A new prediction strategy for long local protein structures using an original description.

Aurélie Bornot, Catherine Etchebest, Alexandre de Brevern

A new prediction strategy for long local protein structures using an original description

Aurélie Bornot[1][§], Catherine Etchebest[1] & Alexandre G. de Brevern[1]

[1] Equipe de Bioinformatique Génomique et Moléculaire (EBGM), INSERM UMR-S726, DSIMB workgroup, Université Paris Diderot - Paris 7, Institut National de la Transfusion Sanguine, 6 rue Alexandre Cabanel, 75739 Paris Cedex 15, FRANCE.

[§]Corresponding author

Correspondence to: Aurelie BORNOT (aurelie.bornot@univ-paris-diderot.fr),
Equipe de Bioinformatique Génomique et Moléculaire (EBGM), INSERM UMR-S726, DSIMB workgroup, Université Paris Diderot - Paris 7, Institut National de la Transfusion Sanguine, 6 rue Alexandre Cabanel, 75739 Paris Cedex 15, FRANCE.
Phone/Fax : 33 1 44 49 30 38 / 33 1 47 34 47 31

Abstract:

A relevant and accurate description of three-dimensional (3D) protein structures can be achieved by characterising recurrent local structures. In a previous study, we developed a library of 120 3D structural prototypes encompassing all known 11-residues long local protein structures and ensuring a good quality of structural approximation. A local structure prediction method was also proposed.

Here, overlapping properties of local protein structures in global ones are taken into account in order to characterize frequent local networks. At the same time, we propose a new long local structure prediction strategy which involves the use of evolutionary information coupled with Support Vector Machines (SVMs). Our prediction is evaluated by a stringent geometrical assessment. Every local structure prediction with a Cα RMSD less than 2.5 Å from the true local structure is considered as correct. A global prediction rate of 63.1% is then reached, corresponding to an improvement of 7.7 points compared to the previous strategy. In the same way, the prediction of 88.33% of the 120 structural classes is improved with 8.65 % mean gain. And 85.33% of proteins have better prediction results with a 9.43 % average gain. An analysis of prediction rate per local network also supports the global improvement and gives insights into the potential of our method for predicting super local structures. Moreover, a confidence index for the direct estimation of prediction quality is proposed. Finally, our method is proved to be very competitive with cutting-edge strategies encompassing three categories of local structure predictions.

Introduction

Three-dimensional (3D) structural information is very critical for understanding the functional properties of proteins. Therein, 3D structures are a valuable source of data for understanding their biological roles, their potential implication in diseases, and for progress in drug design [1-3].

It is now admitted that the folded state of proteins, *i.e.* the native 3D structure, may be described by a limited set of recurring local structures [4,5]. Besides, native-like local structures were experimentally observed during folding pathways and, thus, could have a guiding role in the folding process toward the global native structure [4,6]. These observations support the idea that local structures are relevant in characterizing native folds and that the structural information encoded in sequence segments should be predictable and useful for proposing relevant structural models [7,8].

In addition, costs and difficulties associated with experimental determination of protein 3D global structures led the prediction of pertinent structural models from sequence to become a major area of interest. In this context, prediction of accurate local structure characteristics is a promising field and, can provide relevant information and constraints for global structure prediction [9,10]. Indeed, some of the most successful prediction methods in competition at CASP7 in the hardest category, the new fold prediction, used biased sampling of structural fragments and assembly techniques combined in different strategies [11].

Accordingly, fragment libraries or structural alphabets were designed to characterize in the most suitable way, the local structures of all proteins with known 3D structures. These alphabets consist of a finite set of representative structural fragments. They have been reviewed in [12]. They differ by the number of local

structure groups and representatives, the fragments length and the selected criteria and methods used to design them, *e.g.*, [10,13].

We had previously developed such a structural alphabet named Protein Blocks (PBs), with the aim of not only characterizing the local structures observed among the largest set of known proteins 3D structures but also at identifying optimal sequence-structure specificities for prediction purpose [14,15]. It was developed by using an unsupervised cluster analyser and by relying on series of eight dihedral angles characterizing five-residue long fragments observed in protein structures. Thus, this alphabet is optimally composed of 16 structural classes. Each class is represented by a PB, *i.e.* an average dihedral vector. The Protein Blocks alphabet takes into account the preference of consecutive fragments to occur in a sequence and, has been shown to be highly informative and to have the best predictive ability among those tested by Karchin and co-workers [16]. PBs have been designed to describe protein 3D backbones [15] and to predict local structures [13]. They have been successfully used to predict short loop regions [17], to superimpose protein structures [18,19], to mine PDB [20], to help difficult modelling of transmembrane proteins [21], to study mutation structural consequences [22], to rebuild protein structure [23] and for the for the discovering of functional structural motifs [24].

Recently, we proposed an extension of this description with a novel library consisting of 120 overlapping structural classes of 11-residues long fragments [25]. This library was constructed with an original unsupervised structural clustering method called the Hybrid Protein Model (HPM) [26,27]. The Hybrid Protein principle is similar to a self-organizing neural network [28,29]. It was constructed as a ring of *N* neurons (here *N*=120), each representing a cluster of structurally similar 3D fragments encoded into series of PBs. Its training strategy consisted in learning the similarities

between protein fragments by aligning them according to their PB series. Once the HPM was trained, each neuron or cluster was associated with a set of fragments representing a structural class [25]. For each class, a mean representative fragment, or a "local structure prototype" (LSP), was chosen. The 120 LSPs correctly approximated the local structure ensembles. The major advantage of this library is its capacity to capture the continuity between the identified recurrent local structures. Relevant sequence-structure relationships were also observed. This permitted us to develop an original method for the prediction of long local protein structures from sequence. The principal interest of this method is that, for a target sequence of 11-residue long, it proposes a short list of the best structural candidates among the 120 LSPs of the library. This prediction was assessed by a geometrical criterion, *i.e.* a prediction was considered as correct if the Cα RMSD between the best candidate and the true local structure was less than of 2.5 Å. The prediction rate based on single sequence information reached 51.2%, a satisfying result taking into accounts the very large size of the library and the long length of fragments [25]. Compared to other local structure libraries, the long length of the fragments studied here is worth being stressed. As a matter of fact, until our previous work in 2006, no local structure library handled fragments as long as 11 residues for prediction purposes. Actually, libraries have often been developed for fragments encompassing 4 to 9 residues [12]. Only Bystroff and Baker developed the *I-sites* library characterizing fragments of 3 to 15 residues. A prediction method was associated but the results were assessed for fragments of 8 residue-long only [30]. Recently, Beaten and co-workers constructed the BriX database encompassing more than 1000 frequent local conformations ranging from 4 to 14 residues[31]. Similarly, Sawada and Honda developed a database of structural clusters taking into account fragments of 5, 9, 11 and 15 residues[32]. But, in these cases,

prediction methods were not developed. Indeed, 11 residue fragments correspond to long patterns for analysis and prediction of protein local structures. Thus owing to the greater length of the fragments, the LSP library associated to its prediction method represents the first attempt to trap the consequences of long-range interactions at both sequence and structure levels.

In the present work, three main points were addressed: i) the development of a new strategy for protein local structure prediction using LSPs, ii) the proposal of an index for directly assessing the relevance of the prediction at each sequence site and iii) the evaluation of the quality of prediction using an original description of long fragments that relies on the overlapping properties of prototypes. The novel prediction approach still takes advantage of the features of the LSPs library. In addition, the method benefits from the use of sequence information of proteins homologous to target. Evolutionary information is widely being used for different prediction purposes and is well known to improve the quality of prediction. [33]. Moreover, the learning process is performed using Support Vector Machines (SVMs), a sophisticated classifier which has already been used successfully in various fields related to protein structure prediction [10,34-35]. We take care of preserving a well-balanced prediction rate for each one of the 120 structural classes. The results obtained were compared and discussed with other tested schemas of prediction.

An index that directly estimates the quality of each prediction and the relevance of the structural candidates for each fragment sequences all along a protein sequence was also proposed.

Finally, we analyze the results in the light of a new structural description. Taking into account the most frequent transitions between LSPs in the whole

databank, preferential sub-networks can be delineated. The latter encompass long structural fragments and even super-secondary structures. This representation reminds, at a different scale, the network representation of Protein Blocks we previously designed and that permitted to define Structural Words [36]. Apart from its structural interest for defining rather long structures, this new representation was also used to assess the local structure prediction results.

Methods

Protein structural databank and datasets

A databank of 1041 X-ray structures (available on request) was extracted using PDB-REPRDB [37] such that they share a resolution better than 2 Å, no more than 30% pair wise sequence identity and a Cα RMSD between them higher than 10 Å. Each protein structure was then analysed as overlapping fragments of 11 residues. The assignment of each fragment to a given HPM structural class was based on the minimal Cα RMSD criteria with the representative local structural prototype (named LSP) of the given class. A total of 251,497 fragments were obtained and encoded in terms of LSPs.

Three protein datasets have been extracted from this databank (see Figure 1). *Set 1* (half of the structure databank, *i.e.* 521 proteins, 125,074 fragments), was used as an independent set to deduce the relation between confidence index categories and prediction rate. *Set 2* (the first quarter of the structure databank, *i.e.* 261 proteins, 62,194 fragments, different from *set 1*) was used as a training set for the SVM prediction procedure and the confidence index calibration. *Set 3* was a validation set used to assess the local structure prediction method and the relevance of confidence

index in inferring the prediction quality (the last quarter of the structure databank, *i.e.* 259 proteins, 64,229 fragments).

In addition, for each structural class *s*, a machine learning we named an "expert" was trained to discriminate between fragments associated to the class *s* and those which are not (see Figure 1). Henceforth, for a given class *s*, a "training subset" was derived from *set 2*. Half of this training subset was made of 3D fragments assigned to *s* (positive part) and half was composed of the same number of 3D fragments taken randomly from other clusters and showing a Cα RMSD structural dissimilarity superior to 1.5 Å compared to the LSP of *s* (negative part).

Prediction strategy for long local protein structures

Overview of the prediction method

The prediction method relies on an expert system. For each local structural class *s* represented by its LSP, an expert (LSP-expert) is trained to optimally discriminate between fragment sequences associated to *s* (positive examples) relative to other classes (negative examples).

Each LSP-expert computes a compatibility score of a target sequence window with the class considered. The 120 scores are then ordered and finally, a jury selects the 5 top-scoring classes and proposes their representative LSPs as structural candidates.

Representation of the target amino-acid sequence

Sequence profiles defined by Position-Specific Scoring Matrices (PSSM) were built using PSI-BLAST [38] to gather the evolutionary information. Similarly to our previous studies [13,25], we considered enlarged sequence windows that are 21 residues

in length (marked from -10 to +10 and centred at 0) to take account of the amino acid content in the neighbourhood.

A PSSM was first computed starting with the complete target protein sequence: PSI-BLAST searches were carried out up to four iterations against the non-redundant SWISS-PROT databank [39] with an the e-value threshold of 10^{-4}. The software blastpgp v2.2.13 was used (ftp://ftp.ncbi.nih.gov/blast/executables/release/2.2.13). Then, the PSSM was cut into overlapping matrices corresponding to the sequence fragments composing the whole protein sequence. So, each fragment sequence is represented by a PSSM of dimension 21 x 20 whose elements correspond to the log-likelihood of an amino-acid AA_k substitution at a position y in the template. The values of the PSSM profiles were linearly scaled to the range [-1; +1] with the software LIBSVM [40].

