Improving Interacting Residue Prediction Using Long-Distance Information in Hidden Markov Models

Colin Kern, Alvaro J. González, Li Liao, and K. Vijay-Shanker
Department of Computer and Information Science
University of Delaware
Newark, DE 19716, USA
Email: kern@cis.udel.edu, alvaro@cis.udel.edu, lliao@cis.udel.edu, vijay@cis.udel.edu

***Abstract*—Identification of interacting residues involved in protein-protein and protein-ligand interaction is critical for the prediction and understanding of the interaction and has practical impact on mutagenesis and drug design. In this work, we introduce a new decoding algorithm, ETB-Viterbi, with early trace back mechanism built into interaction profile hidden Markov models (ipHMMs) that can incorporate the long-distance correlations between interacting residues to improve prediction accuracy. The method was applied and tested to a set of domain-domain interaction families from the 3DID database, and showed statistically significant improvement in accuracy measured by F-score. To gauge and assess the method's effectiveness in capturing the correlation signals, sets of simulated data based on the 3DID dataset with controllable correlation between interacting residues were also used, and it was demonstrated that the prediction consistently improves as the correlations increase.**

I. Introduction

Protein-protein interactions (PPI) play essential roles in cellular functions. With the emerging new paradigm of systems biology, much of the research focus has been shifted from studying individual proteins and their functions to studying how they interact with each other and form biological networks while fulfilling cellular processes. Great advancements have been witnessed in experimental technologies, such as yeast two-hybrid (Y2H) systems and coimmunoprecipitation (CoIP), for detecting PPIs [1]. Still, the cost, time and other limitations associated with the current experimental methods have motivated development of computational methods for predicting PPIs.

Identification of interacting residues involved in protein-protein and protein-ligand interaction is critical for the prediction and understanding of the interaction and has practical impact in its own right on mutagenesis and drug design. General domain identification is a highly non-trivial task. Sequence patterns based on amino acid compositions typically lack enough uniqueness to be solely relied upon for domain identification. In fact, multiple sequence alignments of proteins that are known to contain the same domain show variations in sequence composition. Hidden Markov models (HMMs) are among the most successful efforts to capture the commonalities of a given domain while allowing variations. A collection of HMMs covering many common protein domains and families is available in the Pfam database [2].

For interface domains, the interaction sites impose strong constraints, and therefore play a key role in identifying the domains. However, interaction site information is not readily available for many proteins and the dataset of PPIs that have been resolved structurally using crystallography remains relatively small. To tackle this issue, Friedrich and coworkers developed a method, called interaction profile hidden Markov model (ipHMM) [3], which modifies the ordinary profile hidden Markov model (pHMM) [4] by adding to the model architecture new states explicitly representing residues on the interface based on 3D structure of protein complexes. Once trained, the model can be used to predict interacting domains for proteins whose structural information is not experimentally available. This leads to improved accuracy in identification of interaction domains and residues.

Despite the improvement, the ipHMMs, like most hidden Markov models, are not capable of capturing long-distance correlations, without drastically increasing the computational complexity by resorting to high degree Markov chains. Yet, we have observed significant correlations among the interacting residues of a protein, sometimes separated by dozens of amino acids in the primary structure. In this work we introduce a new decoding algorithm with an early traceback mechanism in ipHMMs, a modification to the Viterbi algorithm, which is designed to incorporate long-distance correlations between interacting residues in the input sequences. To gauge and benchmark the effect of long-distance correlations on the prediction accuracy, we use data from the 3DID database, along with simulated data that is generated from 3DID data, containing varying degrees of correlation. It is shown that our method, called ETB-Viterbi, is capable of capturing the long-distance correlations for consistently improved prediction accuracy.

II. Methods

In this section, we describe in detail our method for detecting long-distance correlations among residues and leveraging such information for more accurate prediction

978-1-4673-2560-8/12/$31.00 ©2012 IEEE

of interacting residues in protein-protein interactions. We first introduces the basic concepts of hidden Markov models and summarizes ipHMM, a previous method that predicts interacting residues. Next, we dicuss how to quantitatively measure the long-distance correlation between interacting amino acids in a single protein, using the log-odds score and conditional probability. After a brief review of the Viterbi algorithm and we then describe our Early Traceback (ETB) Viterbi algorithm, which incorporates the long-distance correlations explicitly in decoding the most probable hidden path representing the sequence of interacting residues with the trained hidden Markov model.

