Protein Fold Recognition using a Structural Hidden Markov Model

Bouchaffra, D.; Tan, J.;
Dept. of Comput. Sci. & Eng., Oakland Univ., Rochester, MI

This paper appears in: Pattern Recognition, 2006. ICPR 2006. 18th International Conference on
Issue Date : 0-0 0
Volume : 3
On page(s): 186 - 189
ISSN : 1051-4651
Print ISBN: 0-7695-2521-0
References Cited: 12
INSPEC Accession Number: 9209965
Digital Object Identifier : 10.1109/ICPR.2006.949
Date of Current Version : 18 September 2006

ABSTRACT

Protein fold recognition has been the focus of computational biologists for many years. In order to map a protein primary structure to its correct 3D fold, we introduce in this paper a machine learning paradigm that we entitled "structural hidden Markov model" (SHMM). We show how the concept of SHMM can efficiently use the protein secondary structure during the fold recognition task. Experimental results showed that the SHMM outperforms the SVM with a 6% improvement in the average accuracy. However, because in this application the two classifiers are not correlated, therefore their combination based on the highest rank criterion boosted the SHMM average accuracy with 10%

Protein Fold Recognition using a Structural Hidden Markov Model

D. Bouchaffra (Senior Member IEEE) and J. Tan
Department of Computer Science & Engineering
131 Dodge Hall, Oakland University
Rochester, MI, 48309, USA
{bouchaff,jtan}@oakland.edu

Abstract

Protein fold recognition has been the focus of computational biologists for many years. In order to map a protein primary structure to its correct 3D fold, we introduce in this paper a machine learning paradigm that we entitled "structural hidden Markov model" (SHMM). We show how the concept of SHMM can efficiently use the protein secondary structure during the fold recognition task. Experimental results showed that the SHMM outperforms the SVM with a 6% improvement in the average accuracy. However, because in this application the two classifiers are not correlated, therefore their combination based on the highest rank criterion boosted the SHMM average accuracy with 10%.

1. Introduction

The primary structure of a protein is its linear sequence of amino acids and the location of any disulfide bridges. Each secondary structure is a stretch of a sequence of amino acids that takes on a characteristic structure in the three-dimensional space. Each protein can be considered as a tertiary structure - a sequence of secondary structures folded in a certain way in the three-dimensional space. This folding process of a protein is a global overview of the protein's energy surface [13]. It is a thermodynamically driven process. Proteins fold by reaching their thermodynamically most stable structure. However, many local and non-local interactions take part in the process, and therefore the search space of possible structures becomes enormous. The folding occurs through organizing an ensemble of structures rather than through only a few uniquely defined structural intermediates. As the protein databank grows larger, the proteins classification process and its folding prediction becomes slower and more difficult.

Computational analysis of biological data obtained in genome sequencing is essential for the understanding of cellular functions and the discovery of new drugs and therapies. Sequence-sequence and sequence-structure comparison play a critical role in predicting a possible function for new sequences. Sequence alignment is accurate in detecting relationships between proteins. However, this method is not efficient when two proteins are structurally similar, but have no significant sequence similarity. Protein fold recognition is an important approach to structure discovery that does not rely on sequence similarity. It consists of assigning an amino acid sequence of unknown structure to one of a library of target 3D structures. Understanding the protein three-dimensional structure is one of the many things we need to achieve if we were to decode the human genome or the genome of a given pathogen.

Researchers have been devising new methods to solve this problem and a lot of valuable work has been undertaken. Lawrence Hunter applied heuristic Bayesian classification to define and enumerate structural motifs present in protein macromolecular systems [8]. White et al. applied a nonlinear optimal filtering algorithm to predict a protein's tertiary structure [10]. Dubchak and his colleagues proposed a method for predicting protein folding class based on a global protein chain description and a voting process [6]. Maeda et al. proposed a classification method of protein folds using a structural transformation of one protein to another [12]. Ding et al. worked on multi-class protein fold recognition using support vector machines (SVMs) and neural networks (NNs) [5]. The SVMs approach used by Ding et al. will be compared to ours in this paper. Jason et al. built a protein classification system which depends significantly on the choice of a "good" representation of the input sequences of amino acids [14]. Though their work achieved the state-of-the-art classification performance, their methodology does not handle unknown and unlabeled data.

