Journal of
Data Mining in Genomics & Proteomics

Reynes et al., J Data Mining Genomics Proteomics 2015, 6:1
DOI: 10.4172/2153-0602.1000167

Research article Open Access

Prediction of Structural Patterns of Interest from Protein Primary Sequence through Structural Alphabet: Illustration to ATP/GTP Binding Site Prediction

Christelle Reynes[1*], Leslie Regad[2,3], Robert Sabatier[1] and Anne-Claude Camproux[2,3]

[1]Facult´e de Pharmacie, Université Paris-Sud 5, EA 2415, F-34093, Montpellier, France
[2]Universit´ Paris Diderot, Sorbonne Paris Cit´e, UMR-S 973, F-75205, Paris, France
[3]Inserm, U973, F-75205, Paris, France

Abstract

The prediction of particular structural motifs associated to biological functions or to structure is of utmost importance. Given the increasing availability of primary sequences without any structure information, predictions from amino-acid (AA) sequences are essential. The proposed prediction method of structural motifs is a two-step approach based on a structural alphabet. This alphabet allows encoding any 3D structure into a 1D sequence of structural letters (SL). First, basic correspondence rules between AA and SL are learnt through genetic programming. Then, a Hidden Markov Model is learnt for each beforehand identified motif of interest. Finally, a probability to correspond to a given 3D motif for any given amino-acid sequence is provided. The method is applied on ATP binding sites to compare the efficiency of our method to other ones for a classical function. Then, the method ability to learn motifs corresponding to more rarely predicted functions or to other types of motifs is illustrated.

Keywords: Availability; Structural motifs

Introduction

Analysis of the tri-dimensional (3D) protein structures is often based on its simplification into secondary structures, including the well-known repetitive and regular regions: α-helix and β-strand, which respectively represent about 30 and 20% of protein residues. The remaining residues constitute a category called loop, often considered as strongly variable. Although the prediction of secondary structure types can be achieved with a success rate of 80%, the description of the secondary structure of a protein does not provide an accurate enough characterization to allow characterization of the complete structure of proteins [1].

To avoid this limitation, sereval studies [2–10] have focused on the identification of a detailed and systematic decomposition of protein structures into a finite set of generic protein fragments. These libraries provide an accurate approximation of protein conformation. They are used to classify protein structures [11], to identify structural changes across proteins in the same SCOP class [12], to find compatible folds for amino acid sequences [13] and to analyze the functional local motions during molecular dynamics [14]. However, the majority of these libraries do not consider the rules that govern the assembly process of the local fragments to produce a protein structure. To take into account these rules, Camproux et al.[9] established a library of fragments based on the Hidden Markov Model approach, called HMM-SA (Hidden Markov Model - Structural Alphabet). It is a collection of 27 structural prototypes of four residues, labeled by {A-Z, a} and named structural letters (SLs). It permits the simplification of all 3D protein structures into one-dimensional (1D) sequences of SLs. HMM-SA is an effective and relevant tool for the study of protein structures [15], protein contacts [16], protein deformations [17], to search for 3D similarity across proteins [18] and to predict the conformation of peptides in aqueous solutions from their amino acid (AA) sequences [19, 20].

HMM-SA particularly provides an accurate description of protein loop structures through 18 specific structural letters. Based on this observation, we developed the notion of SL word to rapidly extract structural motifs from protein loops [15]. SL words correspond to four successive SLs, extracted from SL-sequences corresponding to loop structures. It has been shown that recurrent SL words correspond to clusters of seven-residue fragments with similar structures and AA preferences: these words are named structural words [21]. This method does not require pairwise comparisons of fragments and allows to show that protein loops contain repetitive and regular regions.

Recently, the link between structural words extracted from protein loops [21] and protein functions has been studied. In this study, we supposed that a structural word specific to a protein family is likely to ac-count for a structural motif important to the family function. We started from a set of protein structures encoded into HMM-SA and grouped according the SCOP superfamily ID [22]. To quantify the specificity of a structural word to protein families, we computed the over-representation of each structural word in SCOP superfamilies using SPatt [23]. This allowed us to distinguish two types of over-represented structural words within loops according to their importance either for structure or for protein function. Thus, this method allows extracting some functional motifs without pairwise comparisons of fragments [24].

Coupling our previous works, we have recently developed a web server, named SA-Mot (Structural Alpha-bet - Motif, http://sa-mot.mti.univ-paris-diderot.fr) [25]. This webserver allows the analysis of protein loop structures by extracting structural motifs important for both structure and function of protein.

