for the field of bioinformatics in solving the problem of inferring hidden parameters and extracting missing data from a given data set as is the very case in deriving the structure or function

of a protein from its sequence (Perrin). This review will address a few of the most effective tools in this field and their applications through a series of recent publications. Also to be addressed is another important concept underpinning the field of computational biology, that being that there are numerous ways of applying the various techniques of pattern recognition to produce new discoveries and that although finding creative solutions can be rewarding, finding and solving creative problems can also be very contributive.

The general aims of deriving higher order conclusions about proteins from sequence is a sizeable one and, considering the time it takes supercomputers to attempt folding simulations, it may seem as if this problem may be as yet intractable for at least a few more Moore's cycles. In the papers by Mooney, et al and Lee, et al, the general principle is demonstrated that one potential way of overcoming computational intractability is by conceding a trade-off: the perfect applicability one might achieve in attempting to directly extract structure from sequence through brute force for a slightly less applicable set of conclusions which are more computationally tractable. Their approach to the structure/function problem is one of studying secondary structure since a significant step towards establishing the structure and function of a protein is the prediction of the local conformation of the polypeptide chain (Mooney). This was done by deconstruction of structurally well characterized whole polypeptides into fragments; small oligopeptides whose angular ratios had been determined in crystal structure with great accuracy and precision. Submitting these oligopeptide sequences and angles to a multidimensional scaling (MDS) procedure rendered several cluster sets of distinct secondary structure classes depending on the length of the oligopeptide. Further

refinement of these data by pipelining them through a two-layered bidirectional recurrent neural network into the three traditional structure classes indicated that one could classify any given ϕ/ψ angle pair with accuracy outperforming the state of the art predictor Porter by several percentage points (Mooney).

Another recent method by which secondary structure prediction has been refined is based on the extraction and implementation of extra-sequential information from sequence, illustrating the fact that simple sequence is frequently not the only raw data that can be useful. Such extra-sequential information is found as correlations between various related sequences highlighting the varying importance of certain residues. Two previous papers upon which Lee, et al built demonstrate that employing peptide sequence in the form of residue doublet probabilities (that is, the probabilities $P_n(A|A)$ to $P_n(Y|Y)$ for each residue or 400 different values at each residue for combinations the two residues n and $n+1$, each one being one of twenty possible residues) as the input for statistical analysis. This consideration of a limited peptide neighborhood for each residue resulted in an improved accuracy in secondary structure prediction when more classical multiple linear regression was employed by Liu et al, and still another work by Cai, et al demonstrated that a non-linear regression method such as an artificial neural network improved predictive power even further. Finally Lee, et al demonstrate that accuracy can be improved further still by considering evolutionary data found in multiple sequence alignments. With their particular dataset of evolutionary weighted residue maps (derived from a modified PSI-BLAST matrix) applied to a support vector machine (SVM) learning algorithm there was an overall increase in accuracy of prediction of eight secondary structures from a 6.1% overall error rate as reported in Liu et al to 3.3% error

rate. Of particular importance is that the Lee group attempted three different learning tools in maximizing their predictive accuracy: Multiple Linear Regression, Artificial Neural Network and Support Vector Regression indicating that while these may all be perfectly suitable statistical approaches to computer learning, they are each suited to learning from different types of datasets (Lee).

So far, structural approaches have been demonstrated that each focus on a single computer learning algorithm and how such techniques can greatly enhance the accuracy of secondary structure prediction. Much as was reported by Sosa, et al regarding protein localization techniques, a hybrid analysis of any given dataset by several methodologies may be the best approach yet (MIC8006, lec. 17). There are several different ways in which this idea may be implemented, the first of which being a parallel evaluation in which a given dataset is evaluated by several methods and the collective pool of results are integrated by some sort of averaging or simple agreement-disagreement methods to synthesize a final evaluation. Another method for minimizing prediction error is to combine methods in a serial fashion rather than a parallel fashion. Jahandideh et al describe one such hybrid method where a method very similar to the di-peptide methods proposed by the two papers from the Zhou group is first evaluated with multinomial logistic regression (MLR). The results from this initial evaluation are then used as the input set for a non-linear method, in this case an artificial neural network. While this is simply recombination of old methods, it does render an enhanced accuracy in secondary structure prediction. A second paper by Armano et al employs a very novel genetic-neural methodology involving a population of “experts,” evaluational theoretical automata which have an internal structure which takes in raw data if that data meets the

