13

Application of HMM to the Study of Three-Dimensional Protein Structure

Christelle Reynès[1], Leslie Regad[2], Stéphanie Pérot[3], Grégory Nuel[5] and Anne-Claude Camproux[4]
[1,2,3,4]*Molécules Thérapeutiques in silico (MTi), UMR-S973 Inserm, Université Paris Diderot, Paris*
[5]*MAP5, UMR CNRS8145, Université Paris-Descartes, Paris*
France

1. Introduction

Hidden Markov models (HMM) have been successfully applied in molecular biology, especially in several areas of computational biology. For example, it is used to model protein families, to construct multiple sequence alignments, to determine protein domains in a query sequence or to predict the topology of transmembrane beta-barrels proteins (Bateman *et al.*, 2004; Durbin *et al.*, 1998; Krogh *et al.*, 1994; Pang *et al.*, 2010).

Proteins are macromolecules responsible for performing many important tasks in living organisms. The function of proteins strongly depends on their shapes. For example, carrier proteins should recognize the molecules they carry such as hemoglobins should recognize oxygen atoms, anti-bodies their antigens,... Protein misfolding may cause malfunctions such as Parkinson and Alzheimer diseases. Therefore, it is necessary to know the structure of a protein to understand its functions and the disturbances caused by the inappropriate behavior derived from misfolding. Precisely this knowledge makes it possible to develop drugs and vaccines and synthetize proteins which, for example, disable the regions of virus activity, preventing them from acting on the cells.
The exploration of protein structures, in its initial phase, consisted in simplifying three-dimensional (3D) structures into secondary structures, including the well-known repetitive and regular zone - the α-helix (30% of protein residues) and the β-sheet (20%). The remaining elements (50% of residues) constitute a category called loops, often considered as structurally variable, and decomposed into some subcategories such as turns (see Frishman & Argos, 1995, for example). Although the prediction of secondary structure types can be achieved with a success rate of 80%, the description of the secondary structures of a protein does not provide *per se* an accurate enough description to allow the characterization of the complete structure of proteins.
Moreover, protein structures are determined experimentally mostly by X-ray crystallography and nuclear magnetic resonance (NMR) techniques. Both require sophisticated laboratories, are time and cost expensive and cannot be applied to all proteins. On the other hand, the methods consisting in protein sequencing are easier and less expensive than methods

implying structure determination. The recent genome sequencing projects (Siva, 2008; Waterston *et al.*, 2002) have provided sequence information for a large number of proteins. Consequentely, Swiss-Prot database release 57.6 (Boeckmann *et al.*, 2003), a curated protein sequence database provided by the Universal Protein Knowledgebase (UniProt) consortium, contains sequences of more than 13,069,501 distinct proteins (Consortium, 2010), whereas the Protein Data Bank (PDB) (Henrick *et al.*, 2998) contains 70,000 distinct protein structures. 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 3D structure protein prediction.

Thus, an important challenge in structural bioinformatics is to obtain an accurate 3D structural knowledge about proteins in order to obtain a detailed functional characterization and a better understanding. With the increase of available 3D structures of proteins, many studies (Baeten *et al.*, 2010; Brevern *et al.*, 2000; Bystroff *et al.*, 2000; Kolodny *et al.*, 2002; Micheletti *et al.*, 2000; Pandini *et al.*, 2010; Unger *et al.*, 1989), have focused on the identification of a detailed and systematic decomposition of structures into a finite set of generic protein fragments. Despite the fact that some 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 describing protein structures are in fact not independent but governed by a Markovian process. In this chapter, the first part presents the development of a HMM approach to analyze 3D protein architecture. We present the use of a HMM to identify a library of representative fragments, called Structural Letters (SLs) and their transition process, resulting in a structural alphabet, called HMM-SA, decomposing protein 3D conformations. The aim of this part is to assess how much HMM is able to yield insights into the modular framework of proteins, i.e. to encode protein backbones into uni-dimensional (1D) sequentially dependent SLs. The HMM-SA is a very performant tool to simplify 3D conformation of proteins and such a simplification can constitute a very relevant way to analyze protein architecture. Different applications of HMM-SA for structure analysis are listed, such as loop modelling, protein structure comparison, analysis of protein deformation during protein interactions, analysis of protein contacts, detection and prediction of exceptional structural patterns.

The second part of the chapter presents a contribution to the important challenge of protein structure prediction by predicting through another HMM the presence/absence of functional patterns identified thanks to HMM-SA. The method can be decomposed into two steps: in a first time, a simple link between amino-acids (AAs) and SLs is learned through boolean functions, then, a HMM is used to take into account the results of the first step as long as the dependencies between successive SLs. The first step is independent on the studied pattern whereas a new HMM is automatically built for each new pattern. The method will be illustrated on three examples.

2. HMM-SA obtention

2.1 Datasets and description of three dimensional conformations

The data extraction of HMM-SA is performed from a collection of 1,429 non-redundant protein structures (Berstein *et al.*, 1977) extracted from the PDB. The selected proteins have a crystallographic resolution lower than 2.5 Å and less than 30% sequence identity with one another (Hobohm *et al.*, 1992). Because the structure of the model is based on local dependence

between successive residues in each protein, all non-contiguous protein chains (i.e. those containing fragments that spanned gaps) were eliminated from the dataset. The polypeptide chains were scanned in overlapping windows that encompassed four successive α-carbons (C_α), thereby producing a succession of short-backbone chain fragments. As in some previous studies (Pavone *et al.*, 1996; Rackovsky, 1993; Rooman *et al.*, 1990; Smith *et al.*, 1997), we used four-residue lengths, which contain enough information to find basic structural elements: four-residue turns for α-helices, bridges for β-sheets and undefined loop structures. Moreover, a four-residue segment is small enough to keep the number of SL categories reasonable. Increasing the number of residues per segment would introduce larger variability and would lead to a larger number of categories.

The collection of 1,429 proteins represents a total of 332,493 four-residue fragments. Protein structures are described using the distances between C_α, see Figure 1a, as series of overlapping fragments of four-residue length (Camproux *et al.*, 1999a). Let $C_{\alpha_1}, C_{\alpha_2}, \ldots, C_{\alpha_n}$ be the n carbon atoms of the backbone structure of the protein. From these data, we build a sequence $X_0, X_1, \ldots, X_{n-4}$ such as $X_i = (X_i^1, X_i^2, X_i^3, X_i^4) \in \mathbb{R}^4$ with:

$$X_i^1 = ||\overrightarrow{C_{\alpha_{i+1}} C_{\alpha_{i+3}}}||; \quad X_i^2 = ||\overrightarrow{C_{\alpha_{i+1}} C_{\alpha_{i+4}}}||; \quad X_i^3 = ||\overrightarrow{C_{\alpha_{i+2}} C_{\alpha_{i+4}}}||;$$

$$X_i^4 = \frac{\overrightarrow{C_{\alpha_{i+1}} C_{\alpha_{i+2}}} \wedge \overrightarrow{C_{\alpha_{i+2}} C_{\alpha_{i+3}}}}{||\overrightarrow{C_{\alpha_{i+1}} C_{\alpha_{i+2}}} \wedge \overrightarrow{C_{\alpha_{i+2}} C_{\alpha_{i+3}}}||} \times \overrightarrow{C_{\alpha_{i+3}} C_{\alpha_{i+4}}}.$$

The three values X_i^1, X_i^2, X_i^3 correspond to distances between the non consecutive $(C_{\alpha_1}, C_{\alpha_2}, C_{\alpha_3}, C_{\alpha_4})$ and X_i^4 the oriented projection of the last α-carbon C_{α_4} onto the plane formed by the three first ones, as shown in Figure 1a. The three distances between consecutive C_α are not considered in this approach because fewly variable.

The first and third distances (X_i^1, X_i^3) describe the opening of the beginning and end of a four-residue fragment and X_i^2 describes the global length of this fragment. X_i^4 is proportionnal to the determinant of the three vectors defined by the successive C_α pairs normalized by the norm or modulus of the first two vectors. This descriptor is proportional to the distance of the four C_α to the plane P built by the first three C_α. The sign of X_i^4 indicates the topological orientation of the fragment relative to P : trigonometric, i.e. the fourth α-carbon C_{α_4} is located above P (for a positive value of X_i^4) and inverse trigonometric, i.e. the fourth α-carbon C_{α_4} is located below P (for a negative value of X_i^4). X_i^4 not only gives the direction of the fragment fold but also provides direct interpretable information about the volume of the fragment. A *flat* fragment, i.e. with no volume, corresponds to a value of X_i^4 close to 0.