However, it could be really interesting to be able to predict the presence of such identified functional structural motifs directly from the AA sequence. This work focuses on the prediction of a given structural

***Corresponding author:** Christelle Reynes, Université Paris-Sud 5, Facult´e de Pharmacie, EA 2415, F-34093, Montpellier, France, Tel: +33411759683; E-mail: christelle.reynes@univ-montp1.fr

Received December 15, 2014; **Accepted** January 30, 2015; **Published** February 07, 2015

Citation: Reynes C, Regad L, Sabatier R, Camproux AC (2015) Prediction of Structural Patterns of Interest from Protein Primary Sequence through Structural Alphabet: Illustration to ATP/GTP Binding Site Prediction. J Data Mining Genomics Proteomics 6: 167. doi:10.4172/2153-0602.1000167

word of interest directly from AA sequence. Our prediction method is divided into two steps: first, each four AA sequence is assigned a profile of potential SLs thanks to Boolean trees aiming at extracting sequence information. Then, to take into account the successive SL dependencies, the prediction of words of successive SLs is assembled through a Hidden Markov Model (HMM). This step aims at computing a score for the probability of finding a given structural motif behind the considered sequence. To illustrate the efficiency of the method concept, it is first applied to the prediction of a structural word located in ATP-binding sites. Then, two other examples of applications are given.

Methods

The datasets

In the learning stage, we used 7,656 different proteins from the Protein Data Base (PDB, www.pdb.org, [26]) with at most 25% of sequence identity. 92,832 loops were extracted from these proteins. These data were used to learn the parameter values of the prediction method described in Prediction method Section. In the validation stage, 6213 proteins were extracted from the PDB in the same sequence identity conditions and after removal of common proteins. They were splitted into 66508 different loops. The results in Result Section were obtained on the latter dataset.

Data encoding through HMM-SA

HMM-SA [9] aims at discretizing the conformational space of four-residue fragments into 27 structural states called SLs. It is based on a HMM allowing taking into consideration dependencies between successive letters. Indeed, some transitions between certain SLs are favored, so it can be really informative in a prediction objective to take into account such dependencies. Four SLs particularly describe conformation of α-helices (namely A, A, V and W), five SLs describe β-sheets (L, M, N, T and X) and the remaining 18 letters characterize loops.

From this alphabet, it is possible to encode any 3D structure into a 1D string of SLs. In this goal, Viterbi or forward/backward algorithms [27,28] can be used to find the most probable sequence of SLs according to a given structure. In the following sections, focus will be put on four SL words, as illustrated in Figure 1.

From now on, the goal is to be able to model the link between AAs and SLs. Indeed, it is impossible to find a perfect application from the set of four AA sequences ($20^4 \approx 1.6 \times 10^5$ possibilities) onto the 27 possible SLs. Indeed, the same AA sequence can be encoded into different SLs (due to flexibility) and the same SL can be obtained from different four-AA sequences.

Motif extraction method

The identification of motifs of interest, we will focus on in this paper, is presented in [24]. After 3D structure encoding into SLs, this method focuses on four SL-words corresponding to structural motifs consisting of seven-residue fragments sharing similar geometry and AA specificities [21]. This length has been chosen to obtain satisfying representativities [29]. Thus, thanks to HMM-SA, extraction of structural motifs from protein structures is translated into the extraction of four-SL words from the SL sequences.

It has been shown that some of these motifs are important for the protein structure, and other are implied in binding sites of small ligands (ATP/GTP, calcium, NAD(P)) [24].

Prediction method

First step: from four amino acids to one structural letter

The goal of the first step is to build classifiers characterizing AA sequences and allowing distinguishing between the different SLs. A usual one-versus-one classification process is used. In this way, each classifier deals with a simpler problem. Hence, a classifier will be built to compare each pair of SLs leading to optimize 351 classifiers.

Structure of a classifier allowing the discrimination of two SLs In order to find robust information about the relationship between AA sequence and SLs and to avoid any kind of over fitting, a very simple use of information is proposed. The proposed classifier can be seen like a tree. Figure 2 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 no such letter at such 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 out of the two SLs compared through this classifier. A jury of 351 classifiers is finally obtained providing 351 votes. Hence, a kind of profile in SLs is obtained for a four-residue fragment.

Scoring a classifier allowing the discrimination of two SLs To optimize each classifier, we define a *fitness* value, which quantifies its quality. It consists in three parts: entropy gain (related to discrimination ability), tree complexity and representativeness of the obtained decision rule.

Figure 1: Illustration of the way of fragment corresponding to seven AA (AA_1, . . . ,AA_7) is encodedinto a four SL word (SL_1, SL_2, SL_3, SL_4).

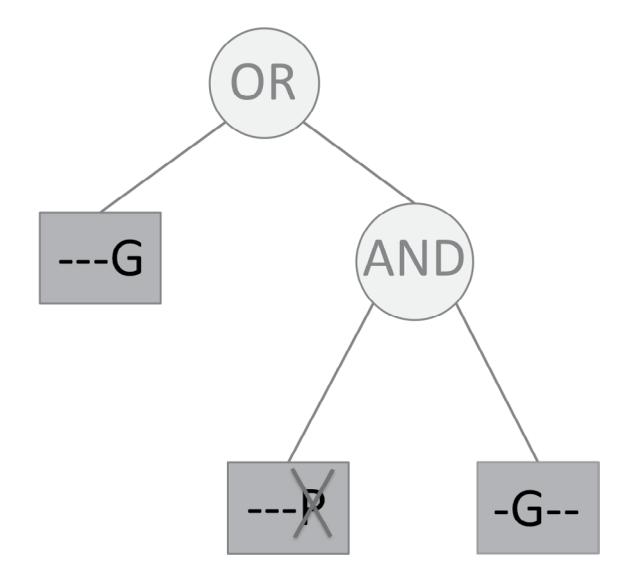

Figure 2: Example of classifier used to discriminate between two letters A and B: if there is *G insecond position AND no P in fourth position* OR a G in fourth position then the sequence is affected to A else to B.