criteria of the expert's selectivity filter which is then evaluated and passed to an output module which gives a prediction as well as a confidence level based on the expert's expertise. A set of supportive entities condense the input from the whole population of experts, each with varying internal mechanisms and selectivity filters. This conceptual population of evaluational automata can then be made subject to evolutionary pressures in order to better meet the demands of the first step data evaluation. Other supportive entities assign rewards, mutate strength of selectivity, expertise and scope (i.e. the amount of the whole data set visible to any particular expert) and create or destroy experts as it deems necessary in order to realistically emulate an environment of selection for better and better experts. In practice, Armano et al demonstrate that indeed these experts can be trained and they do become better with training at parsing data for an ANN to evaluate. Such an approach may serve some benefit as at its inception it rivaled other highly competent evaluational systems in its accuracy and precision (Armano).

To this point I've discussed numerous approaches and strategies by which one can evaluate sequence data and arriving at structural insights all of which are derived from their causative agents, namely the sequence of the corresponding polypeptide. A final work by Perrin et al demonstrates a very different application for computer learning methods in ascertaining the function of a given set of proteins. They apply an Expectation-Maximization (EM) regression algorithm to the learning of a temporally dynamic Bayesian network based on gene expression. They postulate that gene networks are very amenable to representation by Bayesian networks in that there are a number of hidden interdependencies with strings of nodes which represent genes in the network exhibiting the behaviors of a conditional probability function. Additionally, like all

dynamic genetic observations, there is an ambient noise through which Bayesian networks cut with minimal difficulty. After learning of a suitable Bayesian network architecture by the EM algorithm defined in their work, the resultant network was compared to observed genetic relationships in the SOS DNA repair pathway in *E. coli* and was found to reflect the natural behaviors they intended to model.

There are numerous strategies which scientists can employ in order to draw out the gems of structural insights in linear sequence data from the chaff of residues which either play no important role in a protein's overall activity or can be changed more or less drastically than one might expect while still maintaining that protein's functional viability. Additionally, new methods, some very creative and novel such as Armano's expert approach are conceived all the time, each of which bring scientists closer to being able to close the gap between the size of sequence data widely available and the corresponding structural sequence sets. It can only be a matter of time until the natural strengths of computers in finding patterns and performing vast arrays of calculations quickly can be properly applied to the problems of structural biology. This proper harnessing of computational power may require more than simply brute force direct calculations of sequence to structure but might instead focus on answering creative questions which can extract information from less likely sources such as residue neighborhoods, evolutionary relationships, or intracellular signaling dynamics.

Works cited

Giuliano Armano, Gianmaria Mancosu, Luciano Milanese, Alessandro Orro, Massimiliano Saba and Eloisa Vargiu (2005) *BMC Bioinformatics.*, **6**(Suppl 4):S3.

Christopher J. C. Burges. (1995) *Data Mining and Knowledge Discovery* 2:121 – 167.

YU-DONG CAI, XIAO-JUN LIU, KUO-CHEN CHOU (2003) *J Computational Chemistry* 24: 727–731.

Samad Jahandideh, Parviz Abdolmaleki, Mina Jahandideh, Sayyed Hamed Sadat Hayatshahi (2006) *Journal of Theoretical Biology.*, in press.

Soyoung Lee, Byung-chul Lee and Dongsup Kim(2005) *PROTEINS: Structure, Function and Bioinformatics.*, 62:1107-1114.

Derong Liu, Xiaoxu Xiong, Bhaskar DasGupta, and Huaguang Zhang *IEEE TRANSACTIONS ON NEURAL NETWORKS*, 17(4): 919-928.

Catherine Mooney, Alessandro Vullo, and Gianluca Pollastri (2006) *Journal of Computational Biology.*, **13**(8), 1489-1502.

Bruno-Edouard Perrin, Liva Ralaivola, Aur' elien Mazurie, Samuele Bottani , Jacques Mallet and Florence d'Alch'e–Buc (2003) *Bioinformatics* 19(Suppl. 2): ii138-ii148.

Abdi, H. (1994) *Journal of Biological Systems*, 2, 247-281.

Multiple authors. Wikipedia entry: Nonlinear Regression. November 1, 2006. November 4, 2006 http://en.wikipedia.org/wiki/Expectation_Maximization.