From all the previous work, it is worth to underscore that the interaction between secondary structures has not been fully exploited in the literature. The goal in this paper is to discover the protein fold by considering both the amino acid sequence (sequential information) and the 3D

folding of the secondary structures (structural information). *The fusion of sequential and structural information is the basis of the methodology we are proposing*. This fusion is accomplished through the structural hidden Markov model (SHMM) [2, 4, 3]. The core of SHMM is based on the notion of *local structure*. The whole pattern is a sequence of structures. A local structure may have different representations. It can be captured by production rules, classes of equivalence, or any other clustering scheme.

2. Structural HMM

The concept of SHMM emphasizes the relations between parts (eg. secondary structures of a protein) of an entity and the whole [3, 4]. *Our idea is that a complex pattern* $O = o_1, o_2, \ldots o_T$ *can be viewed as a sequence of constituents* O_i *made of strings of symbols* $o_i \in \Sigma$ *interrelated in some way*. Each O_i is assigned to a local structure C_j. A SHMM is then defined as follows.

Definition 2.1 *A structural hidden Markov model is a quintuple* $\lambda = [\pi, \mathcal{A}, \mathcal{B}, \mathcal{C}, \mathcal{D}]$*, where:* π *is the initial hidden state probability vector,* $\mathcal{A}$ *is the hidden state transition probability matrix,* $\mathcal{B}$ *is the hidden state conditional probability matrix of the visible observations,* $\mathcal{C}$ *is the posterior probability matrix of a structure given a sequence of observations, and* $\mathcal{D}$ *is the structure transition probability matrix.*

An SHMM is characterized by the following elements:

- **N**, the number of hidden states in the model. We label the individual states as 1, 2, . . . , N, and denote the state at time t as q_t.
- **M**, the number of distinct observations o_i
- π, the initial hidden state distribution, where $\pi_i = P(q_1 = i)$ and $1 \leq i \leq N$, $\sum_i \pi_i = 1$.
- $\mathcal{A}$, the hidden state transition probability distribution matrix, $\mathcal{A} = \{a_{ij}\}$, where $a_{ij} = P(q_{t+1} = j \mid q_t = i)$ and $1 \leq i, j \leq N$, $\sum_j a_{ij} = 1$.
- $\mathcal{B}$, the hidden state conditional probability matrix of the observations, $\mathcal{B} = \{b_j(k)\}$, in which $b_j(k) = P(o_k \mid q_j)$, $1 \leq k \leq M$ and $1 \leq j \leq N$, $\sum_k b_j(k) = 1$.
- **F**, the number of distinct local structures.
- $\mathcal{C}$ is the posterior probability matrix of a structure given its corresponding observation sequence, $\mathcal{C} = P(C_j \mid O_i) = c_i(j)$. For each particular input string O_i, we have: $\sum_j c_i(j) = 1$.
- $\mathcal{D}$, the structure transition probability matrix. $\mathcal{D} = \{d_{ij}\}$, where $d_{ij} = P(C_{t+1} = j \mid C_t = i)$, $\sum_j d_{ij} = 1, 1 \leq i, j \leq F$.

Unlike the traditional HMM, the SHMM has two additional matrices that convey structural information. Figure 1 depicts a graphical representation of a structural hidden Markov model.