Entropy term: The global entropy (we refer to Shannon entropy [30]) associated with a sample containing the observations of two SLs i and $j (i \neq j \; and \; (i,j) \in \{1,2,...,27\}^2)$ can be defined as:

$$H(i,j) = -(p_{ij}^{(i)} \log(p_{ij}^{(i)}) + p_{ij}^{(j)} \log(p_{ij}^{(j)})) = -(p_{ij}^{(i)} \log(p_{ij}^{(i)}) + (1 - p_{ij}^{(i)}) \log(1 - p_{ij}^{(i)}))$$

where $p_{ij}^{(k)}$ is the proportion of SL $k (k \in \{i,j\}$ in the sample containing all the observations of SLs i and j $(i \neq j \; and \; (i,j) \in \{1,2,...,27\}^2)$ and no other SL. The corresponding entropy gain can be defined as follows:

$$G(i,j) = H(i,j) + \pi_1(p_{ij,1}^i \log(p_{ij,1}^i) + p_{ij,1}^j \log(p_{ij,1}^j)) + \pi_2(p_{ij,2}^i \log(p_{ij,2}^i) + p_{ij,2}^j \log(p_{ij,2}^j)$$

where $p_{ij,l}^{(k)}$ is the proportion of the SL k in the l-th $(l \in \{1,2\}$ subsample of the sample containing all the observations of SLs i and j provided by this classifier, $\pi_k (k \in \{1,2\})$ is the proportion of SLs contained in subsample k. Thus, the entropy gain is the difference between the global entropy and the weighted sum of entropies of the two subsamples.

Parsimony term: Penalizing the tree complexity is a way to avoid overfitting. The number of leaves of a tree, denoted nbf, is chosen as a quantification of its complexity. To normalize the variations of the complexity term in the fitness function, the following term is used:

$$penal_1 = \frac{2^{D-1} - nbf}{2(2^{D-2} - 1)}$$

where D is the maximum authorized depth of any tree.

Representativity term: If the classifier makes a relatively small subsample but containing a majority of only one SL, its entropy will be satisfying whereas it is not representative of either of the two SLs. This kind of behaviour will be penalized by the following term:

$$penal_2(i,j) = \left| \frac{p_{ij,1}^i}{p_{ij}^i} - \frac{p_{ij,1}^j}{p_{ij}^j} \right|$$

The higher this term, the better the classifier.

Global criterion: Finally, the global fitness function can be expressed as follows:

$$fit(i,j) = G(i,j) + \alpha(penal_1 + penal_2(i,j))$$

where α allows to balance entropy gain term and penalizations. In our applications, we experimentally chose α=0.05. It is chosen as the quantification of how much entropy gain we are ready to lose to be able to delete one leaf in the tree.

Optimization of each classifier: The kind of chosen classifier leads to use genetic programming (GP, [31,32]). Indeed, the cardinality of the set of all possible trees is huge. Hence, a heuristic method has to be used.

GP is a symbolic approach to computer programs induction. It is a kind of genetic algorithm [33] where potential solutions are programs defined on a landscape determined by the objective task. In our context, a program will be a classifier. The GP will allow the evolution of a population of potential classifiers through the use of mutation and cross-over operators.

In the choice between two SLs, some of them are easier to discriminate through their sequence than other ones. For example, SLs B and M are very well discriminated: one out of the two subgroups obtained after applying the classifier contains 3.2% of the B SLs and 98.0% of the M. On the other hand, SLs a and M are particularly difficult to distinguish through their sequence.

In fact, partly due to flexibility, it is impossible to build a perfect bijection between those two kinds of objects. That is why structure prediction from sequence is such a difficult task. However, this first step allows to limit the possibilities by giving probabilities to each SL given four AAs.

Then, in the second step, the trick of our method is not to claim to be able to predict the best four SLs encoding for seven given AAs but, on the contrary, to verify how compatible the seven AAs are with a given target four SLs pattern. This considerably simplifies the question while allowing to predict different patterns. In addition to this viewpoint change, the second step takes advantage of another kind of information: the structural dependencies between successive SLs.

Second step: Looking for one specific structural word

The aim of this step is to decide, given the combined results of the first step for four consecutive (and overlapping) SLs and through a scoring function, if the conformation adopted by the considered seven residue fragment is likely to be encoded by a given four SL word of interest.