A. Interaction Profile Hidden Markov Models

Hidden Markov models (HMMs) are probabilistic models for analyzing and simulating sequences of symbols that are emitted from underlying states hidden from direct observation [5]. The models are represented as graphs, called architecture, with the nodes corresponding to the states S_i ($i = 1$ to N), and the directed edges between the nodes corresponding to transitions between states. Symbols from an alphabet can be emitted at each state S_i , which has a set of emission parameters $e_i(x)$ specifying the probabilities for emitting each symbol x in the alphabet at that state. For each directed edge from S_i to S_j, it has a weight specifying the transition probability a_{ij} leaving state S_i and entering state S_j. The parameters $e_i(x)$ and a_{ij} for a HMM can be estimated from training data using standard procedures such as Baum-Welch algorithm [6].

HMMs are very suitable and have been successfully applied to analyzing protein and DNA sequences [7]–[9]. Friedrich *et al.* [3] proposed a model in which the interacting sites within protein domains are modeled by a modified profile HMM, called interaction profile hidden Markov model (ipHMM). The model architecture takes into account both sequence data and structural information about interaction sites. Every ipHMM is, like pHMMs, a probabilistic representation of a protein domain family. The architecture of the ipHMM follows the same restrictions and connectivity of the HMMER architecture [6], with one important exception: the match states of the classical pHMM are split into a non-interacting (M_{ni}) and an interacting match state (M_i), as shown in Figure 1, to differentiate residues that are highly conserved but not involved in interaction and residues that are not only highly conserved but also involved in the interaction. Other than this only difference, the new match state is provided with the same properties as a match state in the ordinary profile hidden Markov model architecture.

The parameters of an ipHMM are estimated using maximum likelihood based on a multiple sequence alignment of the member proteins in the domain family, incorporating the annotation of their interaction sites based on x-ray crystallographic structure of the protein complexes – all residue positions are labeled with the corresponding interaction status (0 for not interacting and 1 for interacting). These trained ipHMMs thus encode relevant statistical information about the domain, especially pertaining to the interaction with other domains. As we will show in the following subsections, there are long-distance correlations between these interacting residues which we believe are not sufficiently captured by the existing ipHMM method.

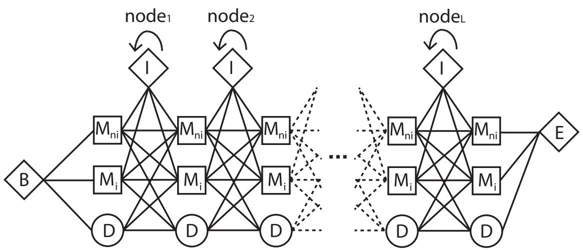

Figure 1. Architecture of the interaction profile hidden Markov model. The match states of the classical pHMM are split into non-interacting (M_{ni}) and interacting (M_i) match states.

B. Long-Distance Correlations

Previous work has found that correlation exists between the amino acids found on the boundaries of protein secondary structure domains [10], [11]. In Crooks and Brenner [10], without associating to specific locations, neighboring amino acids were found to correlate weakly, and mutual information for any pair of neighboring amino acids, even as conditioned with their secondary structure types such as being in helix or beta sheet, were shown to decay quickly as the distance between the two amino acids increases. A characteristic length scale of about 4 is reported for associating amino acids with the secondary structure. However, further study has shown that significantly stronger correlations exist for amino acid pairs when they are associated with key positions in the secondary structures such as the boundaries of alpha helices and beta sheets [11].

Could a similar correlation be found in the context of protein protein interactions? Due to the selection pressure during evolution on the interacting domains of folded proteins, it is reasonable to hypothesize that similar correlations may exist among interacting residues more significantly than in those that do not interact [12], [13]. In other words, the amino acids that occur at interacting residues may not be entirely independent of the amino acids occurring at other interacting residues in the protein. As will be shown later in this section, such a correlation can be shown to exist in a real dataset.