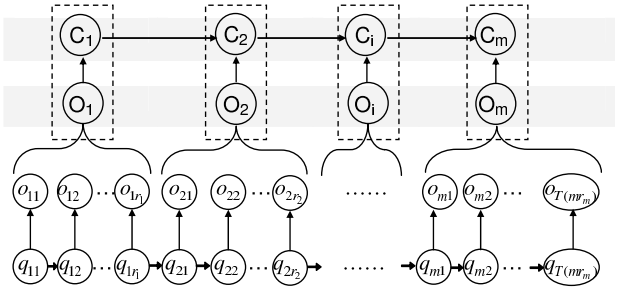

Figure 1. A graphical representation of a structural hidden Markov model.

The evaluation problem in SHMM consists of evaluating the probability for the model $\lambda = [\pi, \mathcal{A}, \mathcal{B}, \mathcal{C}, \mathcal{D}]$ to produce the sequence O. This probability can be expressed as:

$$P(O \mid \lambda) = \sum_C P(O, C \mid \lambda) = \sum_C \Phi \times \sum_q \Psi, \quad (1)$$

where $\Phi = \prod_{i=1}^{s} \frac{c_i(i) \times d_{i-1,i}}{P(C_i)}$,

and $\Psi = \pi_{q_1} b_{q_1}(o_1) a_{q_1 q_2} b_{q_2}(o_2) \ldots a_{q_{(T-1)} q_T} b_{q_T}(o_T)$. The structural decoding problem consists of determining the optimal structure sequence $C^* = < C_1^*, C_2^*, \ldots, C_t^* >$ such that: $C^* = arg \max_C P(O, C \mid \lambda)$.

In Figure 2, the amino acid sequence of protein 2DKB is O, the local structures C_j were determined through an equivalence relation defined on the set of subsequences O_i. The secondary structures of a protein are the local structures C_i assigned to O_i.

3. Experiment

In this section, we discuss data collection, the training and testing phases. We also report the results obtained.

3.1. Data Collection

The dataset that we used during the experiment was obtained from the SCOP (Structural Classification of Proteins) database. It is the PDB-40D set developed by the authors of SCOP database [11]. This data set has also been used by Ding and his colleagues [5]. As outlined in the introduction, one of our goals is to compare the approach taken by Ding's team with ours. The features they used were based on statistical information on amino acids such as "composition", "transition", and "distribution". Details on these features can be found at: *"http://www.nersc.gov/∼cding/protein"*. In Ding's experiment using SVMs, the feature extraction phase did not take into account *the order* of the secondary

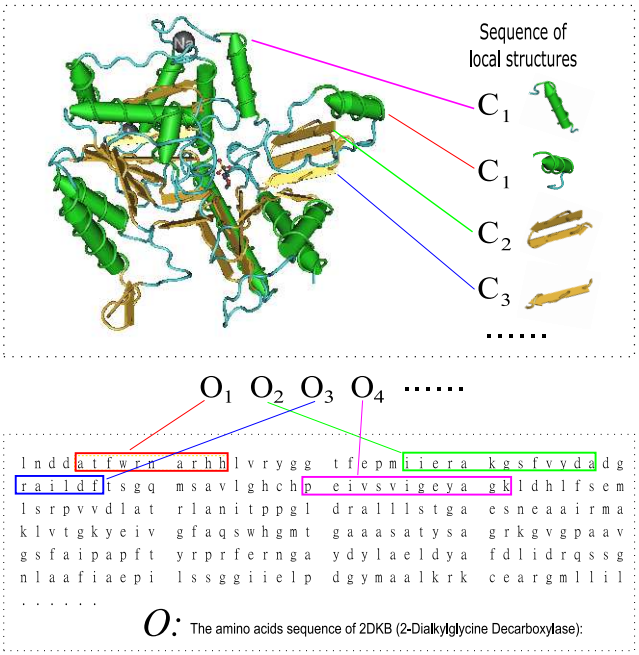

Figure 2. The 3D structure of a protein (fold) is captured by its secondary structure sequence C_j.

structures found in the whole sequence. However, in our approach it is the sequence of secondary structures that captures the whole 3D fold. *Thus, SHMM is capable to model genomic and protenomic data in a more consistent way.*