Structural word modelling and application to prediction : A dependency exists between successive SLs, especially because of overlaps. Hence, a HMM has been chosen to model the link between first step outputs and a given structural word. It is described in Figure 3. In this model, hidden states are the true SLs while observed states are outputs of step 1 for the corresponding sequence. Arrows between X_i and X_{i+1} symbolize the dependency between successive letters called transition probabilities in HMM context and arrows between X_i and O_i represent the link between true SLs and step 1 outputs, namely the output probabilities.

Then, we compute the probability of the four true SLs being the target functional pattern given the step

1 outputs for four successive four AA fragments.

$$P(X_{1:4} | O_{1:4}) = P(X_1, X_2, X_3, X_4 | O_1, O_2, O_3, O_4) \quad (1)$$

According to the chosen model,

$$P(X_{1:4} | O_{1:4}) = P(X_1 | O_1) \prod_{i=2}^{4} P(X_i | X_{i-1}) P(X_i | O_i)$$

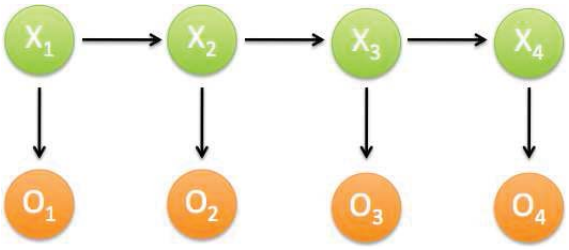

Figure 3: Structure of the HMM used to model the relationship between first step outputs and *true* SLs for a seven residue fragment: (X_1,X_2,X_3,X_4) are the *true* SLs and $O_i = (o^1_i, o^2_i, \ldots, o^{351}_i)$ is the vector of votes obtained from step 1 for the four AA fragment encoded by X_i.

Now, P (X_i |O_i) has to be computed. Assuming that the results of different trees are independent,

$$P(X_i \left| o_i^j \right.) = \frac{P(o_i^j \left| X_i \right.)P(X_i)}{P(o_i^j \left| X_i \right.)P(X_i) + P(o_i^j \left| \overline{X}_i \right.)P(\overline{X}_i)} \quad (2)$$

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,

$$P(X_i \left| O_i \right.) = P(X_i \left| o_i^1, o_i^2, ..., o_i^{351} \right.) = \prod_{j=1}^{351}$$

Finally, $P(X_i)$, $P(X_i \left| X_{i-1} \right.)$ and $P(X_i \left| \overline{X}_{i-1} \right.)$ are estimated on the dataset. The two-step proposed method is summed-up in Figure 4.

Evaluation criteria

In order to evaluate the method efficiency for a given structural word, several criteria can be used. In our context, two classes are defined: the fragments corresponding to the considered structural motif and the fragments that do not. Then, according to the final decision made, four cases can be defined: True Positives (denoted TP) occurring when a fragment really encoded into the considered structural word is predicted as corresponding to this word. In the same way, we have False Negatives (FN), False Positives (FP) and True Negatives (TN). Then, sensitivity (Se) and specificity (Sp) can be defined as follows:

$$Se = \frac{TP}{TP + FN} and\ Sp = \frac{TP}{TN + FP}$$

A threshold has to be chosen to decide whether or not a given structural word corresponds to the consid- ered structural word. Hence, the values of Se and Sp obviously depend on this chosen threshold. More specifically, increasing the threshold will result in improving Sp but decreasing Se. To simultaneously study, Se and Sp, a ROC (Receiver Operating Characteristic, [34]) curve can be built. It consists in plotting the Se according to (1–Sp) values for many threshold values. It is possible to measure the quality of a classifier by the Area Under the Curve (AUC): it will be 1 for perfect classifier and 0.5 for random ones. Then, the closer to 1 an AUC, the better the corresponding classifier.

Results

We have previously developed a method allowing the prediction of seven-AA structural motifs of interest from protein structures based on the 4 SL-words and the structural alphabet HMM-SA [21,24]. However this method is not applicable when the protein structure is not known. To avoid this problem, we propose a new method allowing the prediction of the structural motifs directly from the AA sequences.

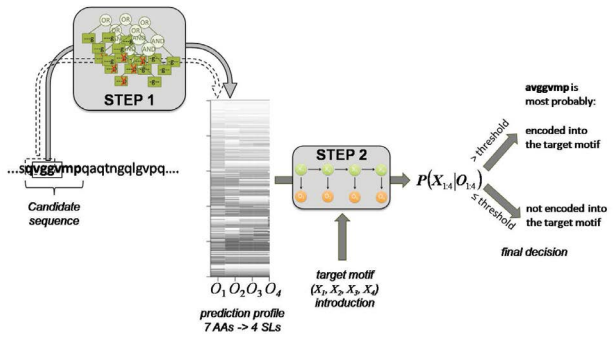

Figure 4: Method global unfolding.

In the following, the new prediction method concept is tested on three different structural motifs of interest previously identified by structural word approach [24] . The first one will be the most detailed. It concerns the structural word YUOD located in ATP-binding sites and allows the prediction of one kind of NP bind sites. This kind of site is quite commonly predicted and will allow to illustrate firstly the overlap between the structural motifs and functional annotations and secondly the comparison with two other prediction methods: SitePredict [35] and PROSITE [36]. The second example concerns a functional structural word RUDO located in the SAM/SAH-binding sites. Finally, in order to show the generality of the method, it will be applied to a structural word HBDS, which describes turn motifs a very recurrent motif in loop structures.