Definition of the expert predictors by support vector machines (SVMs)

In this study, the prediction strategy is based on an expert system defined by SVMs. SVMs are a generalisation of the linear classifiers [41]. We used SVM[light] that provides a fast optimization algorithm for SVMs [42]. The principle of the training stands in two steps. Firstly, the dataset is projected into a different space using a kernel function which defines the similarity between a given pair of objects [43]. We chose a radial basis function kernel (RBF) that was successfully used in recent studies related to protein structure prediction [10,33]. It first implies the selection of a parameter γ :

$$K(x,x') = \exp(-\gamma \|x - x'\|^2) \text{ for } \gamma > 0$$
$$\text{where } x \text{ et } x' \text{ are two examples of the dataset} \tag{1}$$

In addition, SVM method consists in defining the optimal hyperplane that is the farthest from any training example and minimizes the training errors. The

procedure depends on an additional parameter, *i.e.* a parameter C that represents the trade-off between minimizing the training error and maximizing the margins. In addition to C, a parameter λ can also be optimized. It determines the extent by which training errors on positive examples outweigh errors on negative ones and, thus corresponds to an asymmetric regularisation of parameter C.

The SVM experts were trained using the training subsets derived from *set 2* (see Figure 1). Optimal values for γ, C, and λ, were obtained for each structural class *s* through grid-searches and cross-validation. These parameters were optimized two by two, *i.e.* γ *vs.* C and γ *vs.* λ for keeping manageable CPU time calculation. The parameter ranges proposed by Hsu and co-workers [44] were tested and empirically adapted to our optimization procedure. For each structural class *s*, a cross-validation procedure was performed: the training subset was randomly divided in 2 sets containing the same number of positive and negative examples. The first part, representing 2/3 of the data was used to train the SVM with a given parameter pair (γ,λ) or (γ,C); the remaining part was then used as a validation set. For each parameters pair (γ,λ) or (γ,C), this process was repeated 5 times and, the mean of the obtained classification rates was calculated. Finally, for a given class *s*, the selected parameters pair (γ,λ) or (γ,C) maximized the mean classification rate. Once an optimal pair (γ,λ) or (γ,C) was selected for each structural class, the definitive training of the SVM experts was carried out on all subsets.

Only results concerning the parameters pair (γ,λ) leading to the best prediction performance will be presented. Results associated to the pair (γ,C) are presented in supplementary data I.

For the prediction of a given fragment sequence *f*, the LSP-SVM expert trained for class *s* computes a decision value v_s that is proportional to the geometric distance of *f* from the maximum margin separating hyper plane fitted for *s*.

Jury and decision rule for prediction

The compatibility of a fragment sequence *f* with a given LSP is measured by the decision value or score given by each LSP-expert. The 120 scores thus obtained are ranked in descending order and a jury selects the best structural prototypes as candidates for the local 3D structure. This strategy of ranking classifier outputs was used in our previous study which relied on experts defined by logistic functions [25]. In the same way, Kuang and co-workers used the same strategy with SVM while predicting 9-residue local structures, *i.e.* they took the maximum margin given by SVM classifiers specialized for each class [45]. Finally a list of five candidates was proposed reducing considerably the number of possible structures for a given target sequence fragment.

Evaluation of the predicted local structures

The prediction strategies were assessed on *set 3*. Two types of evaluation schemes were tested. In the first one, a prediction for a target sequence window was defined as successful when the assigned LSP *s* was found among the predicted candidates. This evaluation was named Q_{120}. In the second scheme, a geometrical assessment was done, *i.e.* the prediction was considered to be successful if the Cα RMSD between one of the LSP candidates in the list and the true local structure was lower than a given threshold. Three thresholds were considered, *i.e.* 1.5, 2.0 and 2.5 Å. In order to simplify the reading of this study, the results with the two first approximations are summarized in supplementary data II. We will here mainly

concentrate on the latter approximation. As a matter of fact, we had shown in our previous study that this threshold was quite stringent. We studied the distribution of 100.000 Cα RMSD values between 11-residue fragments randomly chosen from the databank. The Cα RMSD was calculated only if these fragments encoded into series of 7 Protein Blocks (PBs) differed by more than 5 PBs. The mean of this distribution was 4.5 Å (σ = 1.1 Å). The probability for a random match with Cα RMSD < 2.5 Å was 10^{-2} [25]. For reference purpose, it is worth noting that this threshold is comparable to the value chosen by Yang and Wang (2.4 Å threshold) to evaluate the proportion of 9-residue fragments that was correctly predicted by their methodology [6].

Analysis of the prediction according to LSPs categories

The 120 LSPs are necessary to give a correct approximation of known local structures. Nevertheless, for analysis purposes, regrouping of LSPs in different coarser categories of LSPs can be done. For instance, secondary-structure-like categories can be defined [25]: Helical (H), Extended core (E), Extended edges (Ext) or Connecting (C) structures that respectively encompass, 16 LSPs, 13, 40 and 51 LSPs respectively.

We defined a new description that is based on the overlapping property of LSPs and consists in grouping LSPs according to transition probabilities from one LSP to another in global structures. A network characterizing the chaining of all LSPs was defined accordingly (see Figure 2). The transition probability $P_{x \to y}$ from a LSP x to another LSP y was calculated as following:

$$P_{x \to y} = \frac{N_{x \to y}}{N_x} \quad (2)$$

where $N_{x \to y}$ was the number of observed transitions from x to y and N_x the number of transitions from x.

This global network was then iteratively divided into sub-networks. A sub-network was definitely defined if composed of less than 20 LSPs and if its inner transition probabilities are higher than a probability p. p had an initial value of 0.20 and was then increased by steps of 0.02. The process was stopped when no sub-network of more than 20 LSPs was left. At this point p equaled 0.40. Hence, the last two sub-networks created, 6 and 7, encompassed very frequent transitions higher than 0.40. Considering all the defined sub-networks, 16.8 % of the transition probabilities considered (higher than 0.20) were higher than 0.50 and 4 % were higher than 0.70. The maximum probability value was 0.86 between LSPs 101 and 102 in network 5.

Like a pruning process, this iterative procedure progressively yielded to 7 sub-networks sufficiently populated and exhibiting significant transition probabilities. 52 LSPs remained as components of small sub-networks composed of less than 5 LSPs and even as isolated LSPs. They were further clustered into three groups according to their size and their inner and outward transition probabilities. Thus, ten substructures composed of different LSPs with preferential transitions finally emerged.

Evaluation of the efficiency of the method compared to other predictions.

We compare the efficiency of our prediction strategy coupling SVM with PSSM (and thus named *SVM_PSSM*) with four kinds of predictions:

i) A random prediction: five candidates for each fragment sequence in *set 3* are selected randomly [25].

ii) A "Naïve prediction": for each sequence of 21-residue long of *set 3*, the 5 most similar sequences in *set 1* were selected. Similarity scores were computed using BLOSUM62 matrix. The corresponding fragment structures were superimposed to that of the target sequence, only considering the central 11-residue fragment

(positions 6 to 16 of the sequence fragment). This strategy is similar to the very first step of prediction associated with the use of fragment libraries like in Rosetta [46].

     iii) Different prediction schemas based on LSPs:

- *SVM_seq*: SVM learning without the use of PSSMs, *i.e.* for a single sequence.

- *LR_seq*: a logistic regression with single sequence, *i.e.* the prediction methodology developed in [25].

- *LR_PSSM*: a logistic regression with PSSM matrices.

     iv) Other cutting-edge methods described in the Discussion section.

All these predictions were assessed with the Q_{120} and the geometrical criterion

Definition of a confidence index

The confidence index (CI) was defined by making use of SVM expert decisions. A local structure prediction was performed on *set 2* using the *SVM_PSSM* strategy. For each target fragment sequence, the 120 decisions of experts were retrieved and associated to correct or incorrect prediction classes according to the geometric criterion. This data was used for training a new SVM (namely SVM_{CI}) to learn which expert decision ensembles lead to correct or incorrect prediction. The procedure was quite similar to what was done for local structure prediction and was carried out with the optimisation of γ and C parameters (for more details see supplementary data III). The Confidence Index CI is defined as the decision value of SVM_{CI} model. This model was assessed on the prediction results obtained on two independent sets, *set 1* and *set 3*. The results obtained with *set 1* were used for deciphering the optimal relationship between the CI categories and the prediction rate. This relationship was checked on *set 3*.

Results

The grouping of LSPs according to their preferential transitions in protein structures is first described. Then, the new local structure prediction strategy using support vector machines coupled with PSSM is assessed (*SVM_PSSM*). The efficiency of the prediction is discussed for the different sub-networks that group LSPs according to their most probable transitions. Finally, a confidence index that gives a direct indication of prediction quality is assessed.

Analysis of preferential structural transitions between LSPs

An analysis of the LSP structural transitions in protein structures, led to the LSP network presented in Figure 2. Using an iterative pruning procedure of this network, we defined ten LSP groups characterizing preferential transitions of local structures in global protein structures. The composition and the transitions of the most significant categories, or sub-networks, are described in Figure 3 and supplementary data IV. Each sub-network comprises a limited number n of LSPs (n ranges from 5 to 16). Owing to possible branching and transitions probabilities, it characterizes super-local structures composed of at most n LSPs.

Sub-networks 1 and 2 can characterize the fold of 15-residue sequences. Sub-network 1 is composed of 5 LSPs and includes a helical LSP followed by four specific different connection LSPs defining a β-turn. Thus, this sub-network can be defined as a super-secondary structure, *α-C^{cap}-β-turn*. It encompasses LSPs characterizing 2.74 % of fragments. The LSPs 44, 45 and 46 are respectively assigned to 0.67, 0.74 and 0.70 % of fragments. They are thus rather frequent LSPs given the great number of classes. As a matter of fact, each one of the 120 is assigned to 0.83 % of fragments in average ($\sigma = 0.56$) and the distribution has a median of 0.71 %. LSPs 47 and 48 are

less frequent, *i.e.* they represent 0.33 and 0.30 % of fragments. This relative rarity strengthens the significance of the succession of LSPs 46 and 47 observed in protein structures ($P_{46 \to 47}$ =27.53 %). Sub-network 2 is composed of 12 LSPs encompassing 6.57 % of fragments. As an example, the super-secondary structure shown in Figure 3, corresponds to a pathway involving transitions from LSPs 51 to 60. It is composed of four LSPs associated to extended edges leading to an extended core LSP. All super-secondary structures defined by this sub-network begin with a β-turn followed by a small β-strand, and then a second one, longer than the first, and nearly orthogonal to this latter. We define this super secondary structure as a *turn-ββ-corner*. The core of this sub-network also encompasses frequent LSPs like 57, 58, 59 or 60 that characterize the more frequent type of corner, with occurrence frequencies of 0.91, 0.78, 0.63 and 0.79 respectively.