2.2 HMM modelling

In the first work, the idea was to consider the model of Figure 2 where all X_i are generated independently from each other conditionally to one out of the K hidden states $S_i \in \mathcal{S} = \{1, 2, \ldots, K\}$ such as:

$$\mathcal{L}(X_i | S_i = r) \sim \mathcal{N}(\mu_r, \Sigma_r) \quad \forall r \in \mathcal{S} \tag{1}$$

with $\mathcal{N}(\mu_r, \Sigma_r)$ four-dimensional multi-normal density with parameters (μ_r, Σ_r) describing the average descriptors, the variability of each descriptor and the covariance between descriptors as estimated on the associated fragments. We additionally consider two types of model to identify a structural alphabet corresponding to K SLs: (i) a simple process without memory or (ii) a HMM process with memory of order 1.

Model without Memory, denoted MM(order 0), assuming independence of the K SLs, is

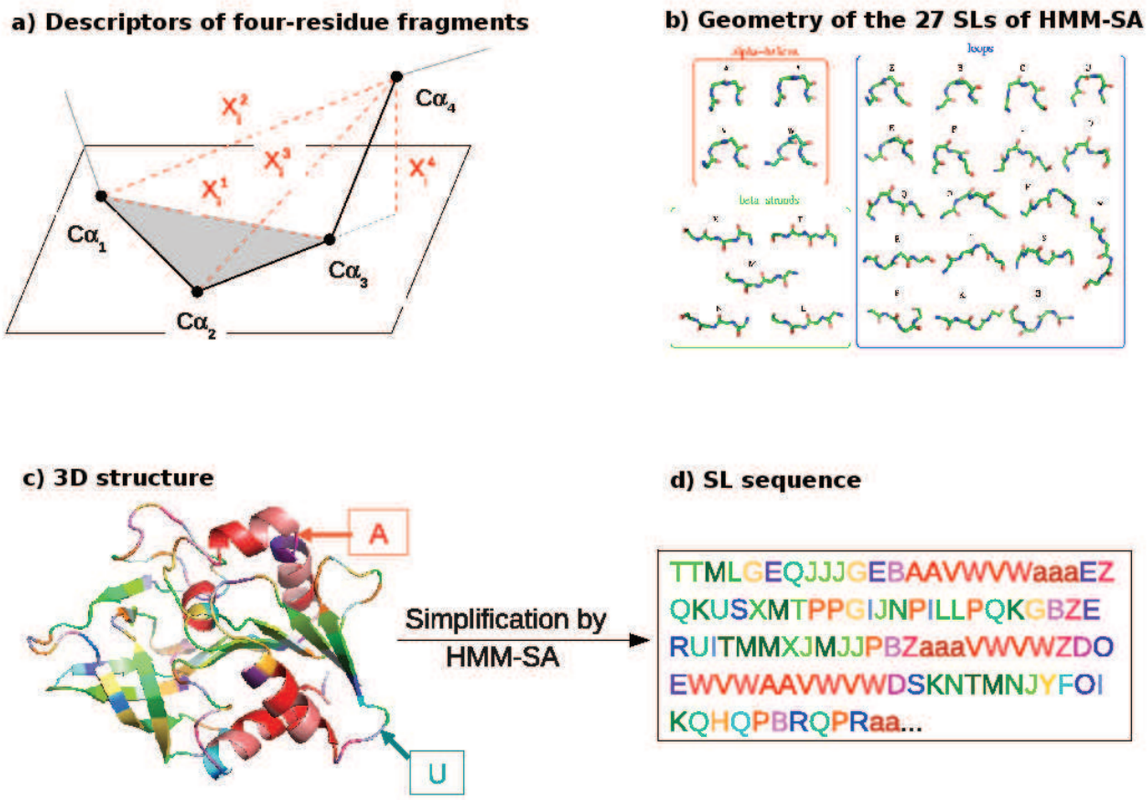


Fig. 1. Encoding of 3D conformation of proteins using HMM-SA with 27 SLs. (a) Representation of the four descriptors $(X_i^1, \ldots, X_i^4)$, used to describe the 3D conformation of four successive C_α fragments. (b) Geometry of the final 27 SLs, ranked by corresponding secondary structures. (c) 3D representation of the B chain of protein 1gpw colored according to its SL encoding. (d) Final 1D SL encoding of the B chain of protein 1gpw.

identified by training simple finite mixture of four-dimensional multi-normal densities. Model assuming that the sequence (S_i) is distributed according to a homogeneous Markov chain with starting distribution ν and transition matrix τ is defined by:

$$\mathbb{P}(S_1 = r) = \nu(r) \quad \text{and} \quad \mathbb{P}(S_{i+1} = s | S_i = r) = \tau(r, s) \quad \forall r, s \in \mathcal{S}. \tag{2}$$

It hence results in a model with $(14K + K^2 - 1)$ parameters.

2.3 Model selection: Statistical criteria to determine the optimal number of SLs

The classical model selection approach is based on the parsimony principle: we want to select the model that better fits the data with the smallest possible complexity. This typically leads to penalized likelihood criteria like the Bayesian Information Criterion (BIC, Schwartz, 1978) which balances the log-likelihood of the model and a penalization term related to the number of parameters of the model and the sample size. In the first work, structural alphabets of

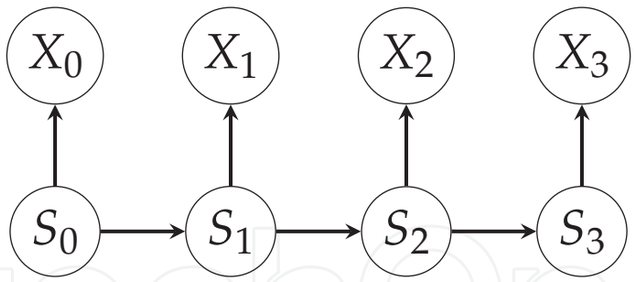


Fig. 2. Graph of dependencies in the simplest model with $n = 7$ C_α hence resulting in a total of $n - 3 = 4$ SLs.

different sizes K, denoted SA$-K$ were learned on two independent learning sets of proteins by progressively increasing K, and using the two types of model detailed in section 2.2 (with or without memory) and compared using BIC.

2.4 Encoding proteins

One ultimate goal is to reconstruct the unobserved (hidden) SL sequence of the polypeptidic chains, given the corresponding four-dimensional vectors of descriptors, and to provide a classification of successive fragments in K SLs. For a given 3D conformation and a selected model (fixed number K of SLs), the corresponding best SL sequence among all the possible paths in $\mathcal{S}^n = \{1, 2, \ldots, K\}^n$ can be reconstructed by a dynamic programming algorithm based on Markovian process.

Once a model has been selected and its parameters estimated, we classically use the Viterbi's algorithm (Rabiner, 1989) in order to obtain the Maximum A Posteriori (MAP) encoding:

$$\widehat{s}^{\text{MAP}} = \arg\max_{s} \ell(\widehat{\theta}|X = x, S = s). \tag{3}$$

We used this approach to optimally describe each structure as a serie of SLs. This process of compression of 3D protein conformation into 1D SL sequence is illustrated in Figure 1c,d, on the structure of chain B of the amidotransferase (pdb ID: 1gpw_B). This protein is coloured after HMM-SA encoding, according to its corresponding series of SLs.

Assessing the discretization of protein structures

For a given SL, the average Root-Mean-Square deviation (RMSd) between C_α coordinates, that is an Euclidean distance of the fragments to their centroid best superimposed, is used to measure the structural variability of each SL. For two given fragments, the RMSd between C_α coordinates of superimposed fragments is used to measure their structural proximity.

To reconstruct the protein 3D structures from their description as a series of SLs, and to keep some possible comparisons, we use the building procedure employed by Kolodny *et al.* (2002). Briefly, the fragments are assembled using an iterative concatenation procedure to adjust 3D conformation.

3. HMM-SA as a general concept to simplify 3D protein structure analysis ?

3.1 Results of HMM-SA identification

HMM-SA is weakly dependent on the learning set

Structural alphabet of increasing sizes using either HMM or MM are learned and compared on the basis of their goodness of fit. The influence of the Markovian process is large. For MM, no BIC optimum is reached until alphabet sizes of 70 whereas, for HMM, a larger optimum is reached for 27 hidden states, which means a better fit of the data using HMM. Interestingly, the Markov classification takes advantage of information implicitly contained in the succession of observations to greatly reduce the number of SLs, while keeping a more homogeneous repartition of fragments into the different SLs (the least frequent SL represents 1.5% of fragments).

Similar results are obtained using two independent learning sets of 250 proteins with similar BIC curves evolution. It follows that, at the optimum, structural alphabet is very weakly dependent on the learning set, which in turn suggests that the learned model can be considered as representative of all protein structures. The optimal structural alphabet, HMM-SA, obtained by using statistical criterion BIC, corresponds to 27 SLs and their transition matrix. Main characteristics of HMM-SA (Camproux *et al.*, 2004) are briefly summarized below.