Prediction of YUOD motif allowing the prediction of ATP-binding sites

YUOD functional specificity

Fragments encoded into YUOD (Figure 5a presents a superimposition of such fragments) have been identified in [24] as over-represented in the SCOP superfamily P-loop-containing nucleotide triphos- phate hydrolase. Subsequently, it has been shown that YUOD is located in binding sites to ATP/GTP (Adenosine/Guanine-5'-triphosphate), which provide by hydrolyze the required energy for chemical and metabolism reactions in cells. YUOD is associated to ATP/GTP binding sites (denoted NP bind) with a sensitivity of more than 35% meaning that more than one third of ATP/GTP binding sites adopt con- formations described by YUOD. This figure has to be interpreted knowing that NP bind sites are mostly divided into two distinct parts, one being very often encoded into YUOD.

In our database, YUOD was found 180 times in 178 different proteins. By comparing the location of YUOD fragments and functional NP bind annotations extracted from Swiss-Prot [37], the NP bind site is NOT overlapping YUOD in only 10 proteins, in all other cases (93.5%) at least one out of the seven positions constituting YUOD is contained in the annotated site. This confirms the strong link between YUOD and NP bind sites. Moreover, this structural word has a high sequence specificity (Figure 5b). Thus, this structural word is a very good candidate for prediction NP bind motif from AA sequence.

YUOD prediction

To test the efficiency of our model to predict the YUOD motif directly from its AA-sequence, we applied our model on the 178 proteins containing this word. The ROC curve associated to the logarithm of the computed probability (given by Eq. 1) is shown in Figure 6a. It displays sensitivity and specificity according to the probability threshold chosen to split the words into YUOD and YUOD (not YUOD). The associated AUC is 0.987. Hence, the computed probability is really efficient to identify YUOD among all other words. Such a 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

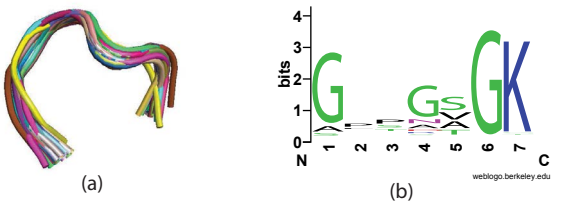

Figure 5: (a) Representation of several fragments encoded into YUOD. (b) Weblogo of the AA sequences encoded into YUOD.

probability thresholds can be defined providing different balances between sensitivity and specificity. Some interesting threshold values and their corre- sponding parameters are enclosed in Table 1. Very high values of specificity have been chosen, indeed the YUOD class is really large and then only 1% of false positive can be a large number when applied to big or several proteins.

An example of YUOD detection is given in Figure 6b. It concerns the Circadian clock protein kinase kaiC, chain A (pdb ID: 2gbl A). It originally contains two true YUOD occurrences and four have been predicted through our model. Two of them (# 1 and 2) are exactly located at co-crystallized ATP binding sites (A and B). Moreover, between 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 SwissProt but false positives may be true positive that have not yet been experimentally verified.

From YUOD to NP bind sites prediction

62.0% of NP bind sites (recovered from either SwissProt or PDB co-crystallization study) are predicted as YUOD by our method by using a threshold of -4732, 91.5% with a threshold of -4805 and 96.0% with a threshold of -4829. Yet, only 93.5% of those sites are really encoded into YUOD. It has been shown that NP bind sites could be encoded into words close to YUOD such as KUOD [24]. Hence, not only our method is able to very efficiently predict the true presence of YUOD but it also enables to recover motifs which are close to YUOD but not exactly encoded into it (6.5% in this case) whereas they are really associated to NP bind sites. In this way, our method takes into account the structural variability of the binding site.

Comparison with other NP bind sites prediction methods

Comparison with SitePredict [35] SitePredict uses Random Forest [38] to combine multiple properties associated with ligand binding sites in order to predict which residues in a protein bind the ligand. It is particularly interesting to compare SitePredict with our method because it also provides a probability score which reflects the likelihood that the site binds a particular ligand.

SitePredict was applied to predict the NP bind site in the 178 proteins. No NP bind site has been pre- dicted in 8.1 % of the proteins containing verified NP bind sites. A potential NP bind site has been predicted at the wrong place in 15.5 % of those proteins. The remaining proteins have been given a positive probability that a NP bind site overlaps the true site. However, 28.1 % of those sites have less than 0.5 probability of being a NP bind site. 28.1% of those sites have a probability higher than 0.6995 which corresponds to the sensitivity obtained by our method for a threshold of -4732, whereas we retrieve 60% of the true NP bind sites with this threshold. The SitePredict AUC is 0.6179. This value is not directly comparable to the one obtained for YUOD recognition by our method (that is to say 0.9866) as it does not concern YUOD but NP bind sites. However, as it has been verified that YUOD is overlapping the true NP bind site in 93.5% of this dataset, our method is likely to perform better. Moreover, for a specificity of 0.95, SitePredict achieves a sensitivity of 0.37 and a specificity of 0.99 corresponds to a sensitivity of 0.22. Hence, it appears that SitePredict is less sensitive than our method to find NP bind sites.