Sub-networks 3 and 4 comprise 6 LSPs. Sub-network 3 describes the chaining of 4 four successive LSPs while sub-network 4 characterizes the transition of at the most 5 successive LSPs, *i.e.* sequences of up to 14 and 15-residue long. Sub-network 3 proposes alternative ways to enter a β-strand after a change in the direction of the backbone, always involving an α-turn. We thus named this sub-network *α-turn-β-strand*. LSPs of this sub-network represent 4.05 % of fragments. LSPs 106 and 96, characterizing this pathway for entering a β-strand, are among the 35 % most frequent ones with frequencies of 0.89 and 0.91 % respectively. On the contrary, sub-network 4 proposes a different situation where β-strand endings lead to a direction change in the backbone due to a β-turn. In general, this direction change leads to another extended structure. Nevertheless, in each case, this pathway includes bulge-like irregularities. Sub-network 4 was identified as an *Irregular β-hairpin-turn*. In this case also, LSPs 118, 84, 85, present in the core of this sub-network, are among the

most frequent ones as they represent respectively 0.73, 0.82 and 1.81 % of fragments respectively. The total percentage of fragments characterized by the LSPs of this sub-network is 4.24 %.

Sub-networks 5 and 6 are the longest (up to 20 residues) and the most complex super-structures. Sub-network 5 is composed of 16 LSPs encompassing 19.95 % of fragments and presents two main pathways for entering extended structures through an α-turn including a β-turn. The first pathway starts with helical structures while the second one begins with extended structures. This sub-network can thus be defined as an *α/β-hairpin*. The structural classes composing this sub-network are highly populated, *i.e.*, 10 characterize more than 1 % of fragments each. Sub-network 6 comprises 14 helical LSPs and connecting LSPs characterizing 12.11 % of fragments. It presents three main pathways to enter a helical structure. The first pathway connects two helical structures with a loop and the two others are constituted of long series of connection structures. In general, the latter follow extended structures. Sub-network 6 can be defined as a *α/β-loop-α* super structure. Like sub-network 5, sub-network 6 encompasses very frequent LSPs, 6 of them represent more than 1 % of fragment each.

Finally, Sub-network 7 comprising 9 LSPs (up to 16 successive residues), presents alternative pathways to enter a helical structure, either only through connecting LSPs or through both extended edges and connecting LSPs. Owing to the orthogonal orientation of the connection or extended edge structures with reference to the helix, this sub-network was identified as a *βα-corner*. 7.14 % of fragments are characterized by the LSPs of this sub-network. LSPs 64, 65, 66, 67 and 68 represent 0.86, 1.02, 1.19, 0.79 and 0.97 % of fragments respectively.

Together, these 7 sub-networks characterize 24.38 % of transitions observed in protein structures.

According to their transition properties, the remaining LSPs were clustered in three categories referred to as groups 8, 9 and 10. Group 8 comprised 11 LSPs presented high inwards or outwards transition probabilities with more than one other LSP whatever its group. Group 9 was composed of 22 LSPs clustered in 9 small sub-networks (2 or 3 LSPs) with very high inner transition probabilities. Finally, the group 10 clustered 19 single LSPs associated to low transition probabilities with other LSPs, for two reasons: either they were quite rare and do not have strong transition preferences, or, they were very frequent and were a sort of hub toward many possibilities (see Methods section and supplementary data IV for details).

SVMs using PSSMs achieve successful prediction of long local structures

Global evaluation of predicted lists of local structure candidates

It must be recalled that the prediction of a fragment sequence is considered as correct if the list of five candidates contains the true LSP (Q_{120}) or if the Cα RMSD between at least one LSP in the list and the true structural fragment is less than 2.5 Å (geometric criterion). This value has to be compared with the intrinsic variability of each LSP that ranges from 0.28 to 2.44 Å [25].

Coupling SVM experts with the evolutionary information represented in the form of PSSMs makes the proportion of true-positive predicted LSP, Q_{120}, reach 38.8% (see Table I and supplementary data II). This result is quite significant given the large number of classes and the length of the predicted sequence fragments. It corresponds to a significant gain of 34.6 % over a random prediction and 18.3% for the naïve prediction, using a list of five candidates in both cases. Considering the geometric

criterion, the prediction rate increases to 63.1%. The gains compared to a random or a naïve prediction, are noteworthy, 38.0 % and 15.2% respectively. Table II (third column, first row) shows that for the fragments predicted as correct (*i.e.* 63.1% of samples), at least one candidate among the five proposed has an average RMSD of only 1.45 Å from the true local structure. For these sequence fragments, the mean approximation over the 5 candidates has an average value of 2.54 Å. By considering all samples, a satisfying 2.09 Å minimal approximation is still available in average among the 5 candidates and the mean approximation for all candidates is 3.03 Å.

It may be noted that the prediction procedure is not biased towards the most frequent and the least heterogeneous classes. Indeed, the prediction of each of the 120 structural classes is largely better than random. The mean true positive rate per class reaches 33.7 %, which corresponds to an average gain of 30.4 % over random. Considering the geometric criterion, the *SVM_PSSM* achieves a 58.1 % mean prediction rate per class, *i.e.* an average gain of 39.2 % over the random prediction. These gains are well balanced over all structural classes. The weakest gain was observed for the LSP 113, which still scores 13.3 points over random. This class presents the largest structural variability, *i.e.* a mean 2.44 Å Cα RMSD between the representative LSP and the fragments of this cluster [25].

Assessing prediction through a secondary structure vision

While considering the four secondary-structure-like LSP categories as done previously [25], the Q_{120} ranges from 28.4 % for Ext LSPs to 50.9 % for H LSPs corresponding to gains of 24.2 % and 46.5 % respectively over a random prediction and of 14 % and 23.5 % over a naïve prediction (see Table I).

The prediction rates based on the geometrical criterion, ranges from 49.5 % to 84.6 % (see Table I). Well-balanced gains of 43, 39.3, 36.5 and 33.9 % were obtained

for H, E, C and Ext local structures, respectively, over random. When compared with the naïve prediction, the gains remain very high, *i.e.* 23.9, 16.2 and 23.2 for E, C and Ext structures respectively, with one exception associated to helical structures, where the gain is only 3.5 %. This less significant performance may be attributed to i) the high sequence specificities of helical structures that make easier the prediction even by a simple sequence similarity search and ii) the choice of the geometric threshold 2.5 Å, that is not stringent enough to accurately assess the prediction of helical structures. Considering a 1.5 Å threshold (see supplementary data II), our method still provides a 67.8 % prediction rate for Helical structures. In this case, a higher 6.5 points gain over a naïve prediction is observed.

While examining the correct predictions defined by the geometric criterion for all predicted fragments, the best available approximations reach in average 0.83, 1.78, 1.75 and 1.93 Å for H, E, C and Ext fragments respectively, while the structural approximation had in average values of 1.21, 2.17, 2.48 and 2.41 Å respectively. These results are quite significant if one considers the structural variability existing inside the H, E, C and Ext LSP categories, *i.e.* respectively 1.29 Å ($\sigma = 0.98$), 2.14 Å ($\sigma = 0.48$), 3.34 Å ($\sigma = 0.61$) and by distinguishing extended structure entrances and exits, 2.87 ($\sigma = 0.68$) and 2.58 Å ($\sigma = 0.53$) [47].

Assessing prediction with the view-point of super-secondary structures

The four categories previously defined help for analysing the LSPs in comparison with well-known polypeptide descriptions. However, these categories are rather crude and fail to capture specific structural features particularly in coil category. To go further, we examined the prediction results obtained for the seven different frequent super-secondary structure types we described above, and, the three

remaining LSP groups characterizing their transition properties (see Result section: Analysis of preferential structural transitions between LSPs).

With such a description, prediction rates based on geometric criterion range from 37% (sub-network 4/*Irregular β-hairpin-turn*) to 64 % (sub-network 1/α-C^{cap}-β-*turn* and 2/*turn-ββ-corner*) for the first seven categories (see Table III) and the rate even reaches 73 % for the three last groups. The gains are quite significant ranging from 25.5 % (sub-network 4/*Irregular β-hairpin-turn*) to 42.8 % (sub-network 5/α/β-*hairpin*) compared to a random prediction, and from 8.8 % (group 9) to 29.1 % (sub-network 2/*turn-ββ-corner*) when compared to a naïve prediction. Interestingly, the prediction rates are not related to the number of LSPs in sub-networks (data not shown).

The accuracy of these predictions is well balanced over the different sub-networks and groups (see Table II). For instance, considering all predicted fragments and the first 7 categories, the average geometric accuracy over the 5 LSP candidates ranges from 3.07 Å to 3.63 Å. And the optimal structural approximation (minimal RMSD over the 5 candidates), ranges from average values of 2.17 to 2.80 Å. The groups 8 and 9 are associated with a better accuracy with a mean RMSD of 2.93 and 2.43 Å respectively. The average minimal RMSD drops in average from 1.94 to 1.56 Å. This result is related to the high helical content of these two groups.

More details on predictions and different representative examples of structural candidates are given in Supplementary data V, VI, and VII. The results for three different proteins belonging to different SCOP classes [48] (α/β, all-α, all-β) are described. The prediction rate for these three cases ranges from 71.5 to 74.5%.

Assessment of a confidence index

All the results described above emphasize the importance of identifying the sequence regions with high predictive power. Thus an index aiming at estimating the quality of the prediction is strongly required. This question was previously addressed for accompanying different prediction methods: for instance PsiPred [49] or HYPROSP II [50] benefit from indices that estimate the probability to obtain the value by chance. Generally, they are related to the information content of the target object. We have previously proposed similar indexes, like *Neq* or confidence index, that effectively quantify the reliability of the prediction rate and permit us to focus on specific regions of the sequence [13,51,52].

In the present study, we propose a Confidence Index CI based on the SVM_{CI} decision values. This index gives an indication of the extent to which a predicted local structure corresponds to a correct prediction.

The SVM_{CI} decision values obtained on *set 1* were divided into 30 equivalent parts. The first 8 and the last 5 categories representing only 1.15 and 0.95% of predictions respectively were gathered, yielding a final number of 19 categories of SVM_{CI} decision values.

The distribution of the prediction rates as a function of these 19 categories for *sets 1* and *3* are given in Figure 4 A. The categories are ordered from low to high confidence. The curve representing the average prediction rates as a function of the categories of SVM_{CI} decision values is clearly sigmoid. Whatever be the set considered, the average local structure prediction rate is high (61.3 % and 63.1 % respectively), and the distribution of rates as a function of SVM_{CI} decision values is almost identical. SVM_{CI} decision values inferior to -0.58 (categories 1 to 5) are associated to poor quality predictions, with prediction rates ranging between 20 and

30 %. Inversely, SVM_{CI} decision values greater than 1.10 (categories 15 to 19) are associated with high quality predictions, the rates ranging from 83.9 to 96.8 %.

Between these two extremes, the prediction rate increases rapidly according to SVM_{CI} decision values categories.

Hence, the SVM_{CI} decision values are related to the reliability of the prediction: the larger they are, the more reliable is the prediction. Consequently, SVM_{CI} decision values can be defined as CI values.

The clear relationship existing between CI categories and the prediction rates can be estimated quantitatively using a linear regression after logarithmic transformation of results from *set 1*. This relationship permits the definition of a theoretical prediction rate (TPR) as a function of the CI category (Equation 3)

$$\text{TPR} = \frac{4}{5} \left(\frac{1}{1 + e^{-0.36(CI \text{ category} - 11.53)}} + \frac{1}{4} \right) \text{ (3)}$$

This relation can explain 96% of *set 1* data (determination coefficient equal to 0.96) and residue distribution followed a normal law (Lilliefors test; R software [53]). The model found to be correct for *set 1,* was assessed on the validation *set 3*. The mean theoretical prediction rate (TPR) calculated on all predictions of *set 3* equals 61.80%, a value very close to the observed rate, *i.e.* 63.13%. Figure 4 B shows the correspondence between the estimated mean TPR for each protein of *set 3* and their observed prediction rate. A significant linear correlation of 0.77 was observed (*p-value* < 2.2 10^{-16}). We also observe a significant correlation (0.86) between the estimated mean TPR per structural class and the observed prediction rate. Therefore, the confidence index defined here, along with the relation described in equation (3), is proved to be very relevant in the estimation of the quality of a prediction.