3.2. Training and Testing

Since there are 27 protein classes in the data set, therefore, we have built 27 SHMM training models λ_i. For this protein application, there are 4 types of secondary structures: "Helix", "Sheet", "Turn", and "Extended". Thus, we have fixed the number of local structures to "4" in each model. This data set contains 990 amino acid sequences. In order to measure the power of generalization of the SHMM's classifier, we used the *m-fold cross-validation* estimation technique. We divided the 990 sequences into 5 sets, each of which contains 198 sequences. Then we selected one set for testing and the other 4 sets for training. We repeated this procedure 5 times with each time selecting a different set for testing. During testing, the optimal model λ^* amongst the 27 is the one that best fits the time series sequence of amino acids. It is defined as: $\lambda^* = \arg\max_{\lambda_i} P(O \mid \lambda_i)$. The global accuracy of the SHMM is the mean of those obtained in the 5 test sets. Each amino acid sequence has been tested on all 27 SHMM models. The one who generates the highest score is the class assigned to that protein sequence.

3.3. Results and Discussion

The results depicted by Table 1, shows that for some protein classes, SHMM performed better than SVM. However, for some other protein classes, SHMM has been outperformed by SVM. When we picked some protein samples from the data set and examined their amino acid sequences, we found out that for those protein classes on which SHMM performed better, their O_i sequence tend to have long subsequences of the same secondary structure. For example, SHMM performed successfully on a protein with a secondary structure sequence like $C_1C_1C_1C_1C_1C_1...C_2C_2C_2C_2...$, and less successfully on secondary structure sequence like $C_1C_2C_1C_2C_2C_3C_2...$. This erratic behavior will be investigated in our future work.

In order to exploit the strengths of SHMM and SVM simultaneously, we combined the results of both classifiers. There has been an extensive amount of research that prove that in most practical cases, a combination of classifiers performs better than a single classifier [1, 7, 15, 9]. *A multiclassifier system is a powerful solution to difficult pattern recognition problems involving large class sets and noisy input*. We have adopted "the highest rank strategy" to determine the final classification results. We assumed both SHMM and SVM have the same weight in decision making. The highest score assigned to an amino acid sequence using SHMM is compared to the highest score using SVM. Then, the maximum score is selected as the classification result. Table 1 shows the results of both classifications and the combined result. The relative improvements of the combined classifier over SVM and SHMM are shown in Table 2.

4. Conclusion and Future Work

In this paper, we have proposed the Structural HMM model as a novel machine learning paradigm that fits seamlessly the protein fold recognition application. We have applied the concept of SHMM in order to exploit the relations between the secondary structures of a protein. This information is vital for the recognition of a protein 3D fold. We combined the classification results of SHMM with those of SVM in order to build a multiclassifier system. Although, the SHMM produced better results than the SVM in the average, the combined classification has outperformed both models when used separately.

It is worth to outline that the incorporation of topological features within the actual SHMM will strengthen the model significantly. This objective will be part of our future work.

Fold Class	SVM	SHMM	Combined	SHMM-SVM
1	87.5	83.3	87.5	-4.2
3	50.9	77.8	88.9	26.9
4	43.7	35.0	50.0	-10.4
7	53.5	100.0	100.0	46.5
9	69.8	50.0	77.8	-19.8
11	50.0	66.7	66.7	16.7
20	48.6	56.6	59.1	8.0
23	15.3	33.3	33.3	18.0
26	46.8	34.7	61.5	-12.1
30	25.0	33.3	33.3	8.3
31	41.9	50.0	75.0	8.1
32	27.4	26.0	42.1	-1.4
33	50.0	75.5	50.0	25.5
35	25.0	25.0	50.0	0.0
39	39.3	50.0	71.4	10.7
46	60.5	50.0	60.4	9.5
47	56.9	58.3	66.7	1.4
48	29.5	34.7	38.4	5.2
51	31.2	30.0	48.1	-1.2
54	47.2	60.0	60.0	12.8
57	25.0	75.0	50.0	50
59	39.3	35.7	35.7	-3.6
62	78.6	85.7	85.7	7.1
69	25.0	50.0	100.0	25.0
72	25.0	50.0	75.0	25.0
87	24.5	33.3	44.4	7.8
110	69.3	33.3	51.8	-36.0
Average	**45.2**	**51.6**	**61.6**	**6.4**

Table 1. Prediction accuracy using SVM, SHMM, and the combination of both.