Comparison with PROSITE [36]

PROSITE is a database of entries which describes not only protein domains, families and functional sites but also associated patterns and profiles to identify them. It is associated to ProRule which is a database of rules whose predictive ability is enhanced by taking into account information about functionally and/or structurally critical AAs. The motif corresponding to NP bind is denoted PS00017. The corresponding consensus pattern is [AG]-x(4)-G-K-[ST]. It can be noticed that this pattern has much in common with YUOD logo (Figure 5b). However, due to the deterministic aspect of the corresponding rule, there is no probability score associated to the prediction. Hence, it is not possible to compute an AUC for this method. Nevertheless, we can retrieve that none NP bind site has been predicted in 13.0% of the proteins containing verified NP bind sites. A potential NP bind site has been predicted at the wrong place in 6.8% of the proteins. Hence, the method performs better than SitePred to identify the real location of the NP bind site but it is less efficient to identify whether or not a given protein contains a NP bind site. This limitation of PROSITE is probably due to its deterministic way of making a prediction, which makes it really precise but a bit too specific.

Finally, it is really difficult to compare different methods of function prediction given that the way a function is defined can be not unique. However, it appears that the proposed method is really powerful. Furthermore, it could be even better by taken into account alternative motifs describing different kinds of NP bind sites.

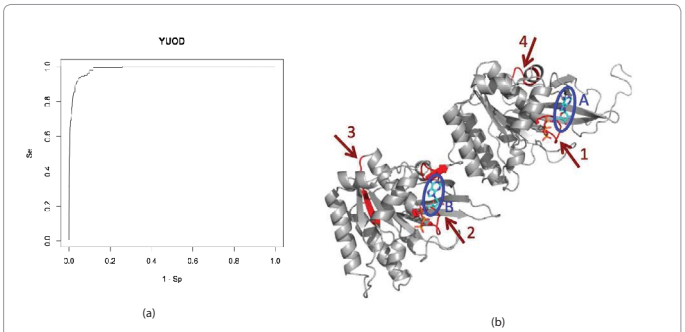

Figure 6: (a) ROC curve associated with the probability of having YUOD for a given seven AA fragment (Se=sensitivity, Sp=specificity). (b) 3D representation of 2gbl A co-crystallized with two ATP molecules (indicated by lettered circles). The fragments identified as YUOD are indicated with numbered arrows.

Other application examples

Prediction of a SAH/SAM-binding site specific motif using RUDO word

In this section, we focus on the RUDO prediction. This structural word has been identified to be most of time associated to SAH/SAM (S-adenosyl-methionine/S-adenosyl-homocysteine)-binding site in Swis- sProt database. SAH/SAM are molecules associated to some methylation processes and are particularly studied in the context of antiviral drugs research. Yet, to the extent of our knowledge, the prediction of their binding to proteins is not proposed by existing method. As RUDO has a good sequence specificity [24], it is a good candidate to predict SAH/SAM-binding sites.

To test the efficiency of our model to predict the RUDO word, which is less studied than YUOD word, we applied our model to the 39 different proteins containing this word. The obtained AUC is 0.9606. The specificity and sensitivity obtained with different thresholds are given in Table 1. Thus, results are satis- fying and allow to recover more than two thirds of the RUDO motifs without wrongly assigning more than 1% of the other words.

An illustration can be found in Figure 7a. It concerns isoloquiritigenin 2'-O-methyltransferase (PDB ID:1fp1) which was co-crystallized with a SAH molecule. Four words were predicted as RUDO with a threshold of -4712 whereas only one has been encoded as RUDO. However, by looking of the 3D conformation, it appears that all four identified fragments are really closed to the ligand. Thus, the method that uses the HMM-SA as a tool to discover patterns, is not limited to the fragments being strictly encoded but is also able to discover fragments with close structures.

Prediction of HBDS word allows the prediction of a specific turn

The four-SL word HBDS (the corresponding fragment conformations are shown in Figure 7b) can be linked to turns [24]. Turns are elements of secondary structure where the polypeptide chain reverses its overall direction. As they may play a role in protein folding, they are likely to allow chains to become closer making some interactions possible [39].

As HBDS presents AA specificities, it is a good candidate to predict this turn motif from AA sequence. To test this hypothesis, we applied our model to the 1363 different proteins containing this word, seen 1633 times in this set of proteins. The obtained AUC is 0.9359. Table 1 indicates the specificities and sensitivities. 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).

