A Hidden Markov Model applied to the analysis of protein 3D-structures

AC. Camproux[1], F. Guyon[1], R. Gautier[2], J. Laffray[1], and P. Tufféry[1]

[1] Equipe de Bioinformatique Génomique et Moléculaire, INSERM U726,
Université Paris 7, case 7113, 2 place Jussieu, 75251 Paris, France
(e-mail:
camproux@ebgm.jussieu.fr,guyon@ebgm.jussieu.fr,tuffery@ebgm.jussieu.fr)

[2] Institut de Pharmacologie Moleculaire et Cellulaire
UMR 6097 CNRS / UNSA
660 route des Lucioles
06560 Sophia Antipolis
(e-mail: gautier@ipmc.cnrs.f)

Abstract. Understanding and predicting protein structures depends on the complexity and the accuracy of the models used to represent them. We have setup a Hidden Markov Model to optimally compress three dimensional (3D) conformation of protein into a structural alphabet, i.e. a library of exhaustive and representative states (describing short fragments) learnt simultaneously with connection logic. The discretization of protein backbone local conformation as a series of states results in a simplification of protein 3D coordinates into a unique unidimensional (1D) representation. We present some evidence that such approach can constitute a very relevant way to the analysis of protein architecture in particular for protein structure comparison or prediction.
Keywords: Hidden Markov Models, structural alphabet, protein structural organization.

1 Introduction

The recent genome sequencing projects [Waterston *et al.*, 2002] have provided sequence information for large number of proteins. In most cases, an accurate 3D structural knowledge of the proteins is necessary for a detailed functional characterization of these sequences. However, even in the days of high-throughput methods, experimental determination of protein structures by X-ray crystallography or NMR is quite time-consuming. Thus, there is an increasing gap between the number of available protein sequences and experimentally derived protein structures, which makes it even more important to improve the methods for predicting protein 3D structures. The structural biology community has long focused on the very hard task of developing algorithms for solving the ab initio protein folding problem - namely, predicting protein structure from sequence. In its initial phase, the exploration of protein structure consisted in simplifying the 3D structure into secondary structures, included the well-known repetitive and regular zone - the α-helix (30%)

of protein residues and the β-sheet (20%). The remaining elements constitute a category, often considered as variable (50% of the structures). With the increasing of available 3D structures of proteins, many studies [Unger *et al.*, 1989],[Rooman *et al.*, 1990],[de Brevern *et al.*, 2000],
[Micheletti *et al.*, 2000],[Kolodny *et al.*, 2002] have focused on the identification of a more detailed but finite set of generic protein fragments. Despite the fact that such libraries provide an accurate approximation of protein conformation, their identification teaches us little about the way protein structures are organized. They do not consider the rules that govern the assembly process of the local fragments to produce a protein structure. An obvious mean of overcoming such limitations is to consider that the series of representative fragments that can describe protein structures are in fact not independent but governed by a Markovian process. For this purpose we have used Hidden Markov Models (HMM). HMM have been applied in several area of computational biology, for example to model protein families, to construct multiple sequence alignment or to determine protein domain in a query sequence [Krogh *et al.*, 1994],[Durbin *et al.*, 1998],[Bateman *et al.*, 2004]. In this study, we apply HMM to identify a library of representative fragments and their transition process, called Structural Alphabet (SA) or HMM-SA. Such an approach can constitute a very relevant way to the analysis of protein architecture in particular for protein structure comparison or prediction.