We can generalize these two examples of long-distance correlation, with the distance measured as separation of residue positions in the primary structure. A pairwise relation is defined for any two positions on a protein sequence where values of some pre-determined attribute are correlated, presumably due to some biological reasons. In both the

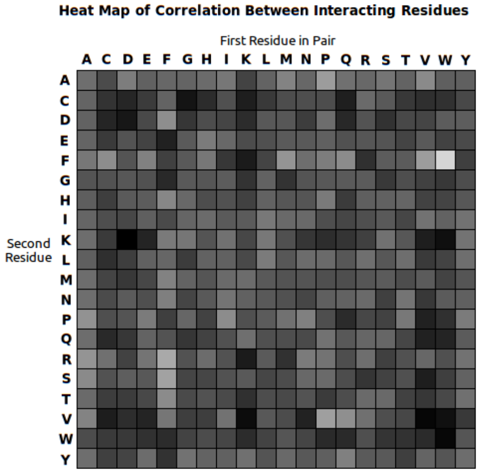

Figure 2. Heat map showing correlation of the amino acids of pairs of consecutive interacting residues.

case of secondary structure and protein protein interactions, the attribute whose values are correlated is the amino acid at those positions. For the convenience of discussion, let us call the set of all such position pairs as R. To quantify this relation, we define $S_R(x, y)$ to be the log-odds score of a pair of residues $(x, y) \in R$, representing their correlation.

$$S_R(x, y) = \log\left(\frac{p(x, y)}{p(x)\,p(y)}\right) \tag{1}$$

where $p(x, y)$ is the probability of the amino acid x appearing in the first position of a pair in R and amino acid y appearing in the second position of the pair. And $p(x)$ is the probability of amino acid x appearing in any pair in R, independent of whether it appears in the first or second position. For brevity, and without confusion, we will refer to $S_R(x, y)$ as simply $S(x, y)$ throughout the rest of this paper.

In the case of interacting residues, we define R to be the set of all pairs of consecutive of interacting residues. For example, if a protein has 3 interacting residues, which we'll label a, b, and c, and their indices on the primary structure sequence are 7, 13, and 22 respectively, then R contains the pairs (a, b) and (b, c), but not (a, c) since those two interacting residues are not consecutive.

The log-odds scores, calculated from our protein interaction data set, for the 20 x 20 pairs of amino acids can be presented as a heatmap, shown in Figure 2. The darker squares represent higher scores, and therefore the corresponding amino acid pairs are more likely to occur at that pair of interacting residues. As can be seen clearly, there are certain pairs of amino acids that are more likely to occur than others between interacting residues, implying a correlation between them. In other words, seeing a certain amino acid at one interacting residue can give us information about the amino acid that might be at the preceding or the following interacting residue in the sequence.

It should be noted that observing the existence of correlation is different from being able to utilize it in prediction. The particular difficulty in our case is that the correlation we are trying to leverage is conditioned on associating the amino acids with the interacting positions, which are not known *a priori*; rather, they are exactly what we want to predict for in the first place. This means that incorporating the long-distance correlation, therefore, must be done during prediction.

C. The Decoding Algorithms

In HMMs, one major task is decoding, namely, for a given sequence of symbols, find out the hidden states that emit the sequence. In our case, the symbols are amino acids and the hidden states are the annotation of the amino acids in terms of their evolutionary and interaction states: delete (D), insert (I), match-noninteracting (Mni), and match-interacting (Mi), as shown in Figure 1. For our purpose, we are mostly interested in predicting Mi positions. In order to incorporate the long-distance correlation during prediction, we develop a new decoding algorithm, which is a modification of the Viterbi algorithm. We next review the Viterbi algorithm briefly and then describe our modification.