Model	Improvement in %
$\frac{Combination-SVM}{SVM}$	36.3
$\frac{Combination-SHMM}{SHMM}$	19.4

Table 2. The relative improvement of the combination of SVM with SHMM over SVM and SHMM alone.

References

[1] D. Bouchaffra and V. Govindaraju. A methodology for mapping scores to probabilities. *IEEE Transactions: Pattern Analysis and Machine Intelligence*, 21(9), 1999.

[2] D. Bouchaffra and J. Tan. Introduction to the concept of structural hmm: Application to mining customers' preferences for automotive designs. In *The 17th International Conference on Pattern Recognition (ICPR)*, Cambridge, UK, August 2004.

[3] D. Bouchaffra and J. Tan. Introduction to structural hidden markov models: Application to handwritten numeral recognition. *Intelligent Data Analysis Journal*, 10(1), 2006.

[4] D. Bouchaffra and J. Tan. Structural hidden markov models using a relation of equivalence: Application to automotive designs. *Data Mining and Knowledge Discovery Journal*, 12(1), 2006.

[5] C. H. Ding and I. Dubchak. Multi-class protein fold recognition using support vector machines and neural networks. *Bioinformatics*, 17(4):349–358, 2001.

[6] I. Dubchak, I. Muchnik, S. Holbrook, and S. Kim. Prediction of protein folding class using global description of amino acid sequence. *Proc. Natl Acad Sci*, (92):8700–4, 1995.

[7] G. Fumera and F. Roli. A theoretical and experimental analysis of linear combiners for multiple classifier systems. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(6):942 – 956, Jun 2005.

[8] L. Hunter and D. J. States. Bayesian classification of protein structural elements. *Hawaiian International Conference on Systems Science*, 1992.

[9] P.-T. Jia, H.-C. He, and W. Lin. Decision by maximum of posterior probability average with weights: A method of multiple classifiers combination. *Machine Learning and Cybernetics, 2005. Proceedings of 2005 International Conference on*, 4:1949 – 1954, Aug 2005.

[10] W. JV, S. CM, and S. TF. Protein classification by stochastic modeling and optimal filtering of amino-acid sequences. *Math Biosci*, 119(1):35–75, Jan. 1994.

[11] L. Lo Conte, B. Ailey, T. hubbard, S. Brenner, A. G. Murzin, and C. Chothia. Scop: a structural classification of proteins database. *Nucleic Acids Res.*, (28):257–259, 2000.

[12] T. Maeda, K. Kamada, N. T.Ohkawa, H. Nakamura, , and A.Kidera. Feature extraction of protein folds based on secondary structure transformation. *Proc. IEEE International Joint Symposia on Intelligence and Systems*, pages 158–162, May 1998.

[13] J. N. Onuchic. Theory of protein folding: The energy landscape perspective. *Annu. Rev. Phys. Chem.*, (48):545–600, 1997.

[14] J. Weston, C. Leslie, E. Ie, D. Zhou, A. Elisseeff, and W. S. Noble. Semi-supervised protein classification using cluster kernels. *Bioinformatics*, 21(15):3241–3247, 2005.

[15] L.-Y. Yang, Z. Qin, and R. Huang. Design of a multiple classifier system. *Machine Learning and Cybernetics, 2004. Proceedings of 2004 International Conference on*, 5:3272 – 3276, Aug 2004.