Discussion

We have previously developed an original approach based on the structural alphabet HMM-SA and the notion of structural words to extract structural motifs of importance for structure and/or function of proteins. We propose a new method to predict these important motifs directly from their AA sequence. The proposed method is based on the identification of structural words of interest on 3D structures simplified through HMM-SA as proposed in [24]. The input data of the described method are only AA sequences and as a consequence, only structural words having sequence specificities will be likely to be handled with this method. But for this kind of motifs, the method is really powerful.

Thanks to the two-step approach, as much information as possible is extracted from data. The dependence between AA sequences and 3D structures is learnt through HMM-SA. It provides, for a given four-AA fragment a profile of potential corresponding SLs. The second step firstly quantifies and uses the strength of dependency between AAs and SLs confusion probabilities: some observations will be really trusted whereas others will be considered with care. Secondly, the dependency between successive SLs is taken into consideration by the computation of transition probabilities. It has to be noticed that the first step is motif independent whereas the model of the second step is motif specific. A complete model is obtained by the combination of both steps providing for each input AA fragment a probability of correspondence with the target motif. This combination of viewpoints between motif independent and motif dependent steps is certainly a strength of our method, the method becoming more and more driven as the process progresses.

This method is able to identify fragments as being close to the target motif even if this fragment would not be encoded into the exact previously identified target word. Hence, relying on sequences has drawbacks as it is less conserved than structure by evolution but it can be a way to overcome some cases of flexibility. Actually, HMM-SA encoding and the proposed prediction method are interestingly complementing each other in the prediction of motifs of interest.

Furthermore, the important adaptability of the prediction method is of big interest. Even if our method has been developed as a continuation of the work performed in [24], it can be immediately generalized to words of more than four SLs, no additional learning would be required compared with any other new four-SL word. Moreover, it is completely possible to identify 3D motifs of interest by any method, to encode it into HMM-SA and to build the model corresponding to the obtained word. For a given loop it took less than one second to compute the probability for one given motif to be present at each position. Concerning learning it will obviously depend on the alphabet and length of the word. For example, in our case it took about 5 hours to learn YUOD on our dataset. It must be noticed that the genetic programming step has to be performed only once for a given alphabet, this is the most time expensive step, then the second most expensive step is the learning of a given word, which is done once for each new word. Those two

SL word	YUOD			RUDO			HBDS		
Threshold	-4829	-4805	-4732	-4903	-4806	-4712	-4013	-3844	-3777
Specificity	90.02	95.07	99.00	90.00	95.00	99.00	90.17	95.11	98.88
Sensitivity	97.81	93.44	69.95	87.18	84.62	69.23	84.71	71.07	28.93

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

steps do not concern daily use of the method which only implies application of learnt most and which takes up to a few seconds according to the protein length.

Moreover, concerning the size of the learning dataset, as illustrated through the examples, it can be really variable (from 35 to 1633 occurrences according to the motif). Obviously, the bigger the learning set, the higher the reliability of the obtained model. But, when the sequence specificity is strong, only a small number of learning observations can be used to learn interesting models.

The only variable parameter depending on the learnt motif is the threshold chosen for the log(probability) to decide whether or not a fragment corresponds to the considered motif. Preliminary studies seem to indicate that this threshold depends on the strength of the sequence specificity of the motif. Anyway, the ROC curve automatically obtained during model building can be used to set the threshold value. In our approach, given the unbalanced ratio between the two groups, we chose to set the threshold thanks to specificity values.

The limits of the method are interlocked with its strengths. First of all, as previously indicated, only motifs with sequence specificities can be predicted. Moreover, a 1D intermediate is necessary. HMM-SA has been used because of its very interesting abilities of precise description especially for loops, but the same methodology could be applied with any other alphabet.

This work shows that the prediction of functional words, i.e., structural words located in a functional site, allows the prediction of functional sites such as ATP, SAM/SAH-binding sites. Thus our method could be used to help the prediction of protein function. Moreover, it is completely possible to learn several motifs linked to a given function and to give a global prediction for all of them. This will be quickly possible by the identification of new motifs which is in progress.

The method codes can be obtained by request to the first author.