Geometrical and logical description of HMM-SA

The 27 identified SLs are denoted as (case sensitive) structural letters: namely $a, A, B, ..., Y, Z$. The set of SLs, sorted by increasing stretches is presented in Figure 1b and their transitions constitute the structural alphabet, HMM-SA. The *local fit approximation* is low, as quantified by the average C_α RMSd to the centroid associated with each SL (0.23 ± 0.14 Å). Concerning description of logic of protein architecture, 66% of the 729 possible transitions between SLs have frequencies of less than 1%. The existence of pathways between SLs is observed, obeying some precise and unidirectional rules. These results are detailed in Camproux *et al.* (1999b; 2004).

Actually, SLs associated with close shapes have been distinguished by different logical rules. SLs A, a, V, W appear almost exclusively in α-helices (more than 92% of associated fragments assigned to α-helices) while five SLs L, N, M, T, X are mostly located in extended structures (from 47% to 78% of associated fragments assigned to strands). Interestingly, the other 18 SLs are involved in loops description, which is particularly interesting given the variability of loop structures and their implication in numerous important processes.

The stochastic HMM approach allows (i) the characterization of different short structural 3D SLs (ii) the description of the heterogeneity of their corresponding short fragments and (iii) the study of their global organization by quantifying their connections. The transition matrix only shows a limited number of transitions between SLs, indicating that the connections by which the SLs form the protein structures are well organized.

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 learned. Our model fits well the previous knowledge related to protein architecture organisation and seems able to grab some subtle details of protein organisation, such as helix sub-level organisation schemes. For instance, the two closest SLs A, a in terms of geometry, close to canonical α-helix, are distinguished by different prefered transitions. Taking into account the dependence between the states results in a description of local protein structures of low complexity. Although we use short fragments, the learning process on entire protein conformations captures the logic of the assembly on a larger scale. HMM-SA shows very reasonable performance in terms of reconstruction of the whole protein structure accuracy, (RMSd value less than 1 Å), compared to other recent fragment libraries optimized in a purpose of reconstruction (Kolodny *et al.*, 2002; Micheletti *et al.*, 2000).

Subsequently, HMM-SA provides some kind of compression from the 3D protein coordinate space into the 1D structural alphabet space (see Figure 1 c,d). From such 1D encoding and the associated logical rules, it is possible to tackle the exploration of 3D protein conformations using 1D techniques, as performed in classical sequence analysis. This widens the perspective of being able to work with a 1D representation of 3D structures much beyond the simple search of exact words, through the use of the classical 1D AA alignment methods. We have explored different directions in which this facility could be of interest.

3.2 Different sucessful applications of HMM-SA

Different applications of HMM-SA have been explored, such as:

- study of conformations of side chains in protein structures (Gautier *et al.*, 2004). It establishes a set of tools for analyzing lateral chain conformations of proteins in the server Ressource Parisienne en Bioinformatique Structurale (RPBS, http://bioserv.rpbs.jussieu.fr/cgi-bin/SCit);
- improvement of protein fold recognition from AA content compared to classical methods by adding Markovian information (Deschavanne *et al.*, 2009);
- performing fast 3D similarity search (RPBS, http://bioserv.rpbs.jussieu.fr/cgi-bin/SA-Search). 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 α-carbon coordinates (Gibrat *et al.*, 1996; Holm & Sander, 1993; Shindyalov & Bourne, 1998). Using HMM, the lod-score matrix of similarity between SLs allows the quantification of the similarity of protein fragments encoded as different series of SLs. It is possible to use it with classical methods developed for the AA sequence similarity search and thus to reduce 3D searches as a 1*D* sequence alignment problem (Guyon *et al.*, 2004);
- analyzing protein contacts (Martin *et al.*, 2008b). This study showed that the description of protein contacts (intra and inter-molecular) by the local structure residues (described by HMM-SA) involved in these contacts is more sensitive than that provided by type of AAs involved in contacts;
- analyzing the deformation of proteins during interaction (Martin *et al.*, 2008a): HMM-SA has also been used to analyze the regions of protein/protein interactions before and after contact (Martin et al., 2008b). This study identified regions undergoing deformation, and the identification of common structural motifs from the strain involved in the interaction of two proteins;
- deciphering the shape and deformation of secondary structures (Baussand *et al.*, in press). The conformation of secondary structures can be further analyzed and detailed thanks to HMM-SA which allows a better local description of protein surface, core and interface in terms of secondary structure shape and deformation. Induced-fit modification tendencies should be valuable information to identify and characterize regions under strong structural constraints for functional reasons;
- in addition, HMM-SA was shown to be a powerful tool for the analysis of protein loops, the most variable and flexible regions in proteins (Camproux *et al.*, 2001; Regad *et al.*, 2006). They are, however, often known to play an important role in protein function and stability

(Fetrow, 1995; Fernandez-Fuentes *et al.*, 2004). The HMM-SA was optimised in terms of 3D local description of proteins and resulted in precise and detailed description of 3D conformations into 27 SLs: 18 SLs being focused on loop description. Indeed, the encoding of loop structures allowed the establishment of a systematic methodology to extract all the structural motifs of seven residues in all the loops, especially long loops. Analysis of these patterns and their environment has enabled a quantification of structural redundancy in loops. An analysis of their distribution in the short and long loops has shown that the short and long loops share a number of structural motifs (Regad *et al.*, 2008);

- concerning the information of AA sequence, all the SLs of HMM-SA have some significant AA sequence specificity compared to the profiles of a collection of protein fragments (Camproux & Tuffery, 2005). This dependence can be used to generate direct candidate folds from AA sequence in a two-steps scheme of prediction. First, the goal is to predict local SL series from AA sequence. For short fragments, available 3D conformation can be found in the PDB (using Guyon *et al.*, 2004); for longer ones (SL-fragments not available in the PDB), local SLs or SL-words could be assembled to generate 3D structures, following the same principles as Maupetit *et al.* (2009);
- concerning the 3D reconstruction, a recent paper (Maupetit *et al.*, 2009) has shown the performance of HMM-SA for peptides (short SL-fragments). Rational peptide design and large-scale prediction of peptide structures from AA sequences remain a challenge for chemical biologists. This paper proposed a *de novo* modelling of 3D conformations for peptides between 9 and 25 amino acids in aqueous solution. Using HMM-SA, PEP-FOLD assembles the predicted SL fragment profiles by a greedy procedure driven by a modified version of the OPEP coarse-grained force field;
- in addition, the HMM methodology of building structural alphabet has been proven to identify a specific structural alphabet for porin proteins (i.e. transmembrane proteins). This alphabet has helped to describe how fine these proteins are, specifically in terms of beta strands composition (Martin *et al.*, 2008c).

Actually, HMM-SA is a very interesting tool to study protein structures and hence function. In particular, it is interesting to identify conserved SL-patterns having particularities such as being associated to a specific function or to turns, for example. Then, the natural continuation of such identifications is to provide a method being able to detect those patterns directly from AA sequence. Thus, it makes it possible to annotate AA sequences with annotations identified from 3D structure without knowing the conformation of the considered sequence.
The last section focuses on the prediction of patterns identified as specific to a function for example.

4. Using HMM to detect interesting HMM-SA patterns

As previously introduced, sequencing technologies are constantly providing new AA sequences with often few functional knowledge. Hence, being able to retrieve information about new protein sequences is a critical problem.
In this context, automatic tools allowing to provide such information are of big interest. The most common way to perform such a search is to identify patterns specific from a given function for example and to design a prediction method. Information taken into account can consist in different levels: only sequence (Ansari & Raghave, 2010; Sigrist *et al.*, 2010), sequence and structure (Halperin *et al.*, 2008; Pugalenthi *et al.*, 2008), only structure (Manikandan *et al.*, 2008; Polacco & Babbitt, 2006) or use of more general classifications: GO (Espadaler *et al.*,

2006), SCOP (Tendulkar *et al.*, 2010) ... In this section, the objective is to design a prediction method only based on AA sequence in order to provide information for only sequenced proteins.
However, sequence-based methods are likely to be limited with regards to structure-based ones as structure is known to be better conserved than sequence (Chothia *et al.*, 2003). Hence, the proposed method will use HMM-SA as a structure-based middle step to identify interesting structural patterns. As loops are very often implied in interactions (Ansari & Helms, 2005; Saraste *et al.*, 1990), stress is laid on patterns of interest found in loops. Those patterns will be defined here as four SL words encoding seven AA residues. This length has been chosen to obtain satisfying representativities (Regad *et al.*, 2006). However, the prediction method is independent on the pattern length and could be applied to any identified pattern.