Discussion

Our results show that the *SVM_PSSM* local structure prediction strategy can give results whose quality and accuracy are far better than a random prediction or a prediction based on sequence similarity searches. By reducing the structure space to 5 relevant candidates for 11-residue long local structures, our method is efficient in simplifying the combinatory problem associated with the generation of structural models for longer protein sequences. Prediction rates and approximation accuracy when compared to simple prediction schemas like random or naïve predictions are noteworthy given the large number of classes, the length of the predicted fragments and the diversity of the structural classes. Moreover, the confidence index is relevant in the identification of regions corresponding to the best predictions.

In the following section, we will discuss and compare the relevance of our strategy to more sophisticated schemas of prediction and state-of-art methods.

Comparing SVM experts with LR experts

The LR method, previously developed in [25], was based on a logistic regression with the use of single sequence information and supervised with a system of experts. These experts estimated the probability for a given sequence fragment to belong to a given LSP. All the candidates with a probability larger than a given threshold were considered. Hence, for each sequence fragment a list of structural candidates belonging to different structural classes was proposed. This list could contain a variable number of candidates (from 1 to 5 and 4.2 on average among 120 classes). The prediction rate of the method based on single sequence information reached 51.2%. Compared to a random prediction, a significant gain of 29.3% was observed. Moreover, this method was also shown to be very competitive with cutting-edge local

structure prediction methods [25], *i.e.* with Yang & Wang prediction method [6] and with HMMSTR predictions based on the I-sites library [54]. We also checked that the prediction was not biased towards a reduced set of structural classes, but well balanced between the whole set of LSPs, confirming the relevance of all structural classes and the sequence information they contained.

In the present *SVM_PSSM* method, we have chosen to fix the number of candidates for each sequence (5 candidates in the list). Accordingly, for comparison purpose we re-assessed the original methodology defined by logistic regression (LR) with a fixed number of 5 candidates. As in our previous study [25], the approach, named *LR_seq* in the following discussion, was carried out using single sequence.

We also evaluated the putative gain yielded by evolutionary information for each of the approaches, LR or SVM (*LR_PSSM*, *SVM_seq* respectively). For comparison purpose, all these strategies were developed using the non-redundant protein databank used in our previous study [25]. In the same way, the datasets defined earlier were used.

SVMs and LR show similar performances with single sequence

When SVM approach was performed with single sequence information, the average Q_{120} equalled 30.61% (see Table I and supplementary data II). This is almost equivalent to the value obtained with the *LR_seq* strategy. With the geometric criterion, the *SVM_seq* prediction rate reached 55.54%. This result was again comparable to the result obtained with *LR_seq*. These similarities were observed for the four secondary-structure-like categories of LSPs. Therefore, LR and SVM experts behave in a similar way and have similar ability to trap relevant structural information with single sequence encoding. This capability is quite interesting in case of orphan

protein sequences since in such cases, *LR_seq* remains an interesting method because of to its low computational cost.

Improvement of prediction coupling SVMs with PSSMs

The introduction of evolutionary information in the SVM strategy leads to a significant improvement of the prediction performance when compared to both approaches performed with single sequence information. This result is achieved as well by considering the true positive rate (Q_{120}) as the geometric criterion (+7.3 and 7.7% respectively). Detailed analyses show that the prediction is improved for all the four categories of LSPs based on secondary structures. The gain ranges from 4.5% for the connecting structures to 16% for extended structures (see Table I and supplementary data II).

Overall, the significant improvements in the mean true positive rate were observed for the majority of the structural classes, independently of their frequency, *i.e.* the Q_{120} of 74.17% of structural classes was increased compared to the *LR_seq* *s*trategy. Clearly, there is no correlation between the improvement in the prediction rate per class and the corresponding class frequencies (correlation coefficient *r* equals to 0.21). For instance, among the 35 weak populated classes (with less than 5% of the local structures), two-thirds exhibit an improved Q_{120} prediction rate. Half of them gain by least 10 % increase. Considering the geometrical criterion, the prediction rate for 88.33% of LSP classes was increased compared to *LR_seq* (see Figure 5). Only, 12 C and Ext classes show limited losses ranging from 1.6 to 6.8 points. These classes represent only 7.4 % of fragments.

Interestingly, when the ten transition categories were considered, significant gains were also obtained (on average 7.6%). The largest value was observed for sub-

network 2 (16%) (see Table III) while the smallest gains of 4.2 and 4.9 % were obtained for sub-networks 4 and 5 respectively.

Finally and more importantly, the prediction of 85.33% of the proteins was improved.

Superiority of SVM with PSSM

For comparison purpose, we also considered the influence of evolutionary information (PSSM) with LR strategy (*LR_pssm*). In this case, a drastic loss of performance was observed. A first explanation comes from the size of the databank which, for certain classes, affected the correct estimation of the coefficients of the logistic functions. Nevertheless, the main explanation for the reduced performance lies on the expert behaviour. Indeed, a detailed analysis showed that the distribution of probabilities, brought by the experts, which measured the compatibility between a target sequence and a given LSP, were not uniformly spread but mainly confined towards extreme values, *i.e.* 0 and 1. Since the selection and the ranking of candidates depend on these values, the introduction of PSSM clearly decreased the discriminative power of experts because a very large set of examples is required to appropriately train the experts (see supplementary data II).

Considering these results, the relevance of coupling SVM experts to PSSMs can be emphasized.

Comparison to cutting-edge other local structure prediction methods

Comparing local structure prediction methods is not trivial due to the large diversity of the existing strategies. Actually, different representations of the local structure space, different protocols of prediction and assessment have been defined.

We compared our results with three categories of local structure analysis and associated prediction methods.

Backbone torsion angle prediction

Bystroff and co-workers defined a backbone torsion angle prediction method (HMMSTR) relying on an alphabet of 11 conformational states for protein backbone[54]. These states characterized 11 torsion angle regions mapped onto the Φ-Ψ Ramachandran plot (see Figure 5 of Bystroff et al.[54]). In the same way, Yang and Wang [6] and Kuang et al. [55] defined local structure prediction methods specialized in the prediction of the backbone conformation of the central amino-acid of nine-residue sequence fragments. Four states, A, B, G, E, were defined. For comparison purpose, they grouped the Bystroff's eleven main states into the four states they defined [55]. Similarly, in our previous study [25], we extracted the ϕ,ψ angles defining each local structure candidate we proposed and allocated to the corresponding A,B,G or E states. We showed that *LR_seq* method with 4.2 candidates in average, evaluated in the context of backbone torsion angle prediction, yielded a prediction rate ranging from 64 to 76%. This result, obtained for a single sequence, was comparable to the 75% accuracy obtained by Yang and Wang and to the 77% obtained by Kuang et al. while their methods profited from the use of information from the results of PSI-PRED [55]. This result was also comparable to the 74% prediction accuracy obtained for HMMSTR. Thus, insofar as our previous strategy, *LR_seq*, was as efficient as these cutting-edge local structure prediction methods, and as the present *SVM_PSSM* strategy gives a larger prediction accuracy compared to *LR_seq* (~12%), so, the methodology we present here, can be considered as a significant contribution to the field of local structure prediction.

Prediction based on Structural Alphabet

Recently, Sander et al. defined a new structural alphabet associated to an efficient prediction strategy [10]. They defined 27 canonical local structures of 7-residue long by a discretization method taking into account both sequence and structural information. Several prediction strategies were tested; the most successful one used a combination of Random Forest (RF) classifiers with a representation of the sequence as profiles of amino-acid properties. Comparison with *SVM_PSSM* is arduous because of very different numbers of classes and lengths of local structures considered. Nevertheless, we tentatively tried to get closer to the Sander et al. prediction conditions and we mapped our 120 LSPs in the 27 classes they defined according to their structural proximities. Our prediction rate was then evaluated accordingly. Table IV reports the prediction rate Q_{27} obtained by Sander and co-workers [10] for 7-residue long local structures. The Q_{27} results ranged from 34 to 64% for 1 to 5 candidates (see [10], Figure 10). These results are comparable with the prediction rates that we obtained with our *SVM_PSSM* strategy, which range from 32 to 61 % for 1 to 5 candidates. These rates are all the more satisfying in the sense that we predict 11-residue long fragments, *i.e.* 4 residues longer than those predicted by Sander and co-workers.

The performance of our prediction method can also be evaluated through Receiver Operating Characteristic (ROC) curves for each structural class. These curves represent the repartition of true positive rate versus false positive rate [56]. They are class distribution-independent, and can be calculated for classifiers with continuous output.

For each fragment, we thus computed the difference between the best SVM decision value and the decision value of the expert specialized for the class considered. The classification performance of each class was then represented by the

area under the ROC curve (AUC) that ranges between 0.5 (random prediction) and 1 (perfect prediction). The AUCs that we obtained range from 0.71 to 0.92, and the average value is 0.82 (see Figure 6). The minimal and maximal values correspond to respectively the classes 55 (connection structure) and 33 (extended edge structure). Thus, even while considering the AUC values, our prediction method performs as good as or slightly better than Sander et al. approach, for which the AUCs range from 0.68 to 0.88.

An interesting application of LSP prediction: the long loop prediction

Loop prediction is a major step and a difficult task (even the most difficult one) confronted in the context of homology modelling methods. Loop prediction is frequently performed on a protein structural framework where the secondary structures are already delineated. Thus, *per se*, the approach benefits from the complete or partial knowledge, of the rest of the protein structure. Interesting protocols have recently been proposed that generally comprise two main strongly imbricated tasks, even concomitant: the sampling of the conformational space of the loop fragment and the scoring of the corresponding sampled conformations (see below). The sampling may be carried out using a set of structures extracted from a dedicated database or using *ab initio* methods. The latter seems to be more effective for long loops [57]. Most recent approaches perform rather well and accurately for loops comprising less than 10 residues. Above this limit, sampling becomes the stumbling block of the procedure because it requires tremendous cpu-time to be efficient and to yield accurate results [58]. Thus prediction of long loops remains a challenging field.

Consequently, the question we address here concerns the efficiency of the LSP description coupled with the use of *SVM_PSSM* to provide rapid predictions of conformations for a loop region. Accordingly, we compare the accuracy obtained with

ab initio prediction methods dedicated to loop prediction to the accuracy that we observe for the connection structures (assigned to C LSPs) using our generic local structure prediction method. We focused on five loop prediction methods that were recently compared: *LoopBuilder* [59], the Modeller loop prediction method [60], LOOPY [57], RAPPER [61] and PLOP [58]. The comparison we propose is rather difficult for two main reasons: the definition of the loops region and the evaluation of the prediction accuracy. Indeed the "C" LSPs connection structures can correspond to loops shorter than 11 residues or, inversely can also be part of longer loops. In addition, the assessment is done for all protein loops, *i.e.* 24 856 fragments while most loop prediction methods carefully focus on specific loop datasets. For instance, *LoopBuider* discards structures crystallized at a non standard pH and loops found to be involved in interaction with a ligand. The datasets may thus be very small *(*for example, *LoopBuilder* is assessed on 54 loops which are 11 residues long*)*. So, our results may be affected by artefacts due to ligands or experimental conditions.