References

1. Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22: 2577-2637.
2. Unger R, Harel D, Wherland S, Sussman JL (1989) A 3D building blocks approach to analyzing and predicting structure of proteins. Proteins 5: 355-373.
3. Camproux AC, Tuffery P, Chevrolat JP, Boisvieux JF, Hazout S (1999) Hidden Markov model approach for identifying the modular framework of the protein backbone. Protein Eng 12: 1063-1073.
4. de Brevern AG, Etchebest C, Hazout S (2000) Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. Proteins 41: 271-287.
5. Bystroff C, Thorsson V, Baker D (2000) HMMSTR: a hidden Markov model for local sequence-structure correlations in proteins. J Mol Biol 301: 173-190.
6. Micheletti C, Seno F, Maritan A (2000) Recurrent oligomers in proteins: an optimal scheme reconciling accurate and concise backbone representations in automated folding and design studies. Proteins 40: 662-674.
7. Pandini A, Fornili A, Kleinjung J (2010) Structural alphabets derived from attractors in conformational space. BMC Bioinformatics 11: 97.
8. Kolodny R, Koehl P, Guibas L, Levitt M (2002) Small libraries of protein fragments model native protein structures accurately. J Mol Biol 323: 297-307.
9. Camproux AC, Gautier R, Tufféry P (2004) A hidden markov model derived structural alphabet for proteins. J Mol Biol 339: 591-605.
10. Sander O, Sommer I, Lengauer T (2006) Local protein structure prediction using discriminative models. BMC Bioinformatics 7: 14.
11. Le Q, Pollastri G, Koehl P (2009) Structural alphabets for protein structure classification: a comparison study. J Mol Biol 387: 431-450.
12. Joseph AP, Valadié H, Srinivasan N, de Brevern AG (2012) Local structural differences in homologous proteins: specificities in different SCOP classes. PLoS One 7: e38805.
13. Mahajan S, de Brevern AG, Sanejouand YH, Srinivasan N, Offmann B (2015) Use of a structural alphabet to find compatible folds for amino acid sequences. Protein Sci 24: 145-153.
14. Pandini A, Fornili A, Fraternali F, Kleinjung J (2013) GSATools: analysis of allosteric communication and functional local motions using a structural alphabet. Bioinformatics 29: 2053-2055.
15. Regad L, Guyon F, Maupetit J, Tuffery P, Camproux AC (2008) A hidden markov model applied to the protein 3d structure analysis. CSDA 52: 3198-3207.
16. Martin J, Regad L, Etchebest C, Camproux AC (2008) Taking advantage of local structure descriptors to analyze interresidue contacts in protein structures and protein complexes. Proteins 73: 672-689.
17. Martin J, Regad L, Lecornet H, Camproux AC (2008) Structural deformation upon protein-protein interaction: a structural alphabet approach. BMC Struct Biol 8: 12.
18. Guyon F, Camproux AC, Hochez J, Tufféry P (2004) SA-Search: a web tool for protein structure mining based on a Structural Alphabet. Nucleic Acids Res 32: W545-548.
19. Maupetit J, Derreumaux P, Tuffery P (2009) PEP-FOLD: an online resource for de novo peptide structure prediction. Nucleic Acids Res 37: W498-503.
20. Maupetit J, Derreumaux P, Tufféry P (2010) A fast method for large-scale de novo peptide and miniprotein structure prediction. J Comput Chem 31: 726-738.
21. Regad L, Martin J, Nuel G, Camproux AC (2010) Mining protein loops using a structural alphabet and statistical exceptionality. BMC Bioinformatics 11: 75.
22. Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247: 536-540.
23. Nuel G, Regad L, Martin J, Camproux A (2010) Exact distribution of a pattern in a set of random sequences generated by a Markov source: applications to biological data. Algorithms Mol Biol 5: 15.
24. Regad L, Martin J, Camproux AC (2011) Dissecting protein loops with a statistical scalpel suggests a functional implication of some structural motifs. BMC Bioinformatics 12: 247.
25. Regad L, Saladin A, Maupetit J, Geneix C, Camproux AC (2011) SA-Mot: a web server for the identification of motifs of interest extracted from protein loops. Nucleic Acids Res 39: W203-209.
26. Berman H, Westbrook J, Feng Z, Gilliland G, Bhat T, et al. (2000) The protein data bank. Nucleic Acids Research 28: 235-242.
27. Viterbi A (1967) Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE Trans Inform Theory IT-13: 260-269.
28. Baum L, Petrie T, Soules G, Weiss N (1970) A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. Ann Math Statist 41(1): 164-171.
29. Regad L, Martin J, Camproux AC (2006) Identification of non random motifs in loops using a structural alphabet. In: Proceedings of IEEE Symposium on computational intelligence in bioinformatics and computational: 28-29 September 2006; Toronto. IEEE, pp. 92-100.
30. Cover T, Thomas J (2006) Elements of information theory. New York: Wiley.
31. Koza J (1992) Genetic Programming: On the Programming of Computers by Means of Natural Selection. London: MIT Press.
32. Langdon W, Riccardo P (2002) Foundations of Genetic Programming. Springer-Verlag, London.
33. Goldberg D (1989) Genetic Algorithms: Search, Optimization and Machine Learning. Addison-Wesley, New York.
34. Zweig MH, Campbell G (1993) Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clin Chem 39: 561-577.

35. Bordner AJ (2008) Predicting small ligand binding sites in proteins using backbone structure. Bioinformatics 24: 2865-2871.

36. Sigrist CJ, Cerutti L, de Castro E, Langendijk-Genevaux PS, Bulliard V, et al. (2010) PROSITE, a protein domain database for functional characterization and annotation. Nucleic Acids Res 38: D161-166.

37. Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, et al. (2005) The Universal Protein Resource (UniProt). Nucleic Acids Res 33: D154-159.

38. Steffens M, Lamina C, Illig T, Bettecken T, Vogler R, et al. (2006) SNP-based analysis of genetic substructure in the German population. Hum Hered 62: 20-29.

39. Fuchs PF, Alix AJ (2005) High accuracy prediction of beta-turns and their types using propensities and multiple alignments. Proteins 59: 828-839.