For a given sequence of symbols, the hidden states that emit the symbols form themselves a sequence and correspond to a path in the graph provided by the model's architecture. Because a state can emit different symbols, "walking" through a path can emit different sequences of symbols. Similarly, a given sequence X of symbols can be emitted from multiple different paths, each with a probability that can be calculated as follows:

$$P(X, \pi) = a_{0\pi_1} \prod_{i=1}^{L} e_{\pi_i}(x_i) a_{\pi_i \pi_{i+1}} \tag{2}$$

where X is the sequence of symbols emitted, π is a path through the HMM and L is the length of the sequence.

In practice, the decoding problem is therefore formulated as finding the most probable path π^*, i.e.,

$$\pi^* = argmax_\pi P(X, \pi) \tag{3}$$

The standard Viterbi algorithm works by creating a dynamic programming table, where rows correspond to the states of the HMM and columns to the symbols of the input sequence, and each cell is filled with a value $v_j(i)$, which is the probability of the most probable path ending at state S_j and emitting x_i, the symbol at the position i of the input sequence X. Because the probability for a path is, as shown in Eq(2), the product of the emission probability and transition probability at each and every step in the path,

$v_j(i)$ can be constructed from the paths ending a step earlier at x_{i-1} as follows.

$$v_j(i) = e_j(x_i)max_k[v_k(i-1)a_{kj}] \quad (4)$$

where max over all possible k is to ensure that $v_j(i)$ is for the most probable path ending at position i and state j. With appropriate initialization, this formula offers a recurrence relation to calculate $v_j(i)$ for all j's and i's, i.e., filling out the dynamic programming (DP) table. And the overall most probable hidden path can be found by identifying the cell in the last column of the DP table with the maximum value and doing a traceback from there. A traceback refers to the process of following pointers stored for each cell in the DP table, which point to the state in the previous position that had the maximum probability (the k term in Eq. 4).

D. ETB-Viterbi Algorithm

It is well known that hidden Markov models are limited to capturing only correlations in a short range, mostly the nearest neighbors, not long enough to cross the span of an interacting domain to connect two interacting residues which, as shown in the previous subsection, possess a significant correlation. To incorporate long-distance pairwise correlation between interacting residues into the prediction, it's convenient to define "R-states", which are states in the HMM that correspond to the correlated sequence positions that make up the pairs in the set R. We define the R-states in the context of ipHMM as any interacting match state, i.e., any M_i state shown in Figure 1.

Ideally, when we are predicting the annotation for a position i in the sequence, we would want to identify the pair in R, of which position i is a member, so that we can use the log-odds scores ($S(x,y)$), as calculated using Eq(1), to modify the probability $v_j(i)$, should position i be annotated as an R-state. But whether or not a position belongs to a pair in R is what is being predicted, so we do not know beforehand what those pairs are.

To unravel this intertwining issue, we develop the following heuristic. Note that we have already filled out the dynamic programming table for all previous positions. Therefore, for those residues occurring before the one we are currently predicting we have already made prediction whether or not they will belong to pairs in R. We can then create a "putative R pair" by following the predicted path backwards until we find a residue predicted to align to an R-state, and pair it with the residue we are currently predicting. We call this an "early traceback", or ETB, to differentiate it from the traceback that's performed at the end of the recursion step, although the process is nearly identical.

Specifically, in the Viterbi algorithm, when a cell in the dynamic programming table for an R-state is being filled, it triggers an early traceback. Given that i is the current position in the sequence, the early traceback follows the pointers recorded at each cell along its way, and ends when it encounters another R-state at sequence position i'. The correlation for the putative R pair is measured by the log-odds score $S(x_{i'}, x_i)$ in Eq(1) and then added to $v_j(i)$.

```
Initialization:  $v_0(0) = 1, v_k(0) = 0$ for $k > 0$
Recursion:       $v_j(i) =$
                     $e_j(x_i)max_k(v_k(i-1)a_{kj})$,
                         if j is not an $R$-state
                     $e_j(x_i)max_k(v_k(i-1)a_{kj})$
                         $+\mathbf{C} * \mathbf{S}(\mathbf{x_{i'}}, \mathbf{x_i})$,
                         if $j$ is an $R$-state
                 $ptr_i(j) = argmax_k(v_k(i-1)a_{kj})$
Termination:     $P(x, \pi*) = max_k(v_k(L)a_{k0})$;
                 $\pi_L* = argmax_k(v_k(L)a_{k0})$
Traceback:       $\pi_{i-1}* = ptr_i(\pi_i*)$
```