Briefly, we recall the main and the common features of the different approaches tested. All of them are all-atom prediction methods and rely on a two-step algorithm: first the sterically allowed backbone conformations are sampled and then they are scored and ranked based on the most favourable energy. Importantly, the selection is done after taking into account the conformation and the side chain coordinates of the rest of the protein.

The Modeller loop prediction method samples all-atom loops in a Cartesian space and optimizes the energy function by coupling conjugate gradients and molecular dynamics with simulated annealing. Before selecting the conformation associated with the lowest energy value, 50 to 500 independent optimizations are done [60]. LOOPY relies on a colony energy filter that favours conformations having many

neighbours in configurational space. For a single loop prediction, 2000 random backbone conformations are generated and then filtered [57]. In the same way, for a target, RAPPER generates 1000 backbone conformations using fine-grained Φ/Ψ tables. The best candidate is then selected by coupling the AMBER force field and the Generalized Born/Surface Area (GBSA) solvatation model [61]. PLOP performs a very extensive multistage Φ/Ψ conformational sampling and scores the generated loop conformations using an OPLS all-atom force field combined with a generalized Born solvation and a new hydrophobic terms [58]. Lastly, *LoopBuilder* relies on LOOPY for sampling and select best conformations using the DFIRE potential before minimizing [62].

Our purpose is not to compete with these elaborated approaches that make use of force fields, minimization techniques and energy functions. They were assessed mostly on small and carefully selected samples while in our case, the assessment was done on a large and diversified set of examples, so with a wider variability. Nevertheless, our point is to show that our approach can provide interesting structural start points for a deeper analysis or a more elaborated algorithm.

For assessing the accuracy, we choose to compute the local $C\alpha$ RMSD between the predicted and the real local structure. Classically, the criteria used in loop prediction methods is a "global RMSD" on the loop main-chain heavy atoms after superposition of the main-chain atoms in the stem residues on each side of the loop[60]. Fiser and al. gave an idea of the relation between the two measures, *i.e.* the global RMSD is about 1.5 times the local RMSD for the heavy main-chain atoms for a 8-residue loop, and, the local $C\alpha$ RMSD is about the same as the local RMSD computed for all the heavy main-chain atoms[60]. For a comparison purpose and to give an idea of the scale, we will use this factor in the following discussion.

For 11-residue loops, after sampling and selection of the best structural candidate, the method implemented in Modeller achieves an average global RMSD value of 5.5 Å, LOOPY 3.52 Å, RAPPER 4.94 Å, PLOP 1.00 Å and *LoopBuilder* 2.50 Å. Over the 5 candidates per target fragment that are predicted by our method, we obtained an average local Cα RMSD of 3.41 Å (see Table III, row untitled "C secondary structure categories"), which corresponds to a global RMSD near 5.12 Å if a scaling factor of 1.5 factor is considered. Among these 5 candidates, the best one succeeded to have an average Cα RMSD prediction accuracy of 2.48 Å that can be associated to a 3.72 Å global RMSD accuracy using a factor of 1.5. Moreover, if we now consider fragments correctly predicted according to our geometrical criteria (49.5% of fragments), the mean local RMSD accuracy over the 5 candidates was 2.93 Å (4.40 Å with a 1.5 factor) and with the best candidates an average accuracy of 1.75 Å was obtained (2.63 Å with a 1.5 factor). This last result is all the more interesting that the confidence index we have developed helps to give a direct assessment of prediction and thus indicates fragments that are well predicted.

These results are comparable with Modeller, LOOPY and RAPPER accuracies. *LoopBuilder* and PLOP perform better. Nevertheless, the computational time for these prediction methods is very high. The average CPU time for one 11-residue loop target is about 12 days for PLOP on a single processor [58]. *LoopBuilder*, for which computational efficiency was a challenge, still perform in few hours [59]. Thus, our results are all the more interesting considering the fact that the prediction is done instantaneously for a target sequence. In addition, it has to be noted that we do not introduce any constraint on the extremities of the loop regions or add any information on the rest of the protein and side-chain position. Moreover, there is no any energetic criterion that filters the results. So, our prediction method is so

competitive with some of the cutting-edge loop prediction methods. More importantly, it would be a noteworthy help for proposing few relevant and accurate candidates that could enrich this field of research by providing filter or by directing the sampling of the most performing strategies.

Demanding parameterization of SVM experts

A parameterisation of SVM experts is required before construction of models. We checked the influence of two pairs of parameters and selected the one that provided the largest prediction rates. We took care to avoid biased prediction that favour of the most populated classes. The results obtained with the pair of parameters (γ,C) that showed a lower performance, are discussed in supplementary data I.

In addition, the choice of the kernel is a matter of discussion. Hsu et al. suggested that RBF kernel was a reasonable choice compared to other kernels because (i) linear kernel is a special case of RBF, (ii) sigmoid kernel behaves like RBF for certain parameters and is not valid under some others, and (iii) polynomial kernel requires the optimization of more parameters and could have more numerical difficulties [44].

Further improvement could also be achieved by considering the fact that sequences and profiles are structured data, *i.e.* information organisation and order do matter. Few kernels specially designed for profiles were already defined [45, 63, 64] and could be used for defining various relevant feature spaces.

Different nature of structural prediction using *LR_seq* or *SVM_PSSM*

The comparison between the two strategies was not limited to the prediction rate. We also examined the overlap between lists of candidates yielded by each approach. Considering *LR_seq* and *SVM_seq* lists with very similar success rate,

78.4% of correct candidates were found in common. This large but incomplete covering rate is probably due to the difference in the mapping of input data by LR and SVM classifiers. The difference in lists covering increases when *PSSMs* are included in the SVM strategy. Indeed, only 51.2% and 50.4% of the correct structural candidates were found in common between the *SVM_PSSM* lists and the *LR_seq* and *SVM_seq* lists. Thus, since correct predictions found by *LR_seq* were not found by *SVM_PSSM* and reciprocally, further improvement is still possible.

Conclusions

In this work, we have focused on the assessment of a new method we developed for predicting long local protein structures. We coupled efficient support vector machines classifiers with an enriched representation of the target segment sequences by using evolutionary profiles obtained from PSI-BLAST. Five structural candidates are predicted per sequence window. This combination yields a prediction rate of 63.1% for 120 classes of 11-residue local structures considering a geometrical approximation better than 2.5 Å as a correct prediction. This result corresponds to a very significant gain over other approaches aiming at predicting long structural prototypes. It is worth noting that this improvement encompassed all LSPs categories. Indeed, the average prediction rate of 88.3 % of structural classes is improved when compared to a previous strategy proposing lists of five structural candidates.

We also focus on the overlapping properties of LSPs and propose transition sub-networks and groups characterizing their implication in frequent super-structures and their properties in a global structural network. This original description is an interesting way of assessing prediction for fragment structures longer than 11

residues. For instance, a significant 60.9 % prediction rate is obtained for the sub-network 5 which can characterize sequences up to 20 residues in length.

Moreover, we address the question of the structural "predictability" of a sequence by defining a confidence index for prediction. This index reflects the "informativity", *i.e.* the information content, of a target sequence according to the experts' ability to give a correct prediction. It is shown to be a valuable tool for estimating a theoretical correct prediction rate along a target protein sequence. Confidence indices can be useful tools to guide the generation of homology or *de novo* protein models involving local structure predictions. A hierarchical procedure for modelling a protein structure could be envisaged where first, the regions considered as accurately predicted would be selected and fixed and then regions with lower informativity would be examined more extensively, by considering for instance a larger number of candidates. As a whole, this procedure would significantly reduce the global conformational space sampling for the entire protein sequence.

These promising results suggest that our approach will prove effective in real biological and biochemical applications. Indeed, applications for our local structure prediction strategy are numerous and diverse. It can be included in *de novo* or homology modelling of global protein structures and for the prediction of loops.

Concerning the representation of the sequence, according to Pei & Grishin [9], combining evolutionary information with structural "preference" in amino acids should be a very promising way to improve local structure prediction methods. In our method this combination is implicitly done. An important point of our strategy is that PSSMs were derived from the complete target sequence. This choice made it possible to take into account the whole structural and physico-chemical environment of sequence fragments, and to learn amino acids preferences in sequence families. It

would be interesting to compare them with PSSMs calculated only on single fragments and to quantify the extent to which such fragment PSSMs capture long range interactions and properties. The quality of PSSM would also be an important point to study further. Actually, a benchmarking of the procedure for taking into account potential misalignments could lead to better prediction performances. As suggested by Altschul and co-workers, one way to refine alignments of PSI-BLAST would be to construct a multiple alignment of the sequences found and to calculate a new PSSM. This matrix can then be used in steps for rescoring and realigning database sequences [38]. Another important point is the handling of sequences with very few homologues for which the use of PSSM could be inappropriate (see supplementary data VIII). An analysis of raw PSI-BLAST alignments could help in identifying these cases and elaborating an adapted strategy for prediction.

Our future work will focus on global 3D structure prediction. Our approach will rely on one hand, on local structures prediction from sequence and on the other hand, on a fragment assembly strategy taking into account predicted global constraints. In this context, the use of the frequent sub-networks defined here will be particularly useful and the intrinsic flexibility of the structures will be taken into account. Beside its utility for better understanding the Hybrid Protein Model and for assessing prediction quality as done in this study, this new description will be particularly fruitful toward the proposition of 3D protein models. As a matter of fact, taking account of LSP transition probabilities should not only help to appropriately filter the results of the prediction by preventing the choice of incompatible LSPs in two successive sites, but also allow predictions for longer local structures.

Acknowledgements

References

1. Blundell TL, Sibanda BL, Montalvao RW, Brewerton S, Chelliah V, Worth CL, Harmer NJ, Davies O, Burke D. Structural biology and bioinformatics in drug design: opportunities and challenges for target identification and lead discovery. Philos Trans R Soc Lond B Biol Sci 2006;361(1467):413-423.

2. Noble ME, Endicott JA, Johnson LN. Protein kinase inhibitors: insights into drug design from structure. Science 2004;303(5665):1800-1805.

3. Doppelt O, Moriaud F, Bornot A, De Brevern AG. Functional annotation strategy for protein structures. Bioinformation 2007;1(9):357-359.

4. Fitzkee NC, Fleming PJ, Gong H, Panasik N, Jr., Street TO, Rose GD. Are proteins made from a limited parts list? Trends Biochem Sci 2005;30(2):73-80.

5. Zhang Y, Hubner IA, Arakaki AK, Shakhnovich E, Skolnick J. On the origin and highly likely completeness of single-domain protein structures. Proc Natl Acad Sci U S A 2006;103(8):2605-2610.

6. Yang AS, Wang LY. Local structure prediction with local structure-based sequence profiles. Bioinformatics 2003;19(10):1267-1274.

7. Bowie JU, Luthy R, Eisenberg D. A method to identify protein sequences that fold into a known three-dimensional structure. Science 1991;253(5016):164-170.

8. Rooman MJ, Rodriguez J, Wodak SJ. Relations between protein sequence and structure and their significance. J Mol Biol 1990;213(2):337-350.

9. Pei J, Grishin NV. Combining evolutionary and structural information for local protein structure prediction. Proteins 2004;56(4):782-794.

10. Sander O, Sommer I, Lengauer T. Local protein structure prediction using discriminative models. BMC Bioinformatics 2006;7:14.