4.1 Looking for interesting patterns

In bioinformatics, it is common to look for a pattern of interest in a potentially large set of rather short sequences (upstream gene regions, proteins, exons, etc.). In DNA sequences, it has been observed that functional sites have unusual frequencies: very frequent or rare. Some methods used this observation to extract functional sites by defining them as over- or under-represented sites. Their identification is usually achieved by considering a homogeneous m-order Markov model of the sequence, allowing the computation of p-values (probability that the expected occurrence of a word is larger than its observed occurrence). Stationarity of the model is often assumed for practical reasons but this approximation can result in some artifacts especially when a large set of small sequences are considered. No specific development has taken into account the counting of occurrences in a large set of short independent sequences as loop trajectories in HMM-SA space. A study aiming at addressing this problem by deriving efficient approaches and algorithms to perform these computations for both low and high complexity patterns in the framework of homogeneous or heterogeneous Markov models has been developped in Nuel *et al.* (2010). More precisely, this article proposed an exact method, enabling to take into account both non stationarity and fragmentary structure of sequences, applied it on simulated and real sets of sequences and actually illustrated that pattern statistics can be very sensitive to the stationary assumption.
Subsequently, a detailed analysis of statistically exceptional motifs, identified by HMM-SA, with regards to SCOP superfamilies, groups of proteins with similar structure and function, shed a new light on candidate patterns. Indeed, this study confirmed the link of those potentially interesting patterns with functional motifs in loops and provide a systematic way of identifying such patterns (Regad *et al.*, in revision).
Then, once a pattern has been confirmed as interesting, it is of big interest to be able to predict its presence and thus, the presence of the identified function, directly from a protein AA sequence, even if the 3D structure is unknown. It is important to notice that among identified patterns only the ones showing AA sequence specificities will be likely to be predicted directly from AA sequences. The proposed prediction method is divided into two steps: the first one aims at assigning to each four-AA sequence a SL profile, this will not be deeply describe as outside the scope of this chapter. The second step makes use of a HMM model to combine the profiles provided by step 1 and compute a final probability of finding the considered motif at each position in the sequence.

4.2 The initial data

We use the AA sequences and corresponding SL encoding of 16,995 loops extracted from the PDB with at most 25% of sequence identity. This limited sequence identity rate aims at avoiding any bias in the learning step. The length of loops ranges from 1 to 1,261 SL (hence from 4 to 1,264 AA) with an average of 116 and a standard deviation of 129. They are extracted from 7,778 different proteins.

4.3 First step outline: from four amino-acids to one structural letter profile

The first step input is a 4-AA sequence fragment. In a practical point of view, each overlapping 4-AA words of the considered AA sequence will successively become input. The goal is to find what should be the SL encoding for this fragment. However, as there is no exact bijection between AA and SL sequences, it would be unappropriate to give only one possible SL for each 4-AA fragment. Hence, a SL profile will be the first step output. It consists in a score quantifying the probability of finding each SL at the considered location. This score is based on votes provided by 351 rules: there is one rule for each SL couple ($27 \times 26/2$).

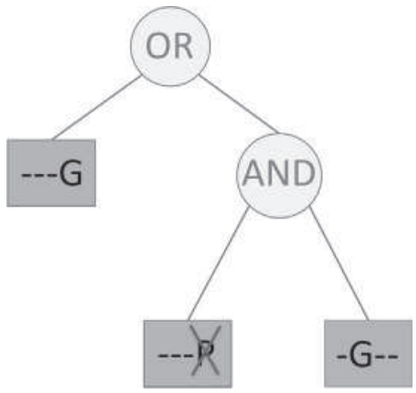


Fig. 3. Example of classifier used to discriminate between two letters A and B: if there is *(G in second position)* AND *(no P in fourth position)* OR *(a G in fourth position)* then the sequence is affected to A else to B.

It is really appropriate to illustrate those rules through a tree-like representation. Indeed, each rule is a combination of binary questions about the presence or absence of a given AA for a given position (between 1 and 4) in the considered 4-AA fragment. For instance, Figure 3 gives an example. Contrary to classical decision trees, this tree has to be read from leaves to root: by sequentially answering to each leaf question (there is or there is not such and such an AA at such and such a position) and combining the answers through the AND/OR operators contained in nodes, a global yes/no answer is obtained allowing to affect the AA sequence to one of the two SLs compared through this classifier. Hence, the 351 rules will provide votes concerning the 4-AA fragment which constitute a kind of profile for the *true* encoding of this sequence into one structural letter. The optimization of the rules is performed through genetic programming (Koza, 1992; Langdon & Poli, 2002) by scoring the rules through their parsimony and the entropy gain they achieve.

4.4 Second step: specific pattern modelling and application to prediction

Due to the complexity of the prediction problem (impossibility to build an easy bijection between AAs and SLs), the first step cannot be sufficient to answer the problem. Indeed, some SLs are easy to discriminate through their AA sequence. For example, SLs B and M are

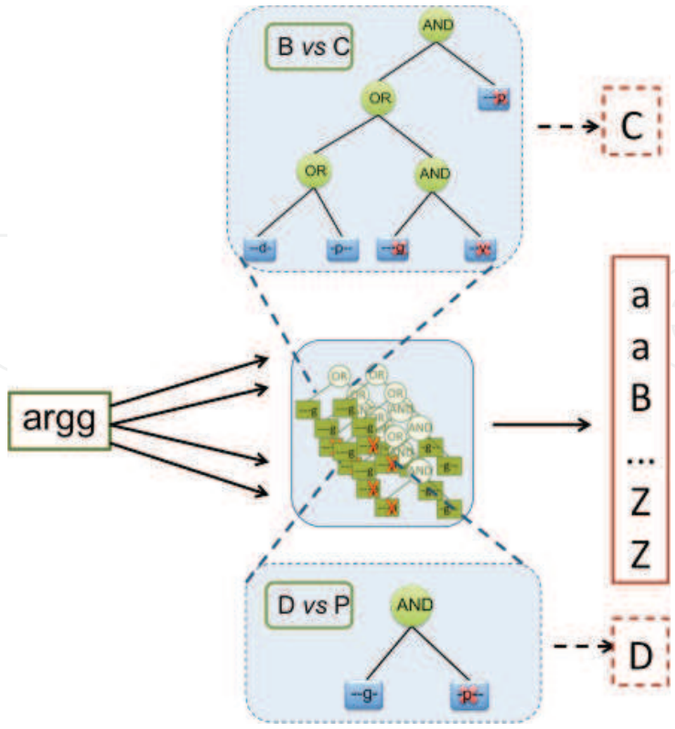


Fig. 4. Global unfolding of Step 1 and two examples of classifiers. A 4-AA long fragment (argg) is given as input of the 351 classifiers, each one voting for one out of the two SLs it compares (ex.: B vs C and D vs P). Finally, a vector of 351 votes is obtained.

very well discriminated through their classifier: one out of the two subgroups obtained after applying the classifier contains 3.2% of the B SL and 98.0% of the M. On the other hand, SLs a and M are particularly difficult to distinguish through their AA sequence: one out of the two subgroups obtained after applying the corresponding classifier contains all the SLs a and 80.6% of SLs M, which is a very poor classification. Hence, further information has to be taken into account to be able to make decisions about a four-SL word. In this context, a particularly interesting knowledge is about dependencies between SLs. It is the goal of the second step.

The aim of this step is to decide, given the results of the first step for four consecutive SLs and through a scoring function, if the conformation adopted by the considered seven residue fragment is likely to be encoded into a given four SL word identified to be linked to a functional pattern.

As emphasized earlier, a real dependency exists between successive SLs, especially because of overlaps. Hence, this dependency can be favourably used to build a model. A HMM has been chosen to model the link between first step outputs and a given four SL word. This HMM is described in Figure 5. In this model, hidden states are the *true* SLs while observed states are outputs of step 1 for the corresponding AA sequence. Arrows between S_i and S_{i+1} symbolizes the dependency between successive letters called *transition probabilities* in HMM context and arrows between S_i and O_i represent the link between true SLs and step 1 outputs, namely the *output probabilities*.