Figure 3. The ETB-Viterbi algorithm. If the state j at current positon i is an R-state, do a traceback following the pointers recorded at each cell along its way until it encounters another R-state at sequence position i'. The correlation for the putative R pair, measured as the log-odds score $S(x_{i'}, x_i)$, is added to $v_j(i)$. See the text for details.

$$v_j(i) = e_j(x_i)max_k(v_k(i-1)a_{kj}) + C * S(x_{i'}, x_i) \quad (5)$$

where C is a constant to regularize the contribution of S. Since the log-odds score S can be either positive or negative, depending on the amino acids seen in the putative R pair, its addition to $v_j(i)$ may be either boosting or suppressing this particular path to be picked as the most probable hidden path by the final traceback. Figure 3

It is important to note that the probability of a path, calculated by the ETB-Viterbi algorithm, can be either boosted or suppressed, from incorporating the correlation among pairs in R. This gives the reason why the log-odds score, rather than the joint probability or mutual information, should be used, because the alternatives, being non negative – and being zero only when the amino acid pair are totally independent – will almost always add to $v_j(i)$ a positive value, even when the amino acids in the putative R pair are less correlated than by chance, leading to an over-prediction of interacting residues.

The ETB-Viterbi algorithm, while generally formulated for any set R, can be easily applied to ipHMM by defining the R-states as the interacting (M_i) match states. In terms of time complexity, the ETB-Viterbi algorithm has an additional linear time term not present in the ordinary Viterbi algorithm. This linear term represents the time required to do the early tracebacks. In ipHMM, an early traceback can go as long as L steps, the length of the entire interacting domain. This traceback occurs in the algorithm when evaluating the interacting residue state, which could potentially be done at every point on the sequence. This makes the total time complexity of the early traceback $O(L^2)$ where L is the length of the domain sequence. The rest of the ETB-Viterbi algorithm has the same time complexity of the classic Viterbi algorithm, so the full time complexity of ETB-Viterbi is $O(N^2 * L) + O(L^2)$ where N is the number of states in the HMM and L is the length of the sequence. In practice, neighboring interacting residues do not span the entire domain, and can be thought of as having a statistically "maximum" length. This would change the additional ETB term of the complexity from $O(L^2)$ to $O(L)$, since the traceback is now a constant time operation. This $O(L)$ term is now overshadowed by the original Viterbi algorithm's complexity, so the two algorithms are asymptotically equivalent.

III. Results and Discussion

A. Data

For training and testing our method, we used a set of domain-domain interactions (DDIs) extracted from the 3D Interacting Domains (3DID) database [14]. Each DDI is a family with interacting domains, *Dom. A* and *Dom. B*, whose members are I pairs of interacting proteins that have been found to physically interact, or in other words, there are I protein complexes with interacting proteins through the domain-domain interface and with distinct pdbid in the Protein Data Bank. We used DDIs with I equal 10 or 11 and where the domain length (number of match states) is smaller than 300. These criteria look for families rich enough in information content for ensuring statistical robustness, but are not too big to avoid prohibitive processing times. With these filtering parameters, 88 DDIs were selected. For every protein sequence that is part of a single DDI, the 3DID database provides information to build binary vectors with the same length of the proteins and where the 1s indicate interacting amino acids. These vectors and the profile hidden Markov models of each domain, extracted from Pfam [2], are used to create an ipHMM for *Dom. A*.

The heatmap shown in Figure 2 was generated using this data. Each protein sequence in the data set was scanned, extracting every pair of consecutive interacting residues. Those pairs were then used to calculate the log-odds scores for all possible amino acid pairs as described in Eq(1), which are represented as the degree of shade of each cell in the heatmap. The probabilities needed for the calculation were estimated from the extracted pairs, for example $p(x, y)$ in the case where x is alanine and y is leucine was determined by the proportion of interacting residue pairs with alanine in the first position and leucine in the second position to the total number of interacting residue pairs in the data.