11. Jauch R, Yeo HC, Kolatkar PR, Clarke ND. Assessment of CASP7 structure predictions for template free targets. Proteins 2007;69 Suppl 8:57-67.

12. Offmann B, Tyagi M, de Brevern AG. Local Protein Structures. Current Bioinformatics 2007;2:165-202.

13. Etchebest C, Benros C, Hazout S, de Brevern AG. A structural alphabet for local protein structures: improved prediction methods. Proteins 2005;59(4):810-827.

14. de Brevern AG, Etchebest C, Hazout S. Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. Proteins 2000;41(3):271-287.

15. de Brevern AG. New assessment of a structural alphabet. In Silico Biol 2005;5(3):283-289.

16. Karchin R, Cline M, Mandel-Gutfreund Y, Karplus K. Hidden Markov models that use predicted local structure for fold recognition: alphabets of backbone geometry. Proteins 2003;51(4):504-514.

17. Fourrier L, Benros C, de Brevern AG. Use of a structural alphabet for analysis of short loops connecting repetitive structures. BMC Bioinformatics 2004;5:58.

18. Tyagi M, Gowri VS, Srinivasan N, de Brevern AG, Offmann B. A substitution matrix for structural alphabet based on structural alignment of homologous proteins and its applications. Proteins 2006;65(1):32-39.

19. Tyagi M, Sharma P, Swamy CS, Cadet F, Srinivasan N, de Brevern AG, Offmann B. Protein Block Expert (PBE): a web-based protein structure

analysis server using a structural alphabet. Nucleic Acids Res 2006;34(Web Server issue):W119-123.

20. Tyagi M, de Brevern AG, Srinivasan N, Offmann B. Protein structure mining using a structural alphabet. Proteins 2007.

21. de Brevern AG, Wong H, Tournamille C, Colin Y, Le Van Kim C, Etchebest C. A structural model of a seven-transmembrane helix receptor: the Duffy antigen/receptor for chemokine (DARC). Biochim Biophys Acta 2005;1724(3):288-306.

22. Etchebest C, Benros C, Bornot A, Camproux AC, de Brevern AG. A reduced amino acid alphabet for understanding and designing protein adaptation to mutation. Eur Biophys J 2007;36(8):1059-1069.

23. Dong QW, Wang XL, Lin L. Methods for optimizing the structure alphabet sequences of proteins. Comput Biol Med 2007;37(11):1610-1616.

24. Dudev M, Lim C. Discovering structural motifs using a structural alphabet: application to magnesium-binding sites. BMC Bioinformatics 2007;8:106.

25. Benros C, de Brevern AG, Etchebest C, Hazout S. Assessing a novel approach for predicting local 3D protein structures from sequence. Proteins 2006;62(4):865-880.

26. de Brevern AG, Hazout S. Compacting local protein fold with a 'hybrid protein model'. Theorical Chemistry Accounts 2001;106(1/2):36-47.

27. de Brevern AG, Hazout S. 'Hybrid protein model' for optimally defining 3D protein structure fragments. Bioinformatics 2003;19(3):345-353.

28. Kohonen T. Self-organizing formation of topologically correct feature maps. Biol Cybernet 1982;43:59-69.

29. Kohonen T. Self-Organizing Maps. Berlin, Germany; 2001.

30. Bystroff C, Baker D. Prediction of local structure in proteins using a library of sequence-structure motifs. J Mol Biol 1998;281(3):565-577.

31. Baeten L, Reumers J, Tur V, Stricher F, Lenaerts T, Serrano L, Rousseau F, Schymkowitz J. Reconstruction of protein backbones from the BriX collection of canonical protein fragments. PLoS Comput Biol 2008;4(5):e1000083.

32. Sawada Y, Honda S. ProSeg: a database of local structures of protein segments. J Comput Aided Mol Des 2008.

33. Song J, Burrage K, Yuan Z, Huber T. Prediction of cis/trans isomerization in proteins using PSI-BLAST profiles and secondary structure information. BMC Bioinformatics 2006;7:124.

34. Ward JJ, McGuffin LJ, Buxton BF, Jones DT. Secondary structure prediction with support vector machines. Bioinformatics 2003;19(13):1650-1655.

35. Eddy SR. Where did the BLOSUM62 alignment score matrix come from? Nat Biotechnol 2004;22(8):1035-1036.

36. de Brevern AG, Valadie H, Hazout S, Etchebest C. Extension of a local backbone description using a structural alphabet: a new approach to the sequence-structure relationship. Protein Sci 2002;11(12):2871-2886.

37. Noguchi T, Matsuda H, Akiyama Y. PDB-REPRDB: a database of representative protein chains from the Protein Data Bank (PDB). Nucleic Acids Res 2001;29(1):219-220.

38. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25(17):3389-3402.

39. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, Schneider M. The

SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res 2003;31(1):365-370.

40. Chang C-C, Lin C-J. LIBSVM: a library for support vector machines. Software available at http://wwwcsientuedutw/~cjlin/libsvm 2001.

41. Hastie T, Tibshirani R, Friedman J. The Elements of statistical learning – Data mining, Inference, and Prediction: Springer Series in Statistic; 2001.

42. Joachims T. Making large-Scale SVM Learning Practical. In: Schölkopf B, Burges C, Smola A, editors. Advances in Kernel Methods - Support Vector Learning: MIT-Press; 1999.

43. Lewis DP, Jebara T, Noble WS. Support vector machine learning from heterogeneous data: an empirical analysis using protein sequence and structure. Bioinformatics 2006;22(22):2753-2760.

44. Hsu CW, Chang CC, Lin CJ. A practical guide to support vector classification. Tech Rep, Department of computer science and information engineering, National Taiwan University, Tapei, Taiwan [Available at http://wwwcsientuedutw/~cjlin/papershtml] 2003.

45. Kuang R, Ie E, Wang K, Wang K, Siddiqi M, Freund Y, Leslie C. Profile-based string kernels for remote homology detection and motif extraction. Proc IEEE Comput Syst Bioinform Conf 2004:152-160.

46. Rohl CA, Strauss CE, Misura KM, Baker D. Protein structure prediction using Rosetta. Methods Enzymol 2004;383:66-93.

47. Benros C. Analyse et prediction des structures tridimensionnelles locales des proteines [Analyse de Genomes et Modelisation Moleculaire]. Paris, France: University Paris 7 - Denis Diderot; 2005. 211 p.

48.   Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247(4):536-540.

49.   Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999;292(2):195-202.

50.   Lin HN, Chang JM, Wu KP, Sung TY, Hsu WL. HYPROSP II--a knowledge-based hybrid method for protein secondary structure prediction based on local prediction confidence. Bioinformatics 2005;21(15):3227-3233.

51.   Hazout S. Entropy-derived measures for assessing the accuracy of N-state prediction algorithms. In: de Brevern AG, editor. In Recent Advances in Structural Bioinformatics. Trivandrum, India: Research signpost; 2007. p pp. 395-417.

52.   de Brevern AG, Etchebest C, Benros C, Hazout S. "Pinning strategy": a novel approach for predicting the backbone structure in terms of protein blocks from sequence. J Biosci 2007;32(1):51-70.

53.   Ihaka R, Gentleman R. R: a language for data analysis and graphics. J Comp Graph Stat 1996;5:229-314.

54.   Bystroff C, Thorsson V, Baker D. HMMSTR: a hidden Markov model for local sequence-structure correlations in proteins. J Mol Biol 2000;301(1):173-190.

55.   Kuang R, Leslie CS, Yang AS. Protein backbone angle prediction with machine learning approaches. Bioinformatics 2004;20(10):1612-1621.

56.   Fawcett T. ROC Graphs: Notes and Practical Considerations for Data Mining Researchers. Intelligent Enterprise Technologies Laboratory 2003.

57. Xiang Z, Soto CS, Honig B. Evaluating conformational free energies: the colony energy and its application to the problem of loop prediction. Proc Natl Acad Sci U S A 2002;99(11):7432-7437.

58. Zhu K, Pincus DL, Zhao S, Friesner RA. Long loop prediction using the protein local optimization program. Proteins 2006;65(2):438-452.

59. Soto CS, Fasnacht M, Zhu J, Forrest L, Honig B. Loop modeling: Sampling, filtering, and scoring. Proteins 2008;70(3):834-843.

60. Fiser A, Do RK, Sali A. Modeling of loops in protein structures. Protein Sci 2000;9(9):1753-1773.

61. de Bakker PI, DePristo MA, Burke DF, Blundell TL. Ab initio construction of polypeptide fragments: Accuracy of loop decoy discrimination by an all-atom statistical potential and the AMBER force field with the Generalized Born solvation model. Proteins 2003;51(1):21-40.

62. Zhou H, Zhou Y. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Sci 2002;11(11):2714-2726.

63. Rangwala H, Karypis G. Profile-based direct kernels for remote homology detection and fold recognition. Bioinformatics 2005;21(23):4239-4247.

64. Mittelman D, Sadreyev R, Grishin N. Probabilistic scoring measures for profile-profile comparison yield more accurate short seed alignments. Bioinformatics 2003;19(12):1531-1539.
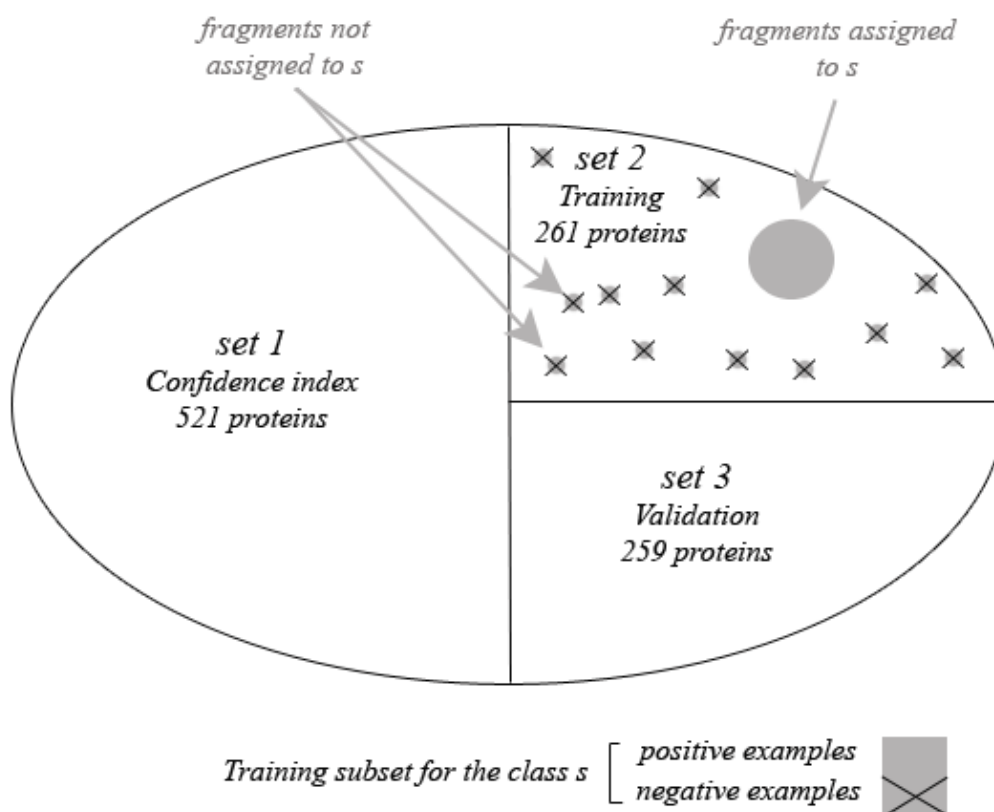
Figures