2 Materials and Methods

2.1 Datasets and describing three dimensional conformations

The extraction of SA is performed from a collection of 1429 non-redundant protein structures presenting less than 30% sequence identity. The structures are described using the α_Carbons (Figure (a.1)), as series of overlapping fragments of 4 residue length (Figure (a.2)) [Camproux *et al.*, 1999]. Each fragment h is described by a 4-descriptors vector $y(h)$ with the three distances between the non consecutive α_Carbons, i.e. $d_1(h)$= d{$C_{\alpha 1}$(h) - $C_{\alpha 3}$(h)}, d_2(h)=d{$C_{\alpha 1}$(h) - $C_{\alpha 4}$(h)}, d_3(h)=d{$C_{\alpha 2}$(h) - $C_{\alpha 4}$(h)}, where $C_{1,..,4}$ denotes the 4 residues of fragment h, and the oriented projection $P_4(h)$ of the last alpha-carbon $C_{\alpha 4}$(h) to the plane formed by the three first ones, as shown in Figure (a.3). The collection of 1429 proteins represent a total of 332493 four-residues fragments.

2.2 Identification of the optimal structural alphabet (SA)

Models
Suppose that polypeptidic chains are made up of representative fragments of (R) different types $\{S_1, S_2, ..., S_R\}$. We then assume that there are (R) states of the model. Each state is associated with a multi-normal function