Thanks to this model, the objective of the second step is to compute the probability of the four true SLs being the target functional pattern given the step 1 outputs for four successive (and

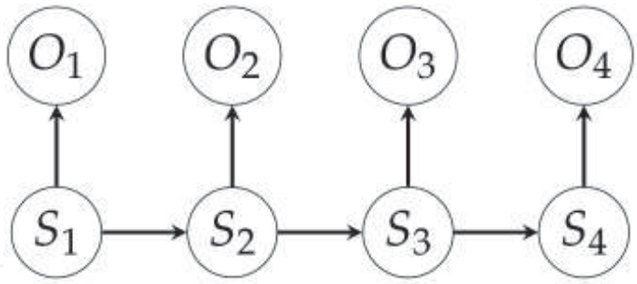


Fig. 5. Structure of the HMM used to model the relationship between first step outputs and *true* SLs for a seven residue fragment: (S_1, S_2, S_3, S_4) are the *true* SLs and $O_i = (o_i^1, o_i^2, \ldots, o_i^{351})$ is the vector of votes obtained from step 1 for the four AA fragment encoded by S_i.

overlapping) four-AA fragments. Hence, we have to compute

$$\mathbb{P}(S_{1:4}|O_{1:4}) = \mathbb{P}(S_1, S_2, S_3, S_4|O_1, O_2, O_3, O_4). \tag{4}$$

High values of this probability will indicate a strong assumption that the considered fragment is likely to be encoded into the identified pattern and then to have the target function.
According to the chosen model,

$$\mathbb{P}(S_{1:4}|O_{1:4}) = \mathbb{P}(S_1|O_1) \prod_{i=2}^{4} \mathbb{P}(S_i|S_{i-1}) \mathbb{P}(S_i|O_i).$$

Now, $\mathbb{P}(S_i|O_i)$ has to be computed. Assuming that the results of the 351 different trees are independent,

$$\mathbb{P}(S_i|O_i) = \mathbb{P}(S_i|o_i^1, o_i^2, \ldots, o_i^{351}) = \prod_{j=1}^{351} \mathbb{P}(S_i|o_i^j).$$

This assumption is wrong for some comparisons (especially comparisons implying a common SL which is well predicted) but most of pairs of comparisons can be considered as independent (results not shown).
Then, by Bayes theorem, and by denoting by $\bar{S_i}$ the absence of S_i, that is to say there is any of the 26 other SLs,

$$\mathbb{P}(S_i|o_i^j) = \frac{\mathbb{P}(o_i^j|S_i)\mathbb{P}(S_i)}{\mathbb{P}(o_i^j|S_i)\mathbb{P}(S_i) + \mathbb{P}(o_i^j|\overline{S_i})\mathbb{P}(\overline{S_i})}.$$

Finally, $\mathbb{P}(S_i)$, $\mathbb{P}(o_i^j|S_i)$ and $\mathbb{P}(S_i|S_{i-1})$ are estimated on the dataset.

4.5 Applications

4.5.1 Prediction of an ATP-binding site specific motif

Previous studies (as described at the beginning of this section) have shown that fragments encoded into the four SLs *YUOD* (see Figure 6(a)) are very often associated to ATP/GTP binding sites. Indeed, in our database, 95% of fragments encoded into *YUOD* are associated to this functional annotation in SwissProt database. Hence, being able to predict the encoding into *YUOD* is really useful to predict this function for a new AA sequence.
The superimposition of several fragments encoded into *YUOD* is shown in Figure 6. Moreover, this structural word has a high sequence specificity as shown in Figure 4, especially

positions 1, 6 and 7. Thus, this structural word, involved in protein function (binding to ATP/GTP), is a very good candidate for our approach.

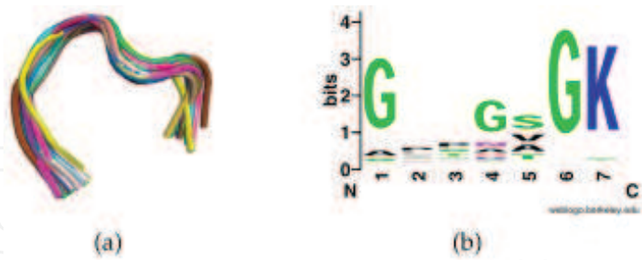


Fig. 6. (a) Representation of several fragments encoded into *YUOD*. (b) Logo of the AA sequences encoded into *YUOD*.

In our dataset, *YUOD* can be found 183 times in 181 proteins (two proteins contain two occurrences). The model is applied on the whole proteins to study the ability of the computed probability (Eq. 4) to discriminate between *YUOD* and $\overline{YUOD}$ (*not YUOD*). The ROC curve associated to the logarithm of this probability is shown in Figure 7. It displays the sensitivity (ability to retrieve *YUOD*) and specificity (ability to recognize $\overline{YUOD}$) according to the probability threshold chosen to split the words into *YUOD* and $\overline{YUOD}$. The AUC (area under curve) associated to this ROC curve is 0.9866. Hence, the computed probability is really efficient to identify *YUOD* among all other words. Indeed, such a discrimination quality is particularly valuable because of the ratio between the two classes: *YUOD* only represents 0.52% of studied words. Then, according to the application requirements, several thresholds can be defined providing different balances between sensitivity and specificity. Some interesting threshold values and their corresponding parameters are enclosed in Table 1. Very high values of specificity have been chosen, indeed the $\overline{YUOD}$ class is really large and then only 1% of false positive ($\overline{YUOD}$ predicted as *YUOD*) can be a large number when applied to big proteins or to several proteins.

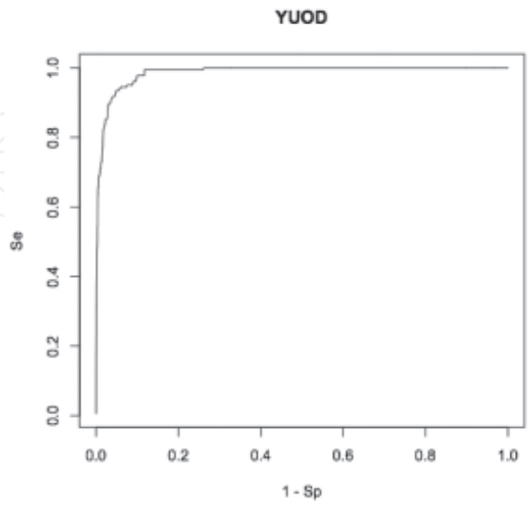


Fig. 7. ROC curve (AUC = 0.9866) associated with the probability of having *YUOD* for a given seven AA fragment (Se=sensitivity, Sp=specificity).

Threshold	-4829	-4805	-4732
Specificity	90.02	95.07	99.00
Sensitivity	97.81	93.44	69.95

Table 1. Sensitivity and specificity obtained for the identification of *YUOD* according to the chosen log(probability) threshold.

An example of *YUOD* detection is given in Figure 8. It concerns the chain A of the Circadian clock protein kinase kaiC (pdb ID: 2gbl_A). It originally contains two true *YUOD* occurrences and four have been predicted through our model. Two out of the four positives (numbers 1 and 2) are exactly located at co-crystallized ATP binding sites (A and B). Moreover, among the two false positives, number 3 adopts a 3D conformation which is really close to the one observed at ATP binding sites. This example demonstrates the difficulty of evaluating a prediction method for annotations. The evaluation of true positive and false negative can be really precise when dealing with manually annotated and reviewed databases such as Swiss-Prot but false positives may be true positive that have not yet been experimentally verified. It is impossible to make a decision in this case.

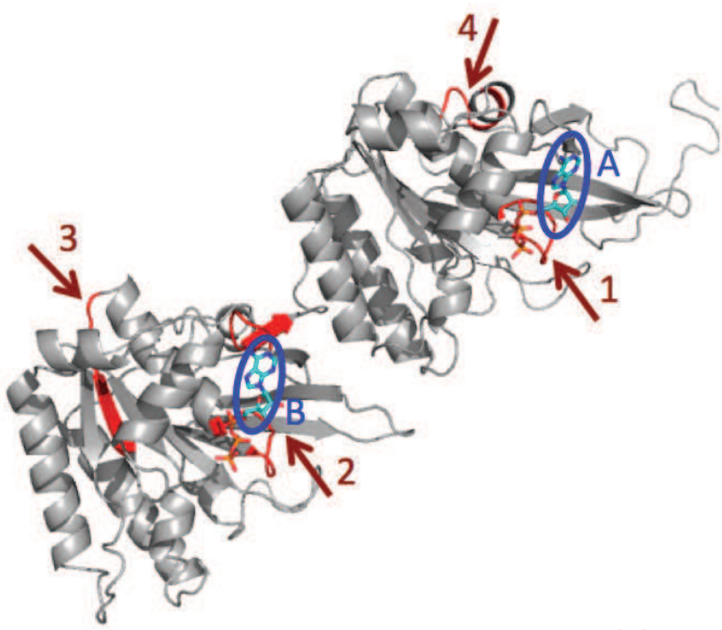


Fig. 8. 3D representation of 2gbl_A co-crystallized with two ATP molecules (indicated by lettered circles). The fragments identified as *YUOD* are indicated with numbered arrows.