Figure 1 - Datasets organisation and usage

S*et 1* was used for learning the relation between confidence index categories and prediction rate. *Set 2* was used as a training set for prediction and for the development of the confidence index. *Set 3* for prediction and confidence index assessment.
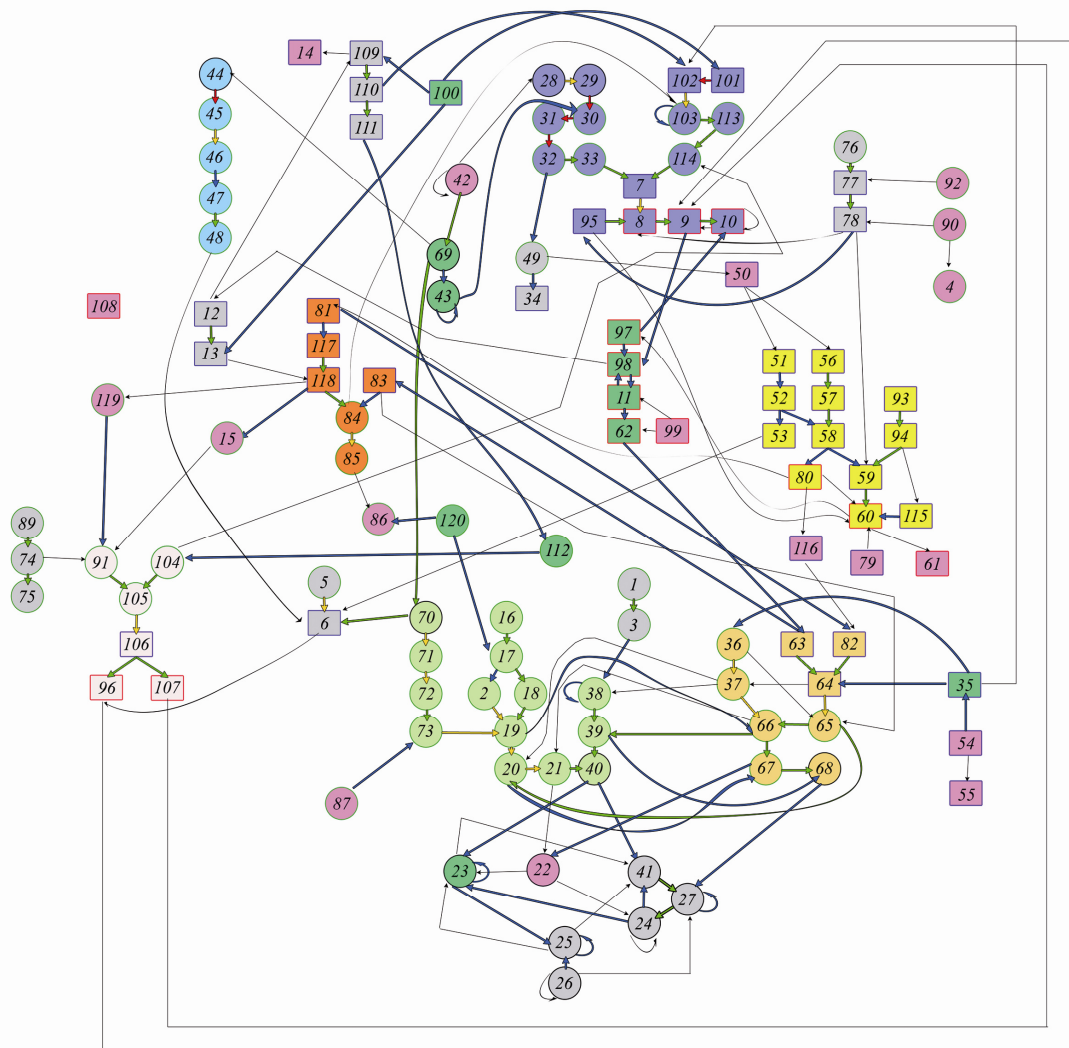
Figure 2 - Global preferential structural transition network between prototypes.

For a better representation, only transitions with a probability greater than 0.15 are shown. These represented transitions, does not represent the whole complexity of the protein structure space but it shows in average 45.56 % (σ ± 22.22) of transitions from a given LSPs to one, two or three others. It takes into account the 51% of the most frequent transitions observed. Actually, each LSP has at least a 0.15 probability to succeed or precede another given LSP. The only exception is the LSP 108.

Transitions with a probability between 0.15 and 0.20 are shown in thin black arrows. Blue, green and yellow arrows show respectively 0.20 to 0.30, 0.30 to 0.50 and 0.50 to 0.70 probability transitions. Red arrows show transitions with probability more than 0.70. Circular nodes with black and green border correspond to helical and connecting LSPs respectively. Rectangular nodes with red and blue border correspond to extended and extended edge LSPs respectively. Seven preferential transition sub-network of LSPs have been defined (see Figure 3). Each node is filled with a color indicating its transition category.
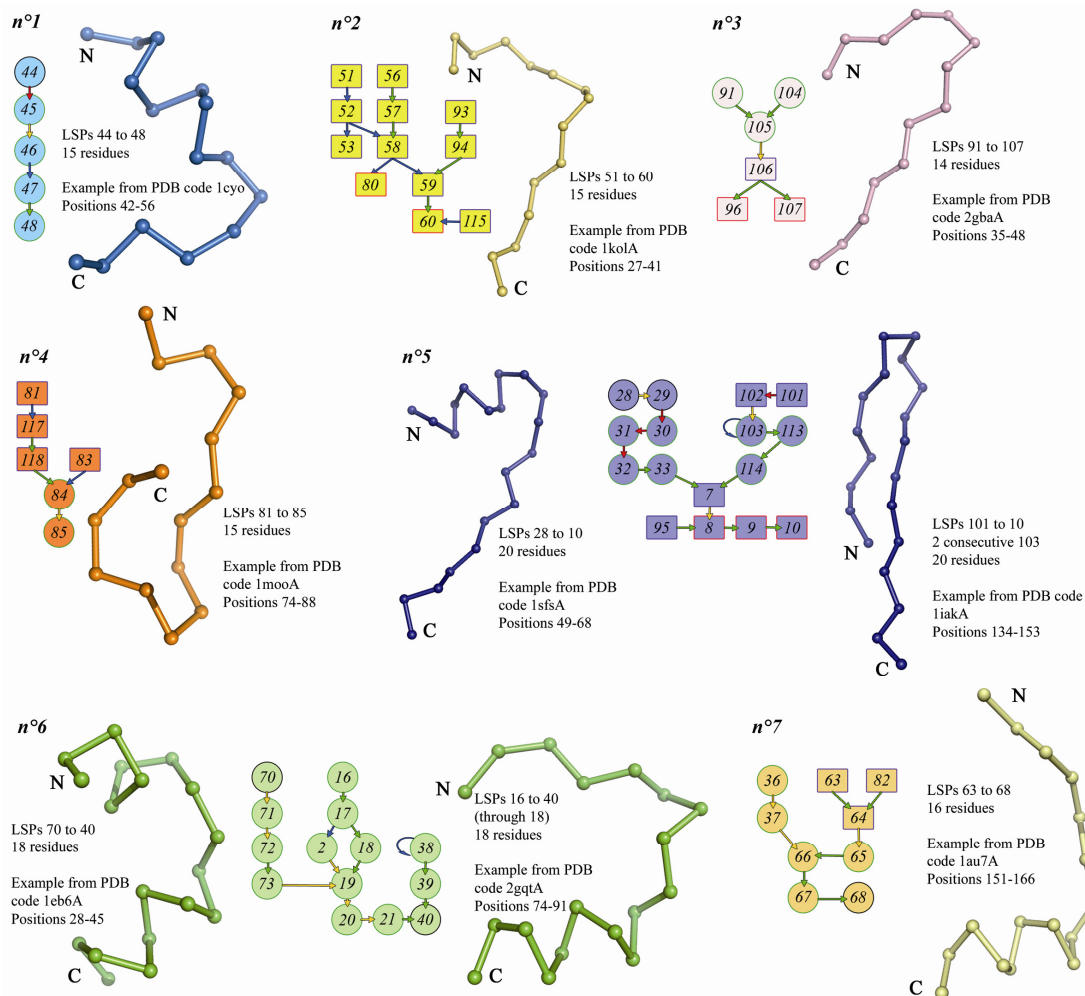
Figure 3 - 7 LSP sub-networks represent significant transition probabilities.

These sub-networks made it possible to assess of the prediction of LSPs involved in frequent super-local protein structures. Examples of these local structures in proteins are given next to the corresponding sub-network. Their position in proteins, their assigned prototypes chain and length are indicated.

Only transitions greater than 0.2 are represented. Blue, green, yellow and Red arrows show 0.2 to 0.3, 0.3 to 0.5, 0.5 to 0.7 and more than 0.7 probability transitions respectively. Circular nodes with black and green border correspond to Helical and Connecting LSPs respectively. Rectangular nodes with red and blue border correspond to Extended and Extended edge LSPs respectively**.**
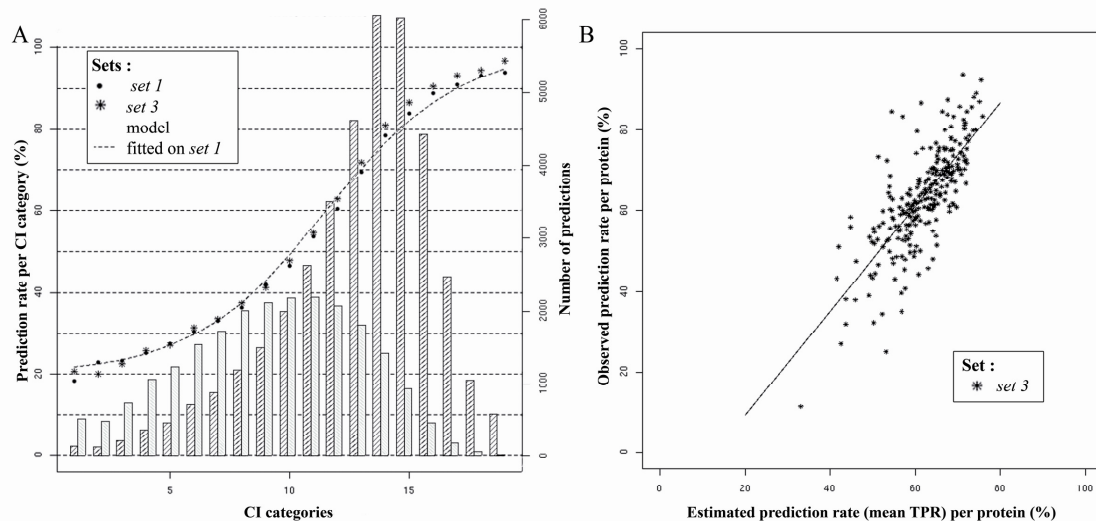
Figure 4 - Definition and assessment of a confidence index (CI).