B. Simulated Data

To further test the effectiveness of our proposed ETB-Viterbi algorithm, we also created simulated data in order to investigate how long-distance information can affect the prediction accuracy for the ETB-Viterbi algorithm as compared to the standard Viterbi algorithm with controllable levels of correlation between interacting residues. The simulated data was created by starting with the real 3DID dataset we used, and applying a set of schemes to change the residues at interacting locations to result in correlations of different strengths. For each protein in the dataset, we located the first interacting residue and noted its amino acid, then moved to the next interacting residue and changed its amino acid based on the amino acid of the first. By using the correlation scores calculated from the real data, we changed the residue to either the single most probable amino acid given the previous interacting residue, or we randomly selected an amino acid from the top 3 or 5 most probable. We repeated the process for every interacting residue pair in each protein, creating 3 different sets of simulated data. These simulated datasets will be referred to as the Top 1, Top 3, and Top 5 datasets, with the number indicated which scheme was used to perturb the real data.

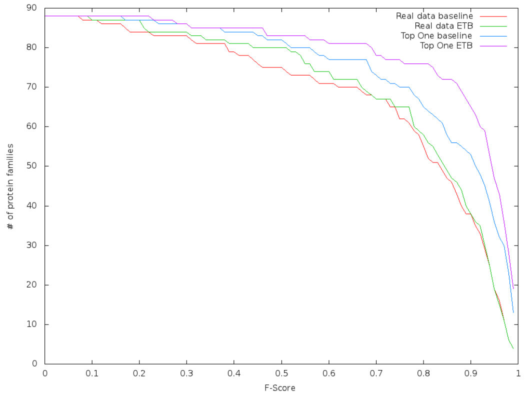

Figure 4. The improvement made by the ETB-Viterbi algorithm on the real data and the Top 1 simulated dataset.

C. Prediction Results

Figure 4 shows the results from running standard Viterbi (labeled as baseline) and ETB-Viterbi on the real data. The y-axis of these plots is the number of protein families that achieve prediction accuracy equal to or above the F-score indicated on the x-axis. Therefore, the higher a curve is, the better the corresponding method is. The F-score was calculated as follows.

$$2 * \frac{precision * recall}{precision + recall} \tag{6}$$

where precision is the proportion of predicted positives that were true positives, and recall is the proportion of true positives that were correctly predicted. A true positive in this context is an interacting residue.

To further quantify the overall performance, we calculate the area under the curve (AUC) for the right half of the graph, i.e., where the x-axis is above 0.5. The left region of the graph is omitted for two reasons. First, predictions with a F-score below 0.5 are not practically useful. Second,

Table I
RESULTS OF RUNNING THE ETB-VITERBI ALGORITHM WITH VARIOUS DATASETS AND CONFIGURATIONS. SEE TEXT FOR DETAILS.

	Real Data	Top Five	Top Three	Top One
Averaged	0.732	0.833	0.847	0.878
Individual	0.759	0.829	0.846	0.858
Standard Viterbi	0.732	0.773	0.781	0.789
Best Improvement	3.69%	7.76%	8.45%	11.28%
p-value	0.00042	1.29e-12	4.33e-9	3.69e-9

also because F-score 0.5 or lower is such a low standard easily achievable, there is not much change to that area with the different datasets and methods. Using this measurement of AUC, the performances for various tests are shown in Table I. It can be seen that ETB-Viterbi outperforms the standard Viterbi by 3.69 percentage points in the real data. Additionally, a 1-tailed t-test on the F-scores from standard Viterbi compared to ETB-Viterbi using the real data results in a very significant p-value of 0.00042.

Note that the calculation of the log-odds score can be done over all proteins in the data set – the “Averaged” section in the table – or individually for each protein family – the “Individual” section. Individual families may have unique correlations that would be diluted if measured over all families, but a single family may not have sufficient data to create accurate log-odds scores.

To get a better understanding of how ETB-Viterbi utilizes long-distance information, we ran the same experiments using the simulated data sets as described in the previous section. Figure 4 shows the difference between ETB-Viterbi and standard Viterbi on the Top One data set, which shows a much larger improvement than on the real data, indicating the stronger correlation is being properly captured by ETB-Viterbi. The improvement over all the simulated data is shown in Table I, and can be seen increasing consistently as the correlation increases in the simulated data, reaching a 11.28% improvement for the Top One data set. All these results were obtained using a default C value of 1 in Eq(5).

IV. CONCLUSION

We have presented a novel decoding algorithm, ETB-Viterbi, for HMMs which can take advantage of correlations between pairs of symbols that may be far apart in the input sequence. This long-distance correlation is shown to be present among interacting residues in protein sequences and its detection and incorporation by the new decoding algorithm is useful in improving the predictions produced by the ipHMMs. Simulated data with modulated correlation signal further illustrates the effectiveness of the method.

V. ACKNOWLEDGEMENTS

This project was supported by the Delaware INBRE program, with grants from the National Center for Research Resources - NCRR (5P20RR016472-12) and the National Institute of General Medical Sciences - NIGMS (8 P20 GM103446-12) from the National Institutes of Health.

REFERENCES

[1] P. Uetz et al., “A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae,” *Nature*, vol. 403, pp. 623–627, 2000.

[2] R. D. Finn, J. Mistry, B. Schuster-Bockler, S. Griffiths-Jones, V. Hollich, T. Lassmann, S. Moxon, M. Marshall, A. Khanna, R. Durbin, S. R. Eddy, E. L. L. Sonnhammer, and A. Bateman, “Pfam: clans, web tools and services,” *Nucleic Acids Research*, vol. 34, pp. D247–D251, 2006.

[3] T. Friedrich, B. Pils, T. Dandekar, J. Schultz, and T. Muller, “Modelling interaction sites in protein domains with interaction profile hidden Markov models,” *Bioinformatics*, vol. 22, pp. 2851–2857, 2006.

[4] S. R. Eddy, “Profile hidden markov models.” *Bioinformatics*, vol. 14, no. 9, pp. 755–763, 1998.

[5] L. Rabiner and B. Juang, “An introduction to hidden markov models,” *ASSP Magazine, IEEE*, vol. 3, no. 1, pp. 4 – 16, jan 1986.

[6] R. Durbin, S. Eddy, A. Krogh, and G. Mitchinson, *Biological sequence analysis*. Camrbidge, UK: Cambridge University Press, 1998.

[7] C. P. Chen and B. Rost, “State-of-the-art in membrane protein prediction,” *Applied Bioinformatics*, vol. 1, no. 1, pp. 21–35, 2002.

[8] A. Krogh, M. Brown, I. Mian, and K. Sjoelander, “Hidden markov models in computational biology. applications to protein modeling,” *Journal of Molecular Biology*, vol. 235, no. 5, p. 1501, 1994.

[9] S. R. Eddy, “Multiple alignment using hidden markov models,” *Intelligent Systems in Molecular Biology Proceedings*, 1995.

[10] G. E. Crooks and S. E. Brenner, “Protein secondary structure: entropy, correlations and prediction,” *Bioinformatics*, vol. 20, no. 10, pp. 1603–1611, 2004.

[11] A. J. Gonzalez, “Protein secondary structure: entropy, correlations and prediction (extending the work by Crooks and Brenner),” *Lab internal technical report*, 2009.

[12] F. Pazos, M. Helmer-Critterich, G. Ausiello, and A. Valencia, “Correlated mutations contain information about protein-protein interaction,” *Journal of Molecular Biology*, vol. 271, pp. 511–523, 1997.

[13] A. Gonzalez, L. Liao, and C. Wu, “Predicting ligand binding residues and functional sites using multipositional correlations with graph theoretic clustering and kernel CCA,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 9, pp. 992–1001, 2011.

[14] A. Stein, R. B. Russell, and P. Aloy, “3did: interacting protein domains of known three-dimensional structure,” *Nucleic Acids Research*, vol. 33, pp. D413–D417, 2005.