4.5.2 Prediction of a SAH/SAM-binding site specific motif

S-adenosyl-methionine (SAM) and S-adenosyl-homocysteine (SAH) are molecules associated to some methylation processes and are particularly studied in the context of antiviral drugs research. It is then interesting to be able to predict their binding to proteins. The four-SL word *RUDO* has been identified to be most of time associated to SAH/SAM in Swiss-Prot. Moreover, it has a certain sequence specificity (results not shown).

In our dataset, *RUDO* is found 39 times in 39 different proteins. The AUC associated to the ROC curve corresponding to the log(probability) computed by our method is 0.9606. The specificity and sensitivity obtained with different thresholds for the log(probability) are given in Table 2. Thus, results are satisfying and allow us to recover more than two thirds of the *RUDO* motifs without wrongly assigning more than 1% of the other words.

Threshold	-4903	-4806	-4712
Specificity	90.00	95.00	99.00
Sensitivity	87.18	84.62	69.23

Table 2. Sensitivity and specificity obtained for the identification of *RUDO* according to the chosen log(probability) threshold.

An illustration can be found in Figure 9. It concerns isoloquiritigenin 2'-O-methyltransferase (pdb ID: 1fp1) which was here co-crystallized with a SAH molecules. Four words were predicted as *RUDO* with a threshold of -4712 whereas only one has been encoded as *RUDO*. However, looking of the 3D conformation, it appears that all four identified fragments are really closed to the ligand. Thus, the method using the HMM-SA as a tool to discover patterns, is not limited to the fragments being strictly encoded into the identified fragments but is also able to discover fragments with close encodings and thus structures, as only sequence is finally taken into account. Hence, fragments which are likely to adopt a *RUDO*-like conformation can be as well identified by the method.

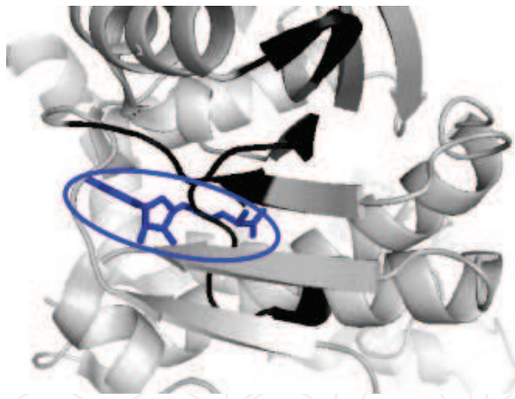


Fig. 9. 3D representation of 1fp1 (light grey cartoons) co-crystallized with a SAH molecule (indicated by a circle). The fragments identified as *RUDO* are black-coloured.

4.5.3 Prediction of a specific β-turn

The prediction of turns is also of special interest in protein study (Fuchs & Alix, 2005). The four-SL word *HBDS* can be linked to β-turns: the corresponding fragment conformations are shown in Figure 10. This is a frequent word, in our database, it was found 1,633 times in 1,363 different proteins (there are one to six occurrences in those proteins). The AUC associated to the prediction of *HBDS* is 0.9359. Table 3 indicates the specificities and sensitivities associated to different log(probability) values. The results are a bit less efficient than previous ones (due to a lower sequence specificity) but enable to locate 85% of those turns with a specificity of 90% (knowing this specificity is likely to be underestimated because of close fragments which have not been strictly encoded into HBDS).

Fig. 10. 3D representation of several fragments encoded into *HBDS*.

Threshold	-4013	-3844	-3777
Specificity	90.17	95.11	98.88
Sensitivity	84.71	71.07	28.93

Table 3. Sensitivity and specificity obtained for the identification of *HBDS* according to the chosen log(probability) threshold.

4.6 Prediction method outcome

The automatic annotation of simply sequenced proteins is a very important task in the present context of high-throughput sequencing programs. The method proposed in this section is based on the identification of motifs of interest directly on structures using HMM-SA. The *input data* of the described method are *only AA sequences* and as a consequence, only patterns having sequence specificities will be likely to be handled with this method. But for this kind of motifs, the method is really powerful. One method (Maupetit *et al.*, 2009) has already been proposed to predict the 3D structure of small peptides through HMM-SA but the motif-oriented aspect of the method proposed here makes it much more precise and time efficient.

As much information as possible is extracted from data. The dependence between AA sequences and 3D structures is learned in the first step through the use of HMM-SA. Then, the second step takes advantage of two different sources of information by building a HMM. Firstly, the strength of dependence between AAs and SLs is quantified and used through observation probabilities: some observations will be really trusted (when a strong link has been found in the first step) whereas others will be considered with care as less reliable. Secondly, the dependence between successive SLs (some SLs favourably follow other ones) is also taken into consideration by the computation of transition probabilities. Finally, a really complete model is obtained by the addition of both steps.

Moreover, as HMM-SA is only an intermediate between sequence and function (or any other interesting pattern), the method, as shown in some illustrations, is able to identify fragments as close to the target word even if this fragment would not be encoded into the exact SL target word. Hence, relying on sequences is a good way to overcome some cases of flexibility: in the crystallization conditions, the fragment has not been found in the strict conformation associated to the target word, but its AA sequence specifities can be recognized by the prediction method. Eventually, HMM-SA encoding and the proposed prediction method are interestingly complementing each other in the prediction of patterns of interest.

Furthermore, the important adaptability of the prediction method is of large interest. Indeed, in this paper we focused on pattern which had been identified directly through HMM-SA but it is completely possible to identify 3D motifs as interesting for any other reason, to encode it into HMM-SA and to build the model on the obtained word. Let us recall here that the size of considered fragments is not limited. Earlier, only seven-residue fragments

have been considered but any length would be possible. Furthermore, as illustrated through the three examples, the size of the learning dataset can be really variable (from 35 to 1633 occurrences of the pattern) as the model is always the same. The only variable parameter is the log(probability) threshold. However, preliminary studies seem to indicate that this threshold depends on the strength of the sequence specificity of the structure. Hence, further work could be able to set this threshold directly from the quantification of this dependence.

5. Conclusion

Interest and limits of the HMM to study 3D protein organisation

In contrast to supervised learning strategies (Levitt and Chothia, 1976; Kabsch and Sander, 1983; Richards and Kundrot, 1988; Prestrelski et al., 1992; Hutchinson and Thornton, 1993; Zhu, 1995), the SLs emerged from the HMM without any prior knowledge of secondary structural classification. In that sense, the HMM is able to classify conformations that template studies must describe as undefined or random structures and also to subdivide conformation classes previously defined as a single class, resulting in a finer description of the 3D conformations. For instance, the HMM approach allows different levels of variability within each SL. Classical methods have recently been used to extract and classify local protein backbone elements but these methods did not take into account any local dependence between SLs: all these studies used only the structural characteristics to identify structural 3D letters and reconstructed a posteriori the organization of these 3D conformations. One major contribution of HMM is that this model implicitly takes into account the sequential connections between the SLs. It is striking that structurally close SLs can have different roles in the construction of 3D structures.

HMM-SA learning has shown to be stable over different protein sets. Our model fits well the previous knowledge related to protein architecture organisation. Using such a model, the structure of proteins can be reconstructed with an average accuracy close to 1.1 Å root-mean-square deviation and for a low complexity of 3D reconstruction (see Camproux *et al.*, 2004, for details). This stochastic HMM approach allows the characterization of different SLs with different fragment heterogeneity by taking into account their global organization and quantifying their connections. It results in a fine and pertinent description of the 3D structures and a very performant tool to simplify 3D conformation of proteins. Different successful applications of HMM-SA for 3D analysis have been performed.

This ability has allowed to design several methods of protein studies, such as the prediction of interesting patterns detailed in this chapter. This method has shown to be really efficient for patterns having a certain AA sequence specificity. Further work should allow to predict the efficiency and the threshold to be used directly from a quantification of this dependency. Moreover, this prediction method has the main advantage to be really adaptive, to different pattern lengths or to different alphabets for example. In this study, HMM-SA has been used because of its very interesting abilities of precise description especially for loops, but the same methodology could be applied on other types of alphabets. Finally, the method is bounded by the function specificity of the pattern. Indeed, a function might be associated to different patterns. Thus, our method is able to predict one type of realization of a given function at a time. Of course, it is completely possible to learn several patterns linked to a function and to give a global prediction for all of them. But for the moment, this limit prevents us to compare with prediction methods for specific function (such as Ansari & Raghave, 2010) encoded through different patterns. This should be quickly possible by the identification of new patterns which is in progress.

6. Further HMM improvements

Concerning the HMM-SA identification, a number of improvements can be brought into the HMM modeling by taking into account deterministic dependency and local descriptors, the criterion of model selection and posterior probabilities of different structural letters in the encoding.