A: representation of the prediction rate associated with each category of CI on *set 1* and on *set 3*. A model of the relation between categories and prediction rate was fitted on *set 1* and is shown by a dashed line. The histogram represents the number of correct (dark stripes) and of incorrect (light stripes) predictions for each CI category on *set 3*. CI Categories numbered from 1 to 19 correspond to the following intervals of SVM outputs:]-∞, -1.32],]-1.32, -1.14],]-1.14, -0.95],]-0.95, -0.77],]-0.77, -0.58],]-0.58, -0.39],]-0.39, -0.21],]-0.21, -0.02],]-0.02, 0.17],]0.17, 0.36],]0.36, 0.54],]0.54, 0.73],]0.73, 0.92],]0.92, 1.10],]1.10, 1.29],]1.29, 1.48],]1.48, 1.67],]1.67, 1.85],]1.85, +∞]. B: representation of the prediction rate per protein observed on *set 3* according to the mean theoretical prediction rate (TPR) estimated per protein based on to CI values categories. The model corresponding to the linear relation between these 2 variables is shown by a solid line.
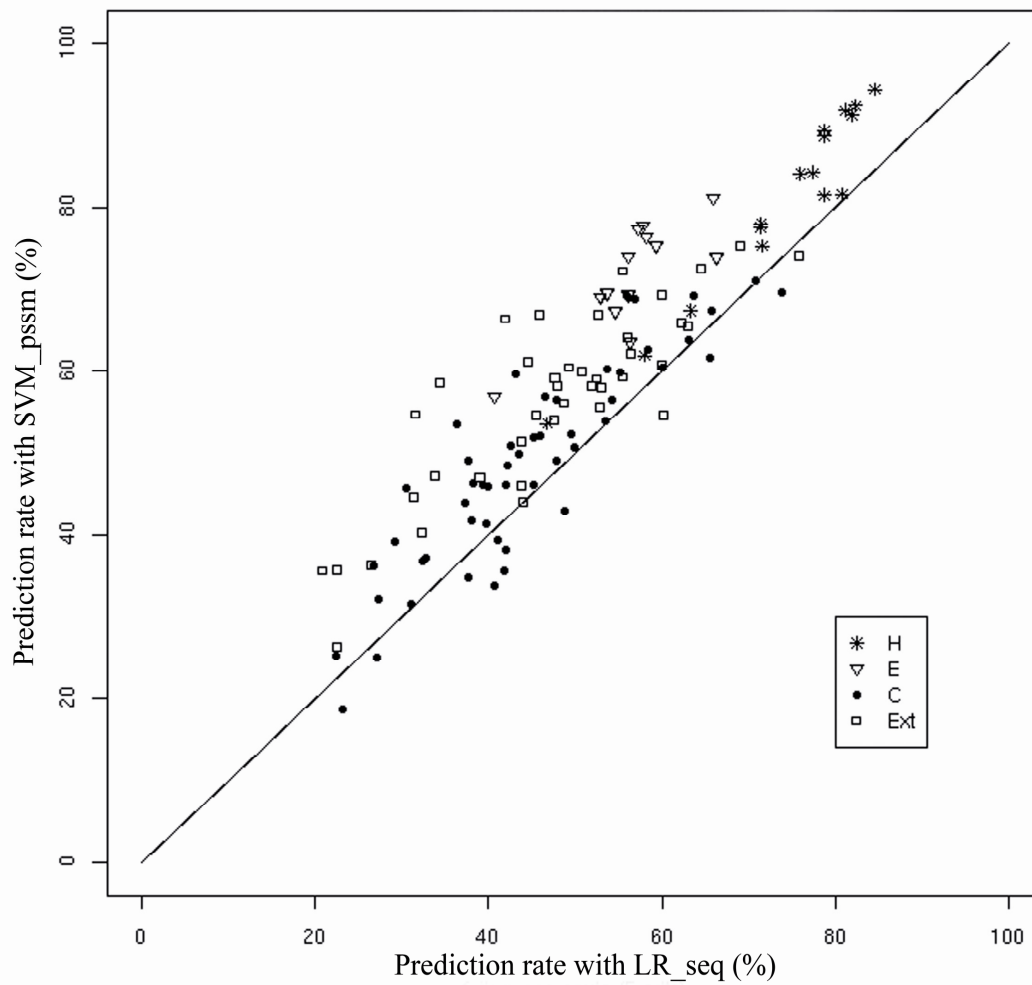
Figure 5 - Global improvement of the local structure prediction per structural classes using the *SVM_PSSM* strategy.

Comparison of the 120 prediction rates obtained using the *LR_seq* and the *SVM_PSSM* strategies.
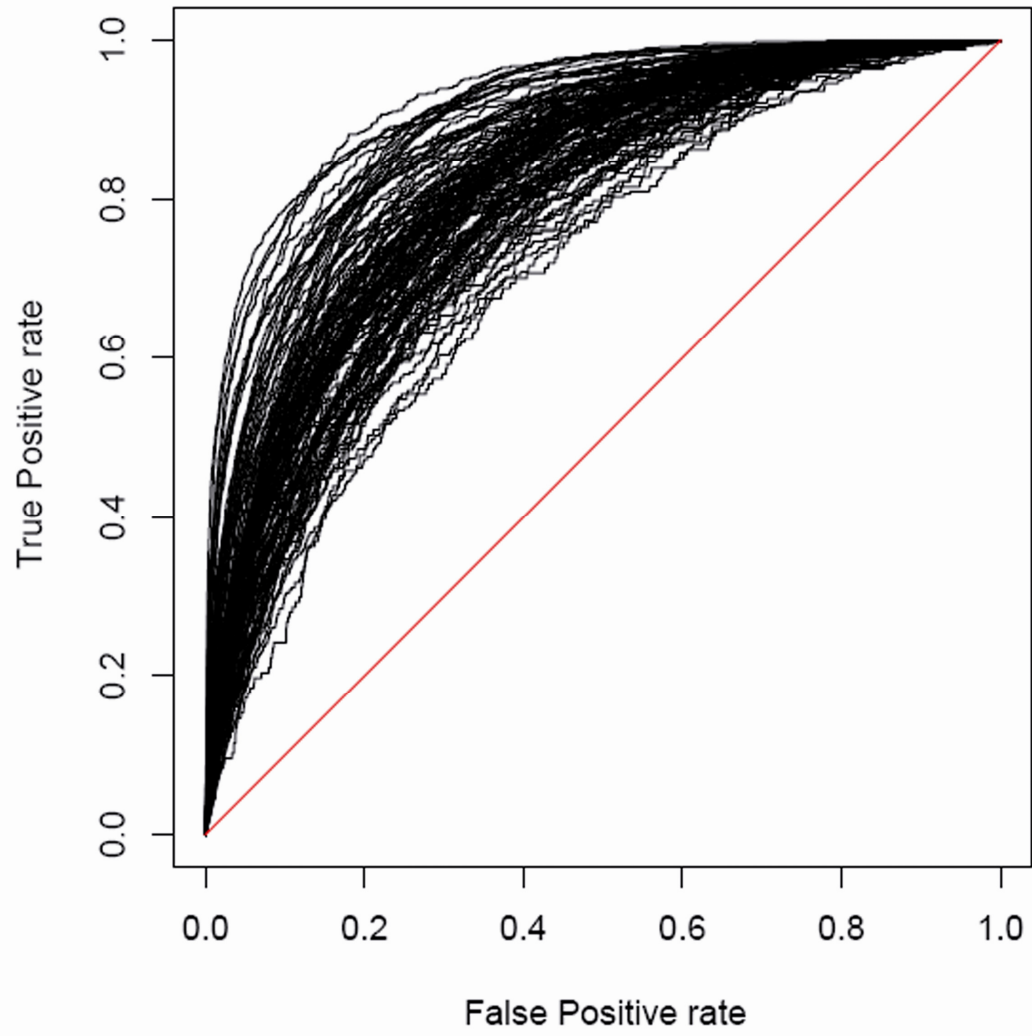
Figure 6 – ROC curves for the 120 structural classes

Tables

Table I - Local structure prediction results

Analysis of the structural prediction results					
Experts definitions and target sequence window representation	*LR_seq*	*SVM _seq*	*SVM _PSSM*	*SVM _PSSM gains over random*	*SVM_PSSM gains over Similar Sequences Search*
Proportion of true positives (%)	31.43	30.61	38.75	34.56	18.34
Prediction rate (%) (approximation < 2.5Å)	**55.48**	**55.54**	**63.13**	**37.95**	**15.18**
Results per secondary structures categories (%)					
H *Proportion of true positives*	40.06	39.10	50.68	46.54	23.47
Prediction rate (approximation < 2.5Å)	**76.80**	**75.45**	**84.60**	**43.01**	**3.46**
E *Proportion of true positives*	20.95	25.45	34.08	30.19	18.82
Prediction rate (approximation < 2.5Å)	**57.43**	**63.18**	**73.03**	**39.26**	**23.94**
C *Proportion of true positives*	33.49	31.46	38.03	33.72	17.31
Prediction rate (approximation < 2.5Å)	**45.06**	**43.85**	**49.47**	**36.52**	**16.17**
Ext *Proportion of true positives*	22.89	21.80	28.35	24.16	13.97
Prediction rate (approximation < 2.5Å)	**47.10**	**48.31**	**56.30**	**33.88**	**23.22**

Table II – Structural accuracy of the Local structure prediction.

Average Geometrical Approximation of the Local Structure Prediction (A)		All predicted fragments		Fragments correctly predicted according to the geometrical criteria (<2.5 Å)	
		Minimal RMSD*	Mean RMSD*	Minimal RMSD*	Mean RMSD*
Set 3		2.09	3.03	1.45	2.54
Secondary Structures Categories	H	1.21	2.28	0.83	1.95
	E	2.17	3.14	1.78	2.77
	C	2.48	3.41	1.75	2.93
	Ext	2.41	3.20	1.93	2.86
Transition Categories	1	2.22	3.18	1.71	2.82
	2	2.29	3.07	1.91	2.78
	3	2.30	3.13	1.85	2.76
	4	2.80	3.63	2.05	3.04
	5	2.27	3.42	1.60	2.96
	6	2.27	3.19	1.58	2.74
	7	2.17	3.11	1.53	2.68
	8	1.94	2.93	1.43	2.51
	9	1.56	2.43	0.98	1.93
	10	2.17	3.14	1.66	2.66

*over the 5 candidates per fragment

Table III – Local structure prediction per transition category.

Analysis of the structural prediction results per transition categories		1	2	3	4	5	6	7	8	9	10
SVM_PSSM	*Proportion of true positives*	45.64	29.46	33.24	23.48	45.62	42.42	39.96	37.11	41.09	27.72
	Prediction rate (approximation < 2.5Å)	64.22	64.22	62.97	36.87	60.88	55.95	59.15	71.71	72.71	57.57
Prediction rate gains over random (approximation < 2.5Å)		41.83	33.85	38.50	25.53	42.82	38.22	40.60	38.71	37.73	31.76
Prediction rate gains over Similar Sequences Search (approximation < 2.5Å)		20.81	29.13	24.73	12.33	15.24	15.83	18.95	14.61	8.79	13.37
Prediction rate gains over LR_seq (approximation < 2.5Å)		6.42	15.98	5.95	4.22	4.90	6.49	6.48	10.97	8.83	6.02

Table IV – Comparison with Sander & al. [10] local structure prediction.

	True Positive rate for the top-k ranked predictions (%)		
	Property Profile+RF	*SVM_PSSM*	
Rank k	*7-residue local structures 27 classes*	*11-residue local structures 120 classes*	*11-residue local structures 27 classes*
1	34	13.56	31.84
2	46	22.24	44.08
3	54	28.91	51.6
4	60	34.22	57.05
5	64	38.75	61.42