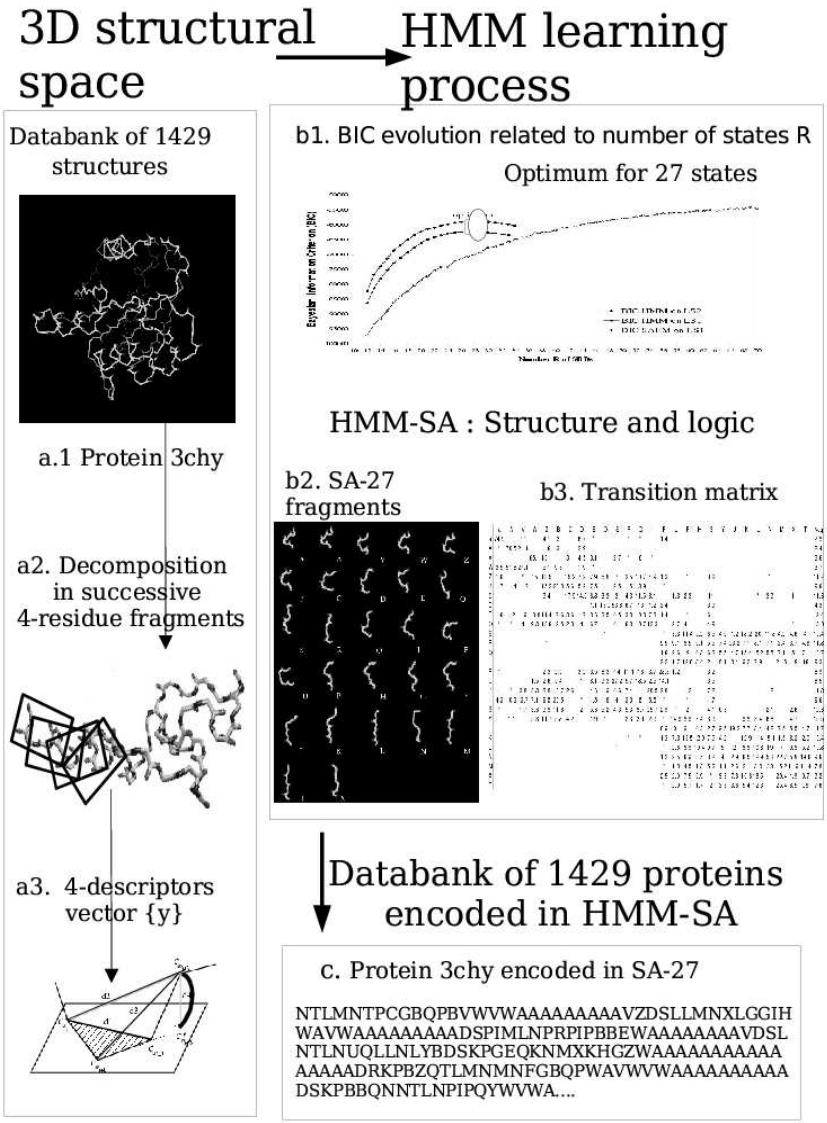

Fig. 1. Encoding of 3D conformation of proteins using HMM-SA with 27 states: right part "3D structural space" represents the polypeptidic chain of protein 3chy (a1) scanned in overlapping windows that encompassed 4 successive-carbons C_α (a2), thereby producing a series of four-residue fragments. Each fragment is described by a vector of four-descriptors (a3). Center part: Figure b1 represents the BIC evolution *versus* the number of states considered, Figures b2 and b3 illustrate the optimal HMM-SA corresponding to both 27 average four-residue fragments associated to 27 states and transition matrix between states. Bottom part represents the corresponding encoded chain 3chy (c1) as a states series.

of parameters θ describing the descriptors and their variability. We consider two types of model to identify a SA corresponding to R states: a process without memory or a process with memory of order 1.
(i) Model without memory *(order 0)*, assuming independence of the R states is identified by training simple finite Mixture Models (MM) of R multi-normal distributions.
(ii) Model with memory *(order 1)* is identified, by training a Hidden Markov Model. Here, the aim is to a learn hidden sequence of states. The succession of underlying states $\{x_1, x_2, ..., x_N\}$ emits the series of vectors $\{y_1, y_2, ..., y_N\}$, describing consecutive overlapping fragments of the proteins, *via* a multi-normal density $b_{S_i}(y)$ of parameters θ_i associated to each state $S_{i,1\leq i\leq R}$. We assume a common state dependence process for all polypeptidic chains governed by a Markov chain. The evolution of the Markov chain is completely described by:
1) the law $V = V(i)$ of the initial state of each polypeptidic chain, i.e. the probability that a polypeptidic chain starts in each of the R different states
2) the matrix of transition probabilities $\Pi = (\pi_{ii'})_{1\leq i,i'\leq R}$ between R different states of the Markov chain, where $\pi_{ii'} = P(X_j = S'_i \mid X_{j-1} = S_i)$ is the probability for different proteins to evolve from state S_i to S'_i at any position j. For a given set of proteins and a given number (R) of states, unknown parameters $\lambda = (\Pi, V, \theta)$ of the selected model are estimated with an Expectation and Maximization (EM) algorithm [Baum *et al.*, 1970] applied on the complete likelihood.

Complete likelihood of N four-residue fragments $\{y_1, y_2, ..., y_N\}$ describing a protein of N+3 residue

$$V(y_1, y_2, ..., y_N|\lambda) = \sum_{\{x_1, x_2, ..., x_N\}} V(x_1) b_{x_1}(y_1) \prod_{t=1}^{N-1} \pi_{x_t x_{t+1}} b_{x_{t+1}}(y_{t+1}) \quad (1)$$

For practical details on application to protein structures, see [Camproux *et al.*, 1999].

Encoding proteins using Viterbi algorithm

Our ultimate goal is to reconstruct the unobserved (hidden) states sequence $\{x_1, x_2, ..., x_N\}$ of the polypeptide chains, given the corresponding four-dimensional vectors of descriptors $\{y_1, y_2, ..., y_N\}$, and to provide a classification of successive fragments in R states. For a given 3D conformation and a selected model (fixed number R of states), the corresponding best state sequence among all the possible paths in $\{S_1, \ldots, S_R\}^N$ can be reconstructed by a dynamic programming algorithm based on Markovian process (Viterbi algorithm [Rabiner, 1989]).

Statistical criteria to determine the optimal number of states

Structural alphabets of different size (R), noted SA$-R$ are learnt using HMM and MM by progressively increasing R and compared using Bayesian Information Criterion (BIC, [Schwartz, 1978]).

2.3 Assessing the discretization of protein structures

For a given state, the average C_α Root-Mean-Square deviation (RMSd) between C_α coordinates, that is an euclidean distance, of the fragments to their centroid is used to measure the structural dispersion of each state. To reconstruct the protein 3D structures from their description as a series of states, and to keep some comparison possible, we use the building procedure employed by Kolodny et al. [Kolodny *et al.*, 2002]. Briefly, the fragments are assembled using an iterative concatenation procedure to adjust 3D conformation.

2.4 Quantifying structure similarity

During the HMM-SA encoding of proteins of known structures, the probabilities of substituting one state for another are directly provided by the forward-backward algorithm [Rabiner, 1989]. A lod-score or substitution matrix is derived from these probabilities:

$$S(i,j) = ln[\frac{P(S_i, S_j)}{P(S_j)P(S_j)}] \tag{2}$$

which can be rewritten as

$$S(i,j) = ln[\frac{P(S_i|S_j)}{P(S_j)}] \tag{3}$$

with $P(S_i|S_j)$, the probability of letter S_i substitutes for letter S_j at one position and $P(S_j)$, the probability of state S_j (computed as the proportion of observed letter S_j). This lod-score matrix quantifying similarity between states is shightly modified. The score values S(i,j) get to $-\infty$ when the substitution of state S(i) by S(j) is impossible. All the finite values of S(i,j) are shifted and made positive, and the infinite one are replaced by large negative values.

2.5 Measuring sequence-structure consistency

Amino acid / state dependence can be learnt *a posteriori* from the database of 1429 proteins encoded in HMM-SA and the corresponding amino acid sequences. The specificity of each state in terms of amino acid is assessed using the "relative entropy" [Kullback and Leibler, 1951].
These amino acid sequence / states dependence can be used to quantify the consistency of a candidate 3D structure encoded in HMM-SA and its corresponding amino acids sequence. Emission probabilities of 20 amino acids $a_{j,1\leq j\leq 20}$ from each state $S_{i,1\leq i\leq R}$: $P(a_j|S_i)$ are introduced in the HMM

to compute the likelihood of an amino acids sequence $\{a_1, a_2, ..., a_N\}$ corresponding to a structure encoded in states sequence x=$\{x_1, x_2, ..., x_N\}$.

$$V(a_1, a_2, ..., a_N, x|\lambda) = V(x_1)P(a_1|x_1) \prod_{t=1}^{N-1} \pi_{x_t x_{t+1}} P(a_{t+1}|x_{t+1}) \qquad (4)$$

3 Results

3.1 HMM-SA validation

HMM-SA is few dependent on the learning set
We learn SA of increasing sizes using either HMM or MM, and we compare them on the basis of their goodness of fit (Figure (b.1)). The influence of the Markovian process is large, as illustrated by the very different behaviors of the BIC associated with MM_0 or HMM_1. For MM, no BIC optimum is reached until alphabet sizes of 70 whereas for HMM, an optimum is reached for a number of states of 27 (SA-27), larger than that obtained using MM, which means a better fit of the data using HMM. Interestingly, the Markov classification takes advantage of information implicitly contained in the succession of the observations to greatly reduce the number of states, keeping a minimal representativity for each (at least 1.5%). Similar results are obtained using two independent learning sets of 250 proteins with similar BIC curves evolution. The optimum is reached for 27 states in both cases, and we find that the two SA-27 very similar. It follows that, at the optimum, the HMM-derived structural alphabet (HMM-SA) is very weakly dependent on the learning set, which in turn suggests that the learnt model can be considered as representative of all protein structures.

Geometrical and logical description of the structural alphabet
The 27 identified states are denoted as structural letters: [a, A, B,..., Y, Z]. The set of letters, sorted by increasing stretches in figure (b.2) and their transitions constitute the SA The "local fit approximation" is low, as quantified by the average alpha-carbon RMSd to the centroid associated with each state (0.23 ± 0.14 Å). SA-27 shown very reasonable performance (RMSd value less than 1Å) in terms of reconstruction of the whole protein structure accuracy, compared to other recent libraries fragments optimized in a purpose of reconstruction [Micheletti *et al.*, 2000, Kolodny *et al.*, 2002]. Concerning description of logic of protein architecture, 66% of 729 transitions between states have probabilities less than 1% (see Transition matrix between 27 states in b.3), i.e.. We observe the existence of pathways between the states, that obey some precise and unidirectional rules. Looking in detail, we observe that the states associated with close shapes have different logical roles. For instance, the two closest states [A, a] in term of geometry, close to canonical alpha helix, are distinguished by different preferred input and output states.

Moreover, the learning process attempts to optimize the likelihood associated with the entire trajectories of the proteins, resulting in propagation of such long range conditioning to the short range constraints that are learnt. For instance, three major types of alpha-helices categorized as linear, kinked or curved by Kumar and Bansal [Kumar and Bansal, 1998], seem identified in HMM-SA by [AAAAAAAAAA] series, [AAAAVWAAAA] series and [aaaaaaaaaa] series. These results are detailed in [Camproux *et al.*, 2004].

3.2 HMM-SA application: HMM structural alphabet as a general concept to simplify 3D protein structure analysis ?

Discretization of 3D structural space of proteins in SA space
Subsequently, HMM-SA provides some kind of compression from the 3D protein coordinate space into the *1D structural alphabet space*, see Figure 1. We have explored two directions in which this facility could be of interest.

Categorizing structural similarity
The detection and analysis of structural similarities of proteins can provide important insights into their functional mechanisms or relationship and offer the basis of classifications of the protein folds. The global 3D alignment of two proteins is NP-hard [Lathrop, 1994]. Therefore, approximate methods have been proposed to achieve fast similarity searching, based on the direct consideration of protein alpha-carbon coordinates [Gibrat *et al.*, 1996], [Holm and Sander, 1993, Shindyalov and Bourne, 1998]. Using HMM model, the lod-score matrix of similarity between states (Eq(2)) allows to quantify the similarity of protein fragments encoded as different series of states. It is possible to use it with classical methods developed for the amino acid sequences similarity search and thus to reduce 3D searches as a 1-dimensional sequence alignment problem [Guyon *et al.*, 2004]. Although we currently obtain performance poorer than pure 3D methods, this approach can perform fast 3D similarity search such as the extraction of exact words using a suffix tree approach, or the search for fuzzy words and is very promising in a perspective of combining with prediction procedure.

Applying sequence-structure consistency measures
All the states of SA-27 have some significant amino acid sequence specificity compared to the profiles of the collection of 1429 protein fragments ("relative entropy", $p<0.001$). Ab initio prediction is commonly viewed as composed of two problems (1) generating candidate folds, called decoys ; and (2) devising a scoring function that discriminates between near native folds and other non-native folds amongst the decoys [Kolodny *et al.*, 2002]. Concerning point (2), we can use significant dependence between states and sequence (Eq(3)) to evaluate the consistency of a set of decoys encoded in SA-27 with its corresponding amino acids sequence. Preliminary results to discriminate 3D decoys proposed in CASP6 (Critical Assessment of Techniques for Protein

Structure Prediction) show some correlation with RMSd for decoys library and this work is in progress.

4 Discussion and perspectives

In the present study, we have discussed an HMM derived 27 states SA based on a Markov process of order 1. Higher order Markovian dependence could be considered, but at the cost of a much larger number of parameters, which may pose practical computational problems. HMM-SA fits well the previous knowledge related to protein architecture organization and seems able to grab some subtle details of protein organization, while using a reduced number of states. Results on dependence between letters and amino acid sequence confirms that, despite we have learnt SA using only geometric information, we have not over-split sequence information and that all states present some sequence signature. The resulting 1D representation of protein structure can be applied to a large variety of problems recurrent to the field of protein structure analysis and prediction. Here, we have presented some evidence of its relevance for categorizing structural similarity, or measuring some sequence / structure consistency. Work is under progress to enlarge this to fold classification and prediction.

References

[Bateman *et al.*, 2004]A. Bateman, R. Coin, L.and Durbin, R.V. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E. L. L. Sonnhammer, D. J. Studholme, C. Yeats, and S.R. Eddy. The pfam protein families database. *Nucleic Acids Research*, 32:138–141, 2004.

[Baum *et al.*, 1970]L. E. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *The Annals of Mathematical Statistics*, 41(1):164–171, 1970.

[Camproux *et al.*, 1999]A. C. Camproux, P. Tuffery, J. P. Chevrolat, J. F. Boisvieux, and S. Hazout. Hidden markov model approach for identifying the modular framework of the protein backbone. *Protein Eng*, 12(12):1063–73, 1999.

[Camproux *et al.*, 2004]A. C. Camproux, R. Gautier, and P. Tuffery. A hidden markov model derived structural alphabet for proteins. *J Mol Biol*, 339(3):591–605, 2004.

[de Brevern *et al.*, 2000]A. G. de Brevern, C. Etchebest, and S. Hazout. Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. *Proteins*, 41(3):271–87, 2000.

[Durbin *et al.*, 1998]R. Durbin, S.R. Eddy, and G.J. Krogh A.and Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Paris, 1998.

[Gibrat *et al.*, 1996]J. F. Gibrat, T. Madej, and S. H. Bryant. Surprising similarities in structure comparison. *Curr Opin Struct Biol*, 6(3):377–85, 1996.

[Guyon *et al.*, 2004]F. Guyon, A. C. Camproux, J. Hochez, and P. Tuffery. Sasearch: a web tool for protein structure mining based on a structural alphabet. *Nucleic Acids Res*, 32(Web Server issue):W545–8, 2004.

[Holm and Sander, 1993]L. Holm and C. Sander. Protein structure comparison by alignment of distance matrices. *J Mol Biol*, 233(1):123–38, 1993.

[Kolodny *et al.*, 2002]R. Kolodny, P. Koehl, L. Guibas, and M. Levitt. Small libraries of protein fragments model native protein structures accurately. *J Mol Biol*, 323(2):297–307, 2002.

[Krogh *et al.*, 1994]A. Krogh, M. Brown, I.S. Mian, K. Sjolander, and D. Haussler. Hidden markov models in computational biology. applications to protein modeling. *J Mol Biol.*, 235(5):1501–31, 1994.

[Kullback and Leibler, 1951]S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematics and Statistics*, 22:79–86, 1951.

[Kumar and Bansal, 1998]S. Kumar and M. Bansal. Geometrical and sequence characteristics of alpha-helices in globular proteins. *Biophys J*, 75(4):1935–44, 1998.

[Lathrop, 1994]R. H. Lathrop. The protein threading problem with sequence amino acid interaction preferences is np-complete. *Protein Eng*, 7(9):1059–68, 1994.

[Micheletti *et al.*, 2000]C. Micheletti, F. Seno, and A. Maritan. Recurrent oligomers in proteins: an optimal scheme reconciling accurate and concise backbone representations in automated folding and design studies. *Proteins*, 40(4):662–74, 2000.

[Rabiner, 1989]L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proc. of the IEEE*, 77(2):257–285, 1989.

[Rooman *et al.*, 1990]M. J. Rooman, J. Rodriguez, and S. J. Wodak. Automatic definition of recurrent local structure motifs in proteins. *J Mol Biol*, 213(2):327–36, 1990.

[Schwartz, 1978]G. Schwartz. Estimating the dimension of a model. *Annals of statistics*, 6:461–464, 1978.

[Shindyalov and Bourne, 1998]I. N. Shindyalov and P. E. Bourne. Protein structure alignment by incremental combinatorial extension (ce) of the optimal path. *Protein Eng*, 11(9):739–47, 1998.

[Unger *et al.*, 1989]R. Unger, D. Harel, S. Wherland, and J. L. Sussman. A 3d building blocks approach to analyzing and predicting structure of proteins. *Proteins*, 5(4):355–73, 1989.

[Waterston *et al.*, 2002]R. H. Waterston, E. S. Lander, and J. E. Sulston. On the sequencing of the human genome. *Proc Natl Acad Sci U S A*, 99(6):3712–6, 2002.