In the previous work, the idea was to consider the model of Figure 2 where all X_i are generated independently from each other conditionally to a hidden state $S_i \in \mathcal{S} = \{1, 2, \ldots, K\}$ as detailed in Section 2.2. One problem of this approach is that it does not take into account the correlation between X_i and X_{i+1}. We alternatively suggest to consider the model where the distribution of X_{i+1} depends on S_{i+1} (like in the previous model) and from X_i. For example, if we assume that $X_{i+1}^1 = X_i^3$ and that $(X_{i+1}^2, X_{i+1}^3, X_{i+1}^4)$ has a Gaussian distribution whose parameters only depend on S_{i+1}, the resulting model both improves the existing one and reduces its number of parameters.

A critical point of the HMM-SA approach is also related to the model selection: how many structural letters should we use in order to get the best structural alphabet ? For this problem however, our objective is not only to select the most parsimonious model providing the best fitting but also to provide a reliable classification of the fragments in order to allocate a given fragment to a specific structural letter with the least possible uncertainty. For that purpose, it might be interesting to replace the used BIC criterion by classification orientated criteria like the ICL (Biernacki *et al.*, 2000; McLachlan & Peel, 2008) or the Discriminative Information Criterion from Biem (2003). The idea of these approaches is to introduce in the penalization a term related to the entropy of the classification which purpose is to avoid to select a model where two structural letters are too close to each other.

Another important issue is related to the encoding of 3D structures into sequences of structural letters. For that purpose, it is both natural and classical to use the Viterbi's algorithm in order to obtain the MAP encoding. However, it often exists many alternative suboptimal configurations that might be of interest. In order to check this, it might be interesting to compute the posterior distribution $P(S = s|X = x, \widehat{\theta})$ using the Forward/Backward quantities and hence to point out regions where the structural alphabet encoding has a low confidence. One may then either exclude this low reliability regions or take into account the uncertainty of the encoding by sampling several encoding for these regions.

It also might be of great interest to introduce in the HMM-SA model a descriptor of the structure flexibility in its learning process. Initially, a unique 3D structure was supposed to correspond to one protein sequence (Mirsky, 1936). The constant and rapid increase in the number of experimentally solved protein structures has shown the flexibility of 3D structure proteins to adapt to different conditions and partners. Thus, this property is at the heart of the fundamental functions of proteins. This flexibility is being quantified by parameters such as the B-factor. The inclusion of this information in the construction of a new alphabet could be used to define classes of structures particularly flexible and to better model the complexity of proteins. For instance, knowledge and prediction of these flexible regions could significantly improve docking protocols, including the choice of starting structures.

Actually, if the HMM-SA original model only considered the 3D structure of the protein, the additional work presented in this chapter has shown that the original AA sequences also bear useful physicochemical information that can improve the prediction. It is hence very tempting to combine these two approaches together by introducing the AA sequence into the HMM-SA model from the beginning like suggested in Figure 11.

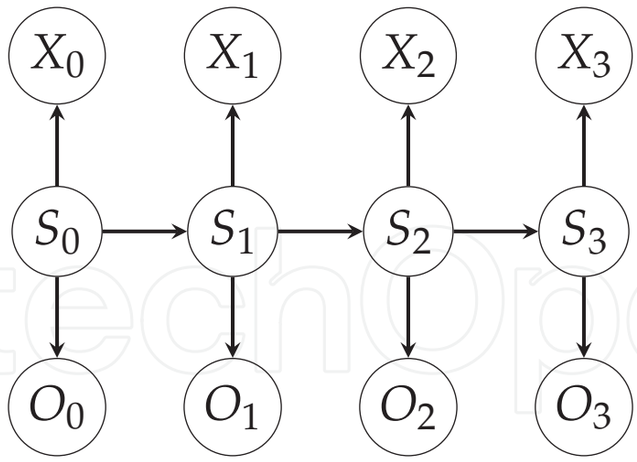


Fig. 11. Graph of dependencies in the model combining both the 3D structure and the primary sequences. The model is drawn for $n = 7$ C_α hence resulting in a total of $n - 3 = 4$ SLs.

7. References

S. Ansari and V. Helms, Statistical analysis of predominantly transient protein-protein interfaces. *Proteins*, 61(2): 344-355, 2005.

H.R. Ansari and G.P.S. Raghava, Identification of NAD interaction residues in proteins. *BMC Bioinformatics*, 11(160), 2010.

L. Baeten, J. Reumers, V. Tur, F. Stricher, T. Lenaerts, L. Serrano, F. Rousseau and J. Schymkowitz, Reconstruction of protein backbones from the BriX collection of canonical protein fragments. *PLoS Comput Biol.*, 4(5), 2010.

A. Bateman, L. Coin, R. Durbin, R.V. Finn, V. Hollich, Griffiths-Jones S., Khanna A., Marshall M., Moxon S., E. L. L. Sonnhammer, D. J. Studholme, C. Yeats and S.R. Eddy, The Pfam Protein Families Database. *Nucleic Acids Research*, 32: 138-41, 2004.

J. Baussand and A.-C. Camproux, Deciphering the shape and deformation of secondary structures using a structural alphabet approach, in press in *BMC structural biology*.

F.C. Berstein, T.G. Koetzle, G.J.B. Williams, E.F. Meyer, M.D. Brice, J.R. Rogers, O. Kennard, T. Shimanouchi and M. Tasumi, The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.*, 112, 535-542, 1977.

A. Biem, A model selection criterion for classification: Application to hmm topology optimization, *Proc. 17th ICDAR*, Edinburgh, U.K, 104-108, 2003.

C. Biernacki, G. Celeux and G. Govaert, Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(7):719-725, 2000.

B. Boeckmann, A. Bairoch, R. Apweiler, M. Blatter, A. Estreicher, E. Gasteiger, M.J. Martin, K. Michoud, I. Phan, R. Pilbout and M. Schneider, The swiss-prot protein knowledgebase and its supplement tremble. *Nucleic Acids Res.*, 31: 365-370, 2003.

A.G. De Brevern, C. Etchebest and S. Hazout, Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. *Proteins*, 41: 271-87, 2000.

C. Bystroff, V. Thorsson and D. Baker, HMMSTR: a hidden Markov model for local sequence-structure. *J. Mol. Biol.*, 301(1): 173-90, 2000.

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.

A.C. Camproux, P. Tuffery, L. Buffat, C. Andr, J.F. Boisvieux and S. Hazout, Using short structural building blocks defined by a Hidden Markov Model for analysing patterns between regular secondary structures. *Theoretical Chemistry Accounts*, 101: 33-40, 1999.

A.C. Camproux, A.G. De Brevern and S. Hazout, Exploring the use of a structural alphabet for structural prediction of protein loops, *Theoretical Chemistry Accounts*, 106: 28-35, 2001.

A.C. Camproux, R. Gautier, and P. Tuffery, A Hidden Markov model derived structural alphabet for proteins. *J. Mol. Biol.*, 339: 591-05, 2004.

A.C. Camproux and P. Tuffery, Hidden Markov Model derived Structural Alphabet for proteins: the learning of protein local shapes capture sequence specificity. *BBA*, 1724(3): 394-403, 2005.

C. Chothia, J. Gough, C. Vogel and S.A. Teichmann, Evolution of protein repertoire. *Science*, 300(5626):1701-1703, 2003.

T.U. Consortium, The Universal Protein Resource (UniProt). *Nucleic Acids Res*, 38, D142-148, 2010.

P. Deschavanne and P. Tuffery, Enhanced protein fold recognition using a structural alphabet. *Proteins*. 76(1):129-37, 2009.

R. Durbin, S. Eddy, A. Krogh and G. Mitchison, Biological sequence analysis. (1998) Probabilistic models of proteins and nucleic acids.

J. Espadaler, E. Querol, F.X. Aviles and B. Oliva, Identification of function-associated loop motifs and application to protein function prediction. *Bioinformatics*, 22(18): 2237-2243, 2006.

J.S., Fetrow, Omega loops: nonregular secondary structures significant in protein function and stability. *FASEB J*, 9: 708-17, 1995.

N. Fernandez-Fuentes, A. Hermoso, J. Espadaler, E. Querol, F.X. Aviles and B. Oliva, Classification of common functional loops of kinase super-families. *Proteins*, 56(3): 539-55, 2004.

D. Frishman and P. Argos, Knowledge-based protein secondary structure assignment. *Proteins*, 23(4): 566-79, 1995.

P.F.J. Fuchs and A.J.P. Alix, High accuracy prediction of β-turns and their types using propensities and multiple alignments. *Proteins*, 59(4): 828-839, 2005.

R. Gautier, A.C. Camproux and P. Tuffery, SCit: web tools for protein side chains conformation analysis. *Nucleic Acids Research*, Web Server issue : W508-11, 2004.

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

F. Guyon, A.C. Camproux, J. Hochez and P. Tuffery, SA-Search: a web tool for protein structure mining based on a Structural Alphabet. *Nucleic Acids Res*, 32: W545-48, 2004.

I. Halperin, D.S. Glazer, S. Wu and R.B. Altman, The FEATURE framework for protein function annotation: modeling new functions, improving performance and extending to novel applications. *BMC genomics*, 9(Sup 2): S2, 2008.

K. Henrick, Z. Feng, WF. Bluhm, D. Dimitropoulos, JF. Doreleijers, S. Dutta, JL. Flippen-Anderson, J. Lonides, C. Kamada, E. Krissinel, CL. Lawson, JL. Markley, H. Nakamura, R. Newman, Y. Shimizu, J. Swaminathan, S. Velankar, J. Ory, EL. Ulrich,

W. Vranken, J. Westbrook, R. Yamashita, H. Yang, J. Young, M. Yousufuddin and H.M., Berman, Remediation of the protein data bank archive. *Nucleic Acids Res*, 36, D426-D433, 2008.

U. Hobohm, M. Scharf, M. Schneider and C. Sandres, Selection of representative protein data sets. *Protein Sci.*, 1: 409-417, 1992.

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

R. Kolodny, P. Koehl, L. Guibas and M. Levitt, Small libraries of protein fragments model native protein structures accurately. *J. Mol. Biol.*, 323(2): 297-07, 2002.

J.R. Koza, Genetic Programming: On the Programming of Computers by Means of Natural Selection, MIT Press, 1992.

A. Krogh, M. Brown, I.S. Mian, K. Sjolander, D. Haussler, Hidden Markov Models in Computational Biology: Applications to Protein Modeling. *J. Mol. Biol.*, 235(5): 1501-31, 1994.

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

G.J. McLachlan, and D. Peel, *Finite mixture models*. New York: Wiley, 2000.

K. Manikandan, D. Pal, S. Ramakumar, N.E. Brener, S.S. Iyengar and G. Seetharaman, Functionally important segments in proteins dissected using Gene Ontology and geometric clustering of peptide fragments. *Genome Biol.*, 9(3): R52, 2008.

J. Martin, L. Regad, C. Etchebest and A.C. Camproux, Inter-residue contact analysing using local structure descriptors. *Proteins*, 9; 73(3): 672-689, 2008.

J. Martin, L. Regad, H. Lecornet and A.C. Camproux, Structural deformation upon protein-protein interaction: a structural alphabet approach. *BMC Structural Biology*, 8:12, 2008.

J. Martin, A. de Brevern and A.C. Camproux, In silico local structure approach: a case study on Outer Membrane Proteins. *Proteins*, 11; 71(1): 92-109, 2008.

W.B. Langdon and R. Poli, *F*oundations of Genetic Programming, Springer-Verlag, 2002.

J. Maupetit, P. Derreumaux and P. Tuffery, PEP-FOLD: an online resource for de novo peptide structure prediction. *Nucleic Acids Research*, 37, 2009.

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: 662-74, 2000.

A.E. Mirsky and L. Pauling, On the structure of native, denatured and coagulated proteins. *Proc. Natl. Acad. Sci.* USA, 22: 439-447, 1936.

G. Nuel, J. Martin, L. Regad and A.C. Camproux, Exact distribution of a pattern in a set of random sequences generated by a Markov source: applications to biological data. *Alg. Mol. Biol.*, 5:15, 2010.

A. Pandini, A. Fornili and J. Kleinjung, Structural alphabets derived from attractors in conformational space. *BMC Bioinformatics*, 11:97, 2010.

X. Pang, L. Zhou, M. Zhang, F. Xie, L. Yu, L. Zhang, L. Xu & X. Zhang. A mathematical model for peptide inhibitor design. *J. Comput Biol.* , 17(8): 1081-93, 2010.

V. Pavone, G. Gaeta, A. Lombardi, F. Nastri, O. Maglio, C. Isernia, and M. Saviano, Discovering protein secondary structures: Classification and description of isolated alpha-turns. *Biopolymers*, 38: 705-721, 1996.

B.J. Polacco and P.C. Babbitt, Automated discovery of 3D motifs for protein function annotation. *Bioinformatics*, 22(6):723-730, 2006.

G. Pugalenthi, K.K. Kumar, P.N. Suganthan and R. Gangal, Identification of catalytic residues from protein structure using support vector machine with sequence and structural features. *Biochemical and Biophysical Research Communications*, 367(3): 630-634; 2008.

L.R. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, *Proc. of the IEEE*, 77(2): 257-85, 1989.

S. Rackovsky, On the nature of the protein folding code. *Proc. Natl Acad. Sci. USA*, 90: 644-648, 1993.

L. Regad, J. Martin and A.C. Camproux, Identification of non random motifs in loops using a structural alphabet. *Proceedings of IEEE Symposium on computational intelligence in bioinformatics and computational*, 92-100, 2006.

L. Regad, F. Guyon, J. Maupetit, P. Tuffery, A.C. Camproux, A Hidden Markov Model applied to proteins 3D structures analysis. *Computational statistics and data analysis*, 52: 3198-3207, 2008.

L. Regad, J. Martin and A.C. Camproux, Dissecting protein loops with a statistical scalpel: functional implication of structural motifs. *BMC Bioinformatics*, in revision.

M.J. Rooman, J. Rodriguez, and S.J. Wodak, Automatic definition of recurrent local-structure motifs in proteins. *J. Mol. Biol.*, 213: 327-336, 1990.

M. Saraste, P.R. Sibbald and A. Wittinghofer, The P-loop: a common motif in ATP- and GTP-binding proteins. *Trends in Biochemical Science*, 15: 430-434, 1990.

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.

G. Schwartz, Estimating the dimension of a model. *Annals of statistics*, 6: 461-64, 1978.

C.J.A. Sigrist, L. Cerutti, E. de Castro, P.S. Kangendijk-Genevaux, V. Bulliard, A. Bairoch and N. Hulo, PROSITE, a protein domain database for functional characterization and annotation. *Nucleic Acids Research*, 38: D161-D166, 2010.

N., Siva, 1000 Genomes project. *Nature biotechnology*, 26(3): 256, 2008.

P.E. Smith, H.D. Blatt, and B.M Pettitt, A simple two-dimensional representation for the common secondary structural elements of polypeptides and proteins. *Proteins*, 27: 227-234, 1997.

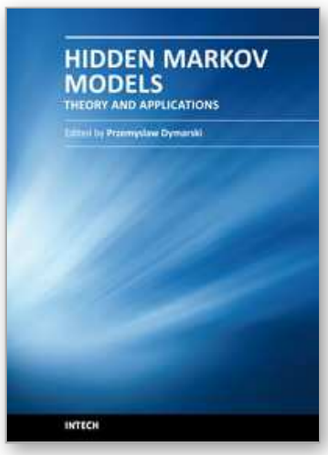
A.V. Tendulkar, M. Krallinger, V. de la Torre, G. Lopez, P.P.Wangikar and A. Valencia, FragKB: structural and literature annotation resource of conserved peptide fragments and residues. *PLoS one*, 5(3), 2010.

R. Unger, D. Harel, S. Wherland and J.L. Sussman, A 3D building blocks approach to analyzing and predicting structure of proteins. *Proteins*, 5: 355-73, 1989.

R.H. Waterston, E.S. Lander and J.E. Sulston, On the sequencing of the human genome. *Proc. Natl. Acad. Sci.USA*, 99: 3712-16, 2002.

C.H. Wu, R. Apweiler, A. Bairoch, D.A. Natale, W.C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M.J. Martin, R. Mazumder, C. O'Donovan, N. Redaschi and B. Suzek, The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res.*, 34: D187-D191, 2006.

L. Zou, Z. Wang, Y. Wang and F. Hu, Combined prediction of transmembrane topology and signal peptide of beta-barrel proteins: using a hidden Markov model and genetic algorithms. *Comput Biol Med.*, 40(7): 621-8, 2010.



Hidden Markov Models, Theory and Applications

Edited by Dr. Przemyslaw Dymarski

ISBN 978-953-307-208-1
Hard cover, 314 pages
Publisher InTech
Published online 19, April, 2011
Published in print edition April, 2011

Hidden Markov Models (HMMs), although known for decades, have made a big career nowadays and are still in state of development. This book presents theoretical issues and a variety of HMMs applications in speech recognition and synthesis, medicine, neurosciences, computational biology, bioinformatics, seismology, environment protection and engineering. I hope that the reader will find this book useful and helpful for their own research.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen