Learning the Pattern-based CRF for Prediction of a Protein Local Structure

Zhalgas Mukanov[1] and Rustem Takhanov[2]
E-mail: mukanovj@mail.ru, rustem.takhanov@nu.edu.kz
[1]Fundamental Mathematics Department, Eurasian National University, 2 Satpayev Str., Nur-Sultan, Kazakhstan
[2]Mathematics Department, Nazarbayev University, 53 Kabanbay Batyr Ave, Nur-Sultan, Kazakhstan

Keywords: pattern-based CRFs, sequence labeling, protein conformation prediction, energy landscape, structural SVM

Received: October 13, 2021

Prediction of protein conformation from its amino acid sequence is widely acknowledged as one of the most important computational biology problems and is considered a source of interesting problem formulations for machine learning. Here methods of supervised learning stay side by side with statistical physics and information theory. According to classical results of Anfinsen, protein conformational structure is fully determined by its primary structure, i.e., amino acid sequence, and energy landscape theory says that the native state of a protein corresponds to the minimum of its free energy [2].
There are two dominating approaches to protein structure prediction, the first is based on minimizing physics-based free energies with some unknown parameters, and the second is a knowledge-based approach that does not necessarily use the notion of free energy and aims only to yield high prediction accuracy [14]. In comparison to these two approaches, there is a deficit in intermediate approaches where the goal is to find such knowledge-based parameterizations of free energy that would approximate real free energy for certain protein families and have a high accuracy of prediction comparable with pure knowledge-based approaches. According to M. Gromov, if energy landscape theory is true, then "probably, free energy can be encoded with a reasonable accuracy by something like $10^4 - 10^6$ bits of information", and the main mathematical problem here is the lack of "general mathematical "parameter fitting" method(s), which, when applied to proteins, could provide (an effective version of) the total inter-residue interaction energies" [10]. In this paper, we introduce a probabilistic model based on a certain parametrization of free energy that we expect could be fruitful both for predicting protein dihedral angles and investigating the structure of the energy landscape. This model is based on the idea that free energy is largely determined by pairwise interactions of amino acids that are located near each other on a protein sequence. Though this approach is far from reality for general proteins, we expect it to approximate an all-alpha protein's energy landscape.

Povzetek: Za določanje strukture beljakovin je bila razvita nova metoda, ki se uči pogostih vzorcev.

1 Introduction

Prediction of protein conformation from its amino acid sequence is widely acknowledged as one of the most important computational biology problems and is considered a source of interesting problem formulations for machine learning. Here methods of supervised learning stay side by side with statistical physics and information theory. According to classical results of Anfinsen, protein conformational structure is fully determined by its primary structure, i.e., amino acid sequence, and energy landscape theory says that the native state of a protein corresponds to the minimum of its free energy [2].

There are two dominating approaches to protein structure prediction, the first is based on minimizing physics-based free energies with some unknown parameters, and the second is a knowledge-based approach that does not necessarily use the notion of free energy and aims only to yield high prediction accuracy [14]. In comparison to these two approaches, there is a deficit in intermediate approaches where the goal is to find such knowledge-based parameterizations of free energy that would approximate real free energy for certain protein families and have a high accuracy of prediction comparable with pure knowledge-based approaches. According to M. Gromov, if energy landscape theory is true, then "probably, free energy can be encoded with a reasonable accuracy by something like $10^4 - 10^6$ bits of information", and the main mathematical problem here is the lack of "general mathematical "parameter fitting" method(s), which, when applied to proteins, could provide (an effective version of) the total inter-residue interaction energies" [10]. In this paper, we introduce a probabilistic model based on a certain parametrization of free energy that we expect could be fruitful both for predicting protein dihedral angles and investigating the structure of the energy landscape. This model is based on the idea that free energy is largely determined by pairwise interactions of amino acids that are located near each other on a protein sequence. Though this approach is far from reality for general proteins, we expect it to approximate an all-alpha protein's energy landscape.

1.1 Related work

There are plenty of publications in literature dedicated to the problem of protein backbone dihedral angles prediction. This problem is interesting in two contexts. First, since secondary structure and backbone dihedral angles correlate (especially in the alpha-helix region), these problems are often tackled together and considered adjacent research themes [13]. Second, it has been shown that high accuracy prediction of secondary structure/dihedral ϕ, ψ angles improves the recognition of the so-called fold of a protein [13]. Pioneering works on secondary structure prediction were published in 70-s [5, 8]. The highest prediction accuracies were achieved in the middle of the 90s by machine learning techniques that use multiple sequence alignment with proteins from the PDB database. PSIPred is a popular and high-scoring example of an algorithm of this kind [11]. Later, using a similar representation of a protein as in PSIPred, the idea of simultaneous prediction of secondary structure and backbone dihedral angles was implemented in DISSPred, which has one of the highest accuracies for both these problems [13]. A survey of the most recent advances in the problem can be found in [25]. Such approaches improve their accuracy as the number of resolved proteins grows, and they weaken if the template protein(the one for which a prediction is made) does not have close homologs among resolved proteins.

In the absence of close homologs, prediction methods based on sequence-structure analysis are considered as one of the promising [6, 4]. The success of such techniques is based on the fundamental fact that the complexity and diversity of local conformational structures observed in proteins are much less than sequence complexity. The first and critical step in such techniques is a choice of structural motives alphabet; then, correlations between sequence and corresponding structure are retrieved from data, and a prediction of secondary/local structure is made. One of the ways to formalize such a scheme is the Hidden Markov Models and their modifications. In the majority of such algorithms, a state of hidden Markov process formalizes structural information associated with a certain amino acid of a protein and transition probabilities between states of adjacent amino acids computed by standard formulas from data. HMMSTR is one of the most popular methods based on HMMs [4]. The idea of applying structural SVM machinery for computation of HMM parameters for secondary structure recognition was implemented in [9, 1].

2 Pattern-based energy

A *pattern-based* energy potential over words $x \in D^n$ in some alphabet D is defined as

$$E(x) = \sum_{\alpha \in \Gamma} \sum_{\substack{[i,j] \subseteq [1,n] \\ j-i+1=|\alpha|}} f_{ij}^{\alpha} \cdot [x_{i:j} = \alpha] \qquad (1)$$

where $\Gamma \subseteq D^*$ is a fixed set of non-empty words, $|\alpha|$ is the length of word α and $[\cdot]$ is the *Iverson bracket*. There $x_{i:j}$ denotes a subword $x_i \cdots x_j$ of x. A pattern-based conditional random field is defined as the probability distribution over words $p(x) \propto e^{-E(x)}$.

Intuitively, pattern-based CRFs allow modeling long-range interactions for selected subsequences of labels. Inference algorithms for pattern-based CRFs were developed in [26, 21] and they include (i) computing the partition function $Z = \sum_x \exp\{-E(x)\}$; (ii) computing marginal probabilities $p(x_{i:j} = \alpha)$ for all triplets (i, j, α) present in (1); (iii) computing MAP, i.e. minimizing energy (1). Note that MAP problem is a special case of hybrid valued constraint satisfaction problems [12, 18, 19]. Hybrid VCSPs in general can be NP-hard, but minimizing the energy (1) is not only tractable but even solvable in time $O(n)$.

Pattern-based CRFs were already applied in such contexts as handwritten character recognition, identification of named entities from text [26], optical character recognition [16] and the language modeling [3, 20, 22].

3 Pattern-based CRFs for the sequence labeling problem

One of the important classes of pattern recognition problems with structural outputs is sequence labeling problems. In such problems, we are given two finite alphabets, A (input alphabet) and L (output labels). A training set consists of pairs of the form (x, y) where x (y) is a word over A (L) and both words have the same length, i.e. letters of the first word are tagged by labels. The goal is to construct mappings $m : A^n \to L^n, n = 1, 2, ...,$ that both consistent with a training set and some supplementary model requirements. Examples of such formulations can be found in protein folding (secondary structure prediction), natural language processing (part-of-speech tagging), and speech recognition. Popular methods used for sequence labeling learning include hidden and maximum entropy Markov models. In this paper, we will describe a generalized version of our model and show how two key problems of inference and learning can be efficiently solved in this framework.

Definition. A pair (α, β), where α is a word over A and β is a word over L, is called a pattern pair.

Suppose we are given a pattern pair (α, β) and tuples $x = (x_1, ..., x_n) \in A^n$ and $y = (y_1, ..., y_n) \in L^n$. Then by $(\alpha, \beta) \vdash^i (x, y)$ we say that $x_{i:i+|\alpha|-1} = \alpha$ and $y_{i:i+|\alpha|-1} = \beta$.

Suppose we are given a finite set of pattern pairs $\mathfrak{A}$. Then, for any tuples $x = (x_1, ..., x_n) \in A^n$ and $y = (y_1, ..., y_n) \in L^n$, $\Psi_{\mathfrak{A}}(x, y)$ denotes a vector with components indexed by $\mathfrak{A}$, and for $(\alpha, \beta) \in \mathfrak{A}$, (α, β)-component is equal to the number of $i \in \{1, ..., n\}$ for which $(\alpha, \beta) \vdash^i (x, y)$.

If any pattern pair from $\mathfrak{A}$ is assigned a real value, then parameters $\overline{w} = \{w_i\}_{i \in \mathfrak{A}}$ define a conditional probability distribution $p\left(y|x,\overline{w}\right)$ where $x\in A^n, y\in L^n$:

$$p\left(y|x,\overline{w}\right)=\frac{e^{-\langle\overline{w},\Psi_{\mathfrak{A}}(x,y)\rangle}}{Z\left(x,\overline{w}\right)} \tag{2}$$

where $Z\left(x,\overline{w}\right)=\sum_y e^{-\langle\overline{w},\Psi_{\mathfrak{A}}(x,y)\rangle}$. It is easy to see that this family of conditional probability distributions is a special case pattern-based CRF.

Two computational problems are of first interest in the framework of conditional random fields: inference and learning. Inference for pattern-based CRF is equivalent to minimization of energy:

$$\max_y p\left(y|x,\overline{w}\right)=\min_y\langle\overline{w},\Psi_{\mathfrak{A}}\left(x,y\right)\rangle. \tag{3}$$

If we rewrite our energy term as

$$\langle\overline{w},\Psi_{\mathfrak{A}}\left(x,y\right)\rangle= \sum_{p\in\mathfrak{A}}\sum_{i:p\vdash^i(x,y)}w_p=\sum_{i=1}^{n}\sum_{p\in\mathfrak{A}:p\vdash^i(x,y)}w_p \tag{4}$$

we will see that the inference is equivalent to minimization of pattern-based functional over y with $f_{ij}^{\alpha}=\sum_{(\alpha,\phi)\in\mathfrak{A}:(\alpha,\phi)\vdash^i(x,y)}w_{(\alpha,\phi)}$.

Given a training sample $\left\{\left(x^i,y^i\right)\right\}_{i=\overline{1,n}}$ maximum likelihood parameter learning is the following task:

$$\prod_{i=1}^{n}p\left(y^i|x^i,\overline{w}\right)\to\max_{\overline{w}} \tag{5}$$

which after standard operations is equivalent to

$$L\left(\overline{w}\right)=\sum_{i=1}^{n}\langle\overline{w},\Psi_{\mathfrak{A}}\left(x^i,y^i\right)\rangle+\log Z\left(x^i,\overline{w}\right)\to\min_{\overline{w}} \tag{6}$$

Let us consider generalized version of pattern-based CRF by adding equivalence relation on patterns set $\mathfrak{A}$, $\sim$. Then we have additional constraints on learned weights: if $p\sim p'$ then $w_p=w_{p'}$. Such models appear to address the problem of overfitting when $|\mathfrak{A}|$ becomes high. From now, we will assume that weights are indexed by equivalence classes of $\sim$.

Since our energy is parameterized linearly, $L\left(\overline{w}\right)$ is convex. Therefore we can optimize it using gradient-based optimization(using first order or second order gradient descent) [15]. Elements of Jacobian and Hessian of $L\left(\overline{w}\right)$ are equal to the following sums:

$$\left[\nabla_{\overline{w}}L\left(\overline{w}\right)\right]_c=\sum_{i=1}^{n}\Big[\sum_{p\in c}\Big(\left[\Psi_{\mathfrak{A}}\left(x^i,y^i\right)\right]_p-\mathbb{E}_{y\sim p(y|x^i,\overline{w})}\left[\Psi_{\mathfrak{A}}\left(x^i,y\right)\right]_p\Big)\Big] \tag{7}$$

$$\left[H_{\overline{w}}L\left(\overline{w}\right)\right]_{cc'}=\sum_{i=1}^{n}\sum_{p\in c,p'\in c'}\mathbb{E}_{y\sim p(y|x^i,\overline{w})}\left[\Psi_{\mathfrak{A}}\left(x^i,y\right)\right]_p\times\mathbb{E}_{y'\sim p(y|x^i,\overline{w})}\left[\Psi_{\mathfrak{A}}\left(x^i,y'\right)\right]_{p'} \tag{8}$$

where

$$\left[\Psi_{\mathfrak{A}}\left(x^i,y^i\right)\right]_p=\left|\left\{k:p\vdash^k\left(x^i,y^i\right)\right\}\right| \tag{9}$$

It is easy to see that computing Jacobian and Hessian can be reduced to computation of marginal probabilities $p\left(y_{k:l}=\alpha|x^i,\overline{w}\right)$ in the pattern-based CRF $p\left(y|x^i,\overline{w}\right)$. Thus, we can apply BFGS algorithm [7] to solve (6) and to learn the weights of patterns in pattern-based CRFs (3).

4 Pattern-based conditional random field for dihedral angles prediction

Suppose we have a two ordered sets of variables $X=\{X_1,...,X_n\}$ and $Y=\{Y_1,...,Y_{2n}\}$. Variables X_i take their values in a set of amino acids

$$\{\mathrm{Ala},\mathrm{Arg},...,\mathrm{Val},\mathrm{Sec},\mathrm{Pyl}\}$$

(amino acid symbols) and we interpret any initialization of X as an amino acid sequence. Any amino acid in a protein corresponds to two dihedral angles ϕ and ψ in a protein (figure 1), and we interpret Y_{2i-1} and Y_{2i} as ϕ and ψ for amino acid X_i. Before any experiment, we will discretize an interval $[-180,180]$ and, therefore, will always imply that Y_i takes its values from some fixed finite set D.

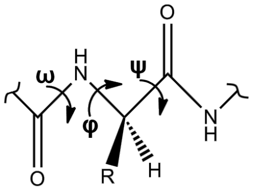

Figure 1: Angles ϕ, ψ and ω.

A triple (A_1,A_2,α), where A_1,A_2 are amino acid symbols and α is a word of even length over alphabet D, is called a local contact triple. The name contact triple is related to the following interpretation: suppose that values of ϕ,ψ angles within certain accuracy can be predicted based on the local interaction of amino acids (by local interaction, we understand the interaction between amino acids that are closely located on a sequence). Then the mutual location of two amino acids can be described by a list of dihedral angles between them. Here we neglect ω angles and torsion angles of these amino acids.

Suppose we are given a finite set of contact triples $\mathfrak{A}$. Then, for any pair $X=\{X_1,...,X_n\}$ and $Y=\{Y_1,...,Y_{2n}\}$, $\Psi_{\mathfrak{A}}\left(X,Y\right)$ denotes a vector with components indexed by $\mathfrak{A}$, and for $(A_1,A_2,\alpha)\in\mathfrak{A}$, (A_1,A_2,α)-component is equal to the number of $i\in\{1,...,n\}$ for which $X_i=A_1,X_{i+|\alpha|/2-1}=A_2,Y_{2i-1}Y_{2i}...Y_{2i+|\alpha|-3}Y_{2i+|\alpha|-2}=\alpha$.

Suppose pairs X,Y are obtained from some probability distribution $p\left(Y|X,\overline{w}\right)$ with unknown parameters $\overline{w}=$

$\{w_i\}_{i\in\mathfrak{A}}$. The family of probability distributions

$$-\log p\left(Y|X,\overline{w}\right) = C\left(X\right) + \left\langle \overline{w}, \Psi_{\mathfrak{A}}\left(X,Y\right)\right\rangle \tag{10}$$

where $C\left(X\right) = \log\sum_{Y} e^{-\langle \overline{w}|\Psi_{\mathfrak{A}}(X,Y)\rangle}$, is our pattern-based conditional random field.

After denoting $E_{\overline{w}}\left(X,Y\right) = \left\langle \overline{w}, \Psi_{\mathfrak{A}}\left(X,Y\right)\right\rangle$, it is easy to see that

$$\max_{Y} p\left(Y|X,\overline{w}\right) = \min E_{\overline{w}}\left(X,Y\right). \tag{11}$$

Therefore, if parameters are known ($\overline{w} = \overline{w^{*}}$), for input string X inference is equivalent to minimization of $\left\langle \overline{w}, \Psi_{\mathfrak{A}}\left(X,Y\right)\right\rangle$.

In the following section, we will describe the structural SVM procedure that we used to obtain weights.

4.1 Learning parameters by structural SVM

Given a training set $\{(x_i,y_i)\}_{i=1}^{l}$, the structural risk minimization approach reduces the problem of CRFs learning to minimization of the following functional:

$$\frac{C}{l}\sum_{i=1}^{l}\Delta\left(y_i, f_w(x_i)\right) \to \min \tag{12}$$

where $f_w(x_i) = \arg\min\limits_{y}\left\langle w, \Psi\left(x_i,y\right)\right\rangle$ and $\Delta : Y\times Y \to \mathbb{N}$ is a loss function on a set Y. The latter problem is very hard in general, and a standard way to tackle it is to replace a function with its convex upper bound [15]. Following this scheme,we used structural SVM technique to learning pattern-based CRFs. After changing $\Psi\left(x,y\right) \to -\Psi\left(x,y\right)$, structural SVM functional is the following one

$$\begin{gathered}\|w\|^2 + \frac{C}{l}\sum_{i=1}^{l}\max_{y\in Y}\big(\ \Delta\left(y_i,y\right) + \\ \left\langle w, \Psi\left(x_i,y\right) - \Psi\left(x_i,y_i\right)\right\rangle\ \big)\end{gathered} \tag{13}$$

We used software by T. Joachims [24], which is based on the following quadratic programming reformulation:

$$\begin{gathered}\|w\|^2 + C\textstyle\sum_{i=1}^{\ell}\xi_n \to \min\limits_{w,\xi}\\ \left\langle w, \Psi\left(x_i,y_i\right) - \Psi\left(x_i,y\right)\right\rangle \geq \Delta(y_i,y) - \xi_i,\\ n = 1,\ldots,\ell, \quad \forall y\in\mathcal{Y},\\ \xi_i \geq 0, \qquad i = 1,\ldots,\ell.\end{gathered}$$

Very roughly, this functional can be interpreted as follows. For optimal parameter w we want $y_i \approx \arg\max\limits_{y}\left\langle w, \Psi\left(x_i,y\right)\right\rangle$, and therefore

$$\left\langle w, \Psi\left(x_i,y_i\right) - \Psi\left(x_i,y\right)\right\rangle \geq 0$$

if $\Delta\left(y_i,y\right)$ is large(greater than resulting slack ξ_i), and in the neighborhood of y_i (i.e. when $\Delta\left(y_i,y\right) \leq \xi_i$ — we can call such neighborhoods as "uncertainty neighborhoods") the difference $\left\langle w, \Psi\left(x_i,y_i\right) - \Psi\left(x_i,y\right)\right\rangle$ can be negative. The goal is to lower the sum of radiuses of "uncertainty neighborhoods" plus regularization on $\|w\|^2$.

5 Experiments and discussion

Experimental data were taken from the PDB database. From 81756 available protein structures, we extracted 7631 all-alpha proteins. This list was filtered by PISCES (A Protein Sequence Culling Server) with requirements of sequence identity to be less than 25%, of resolution to be less than 3 , and of R-factor to be less than 0.3. The resulting sample contained 908 proteins.

Each protein in dataset was represented as a pair of words, the first word is its amino acid sequence, and the second word is a sequence if ϕ,ψ angles of amino acids that were discretized with step of 20 degrees, i.e. $\phi_{discr} = [\phi/20], \psi_{discr} = [\psi/20]$. Therefore, the second word was in an alphabet of 18 symbols. A set of contact triples $\mathfrak{A}$ that is the basis of our model was chosen by a simple procedure: for each triple (A_1, A_2, α), where A_1, A_2 are amino acids, we counted the number of times it occurs in our database (i.e. segment $A_1 ... A_2$ in the amino acid sequence corresponds to segment α in the second word) under the condition that C_α atoms of amino acids A_1, A_2 are located less than 10 from each other according to the corresponding PDB file; then we took all triples that were counted more than ten times.

We have three variants of loss functions. Since ϕ,ψ plane has torical structure, we define $H_\phi\left(\phi_1,\phi_2\right) = \min\left\{\left|\phi_1-\phi_2\right|, 360-\left|\phi_1-\phi_2\right|\right\}$. H_ψ is defined analogously. For discretized ϕ,ψ we define $H_\phi^{discr}\left(\phi_1,\phi_2\right) = \min\left\{\left|\phi_1-\phi_2\right|, 18-\left|\phi_1-\phi_2\right|\right\}$. Then,

$$\begin{gathered}H\left(\{y_i\}_{i=\overline{1,2n}}, \{y'_i\}_{i=\overline{1,2n}}\right) \triangleq \\ \sum_{i=1}^{n} H_\phi\left(y_{2i-1}, y'_{2i-1}\right) + H_\psi\left(y_{2i}, y'_{2i}\right)\end{gathered} \tag{14}$$

H^{discr} is defined analogously. During learning we can use only the discrete version, but when accessing accuracy on the test set we will use the continuous version of the loss function H.

The second criterion is **MDA**, and it is commonly used for accessing the accuracy of dihedral angles prediction. In a continuous case, it is defined as a percentage of amino acids in a sequence that belongs to any continuous segment of length not less than 8 for which predicted dihedral angles do not differ from real by more than 120 degrees. I.e.:

$$\mathbf{MDA}\left(\{y_i\}_{i=\overline{1,2n}}, \{y'_i\}_{i=\overline{1,2n}}\right) \triangleq \frac{|\Omega|}{n} \tag{15}$$

where $\Omega = \{i|\exists k,l : k \leq i \leq l, l-k \geq 7, \forall s = \overline{k,l} |y_{2s-1} - y'_{2s-1}| \leq 120, |y_{2s} - y'_{2s}| \leq 120\}$. The discrete version of **MDA** can be defined by changing 120 to 6; we denote it $\mathbf{MDA^{discr}}$.

Recall that structural SVM calls a procedure to maximize $\Delta\left(y,y'\right) + \left\langle w, \Psi\left(x_i,y\right)\right\rangle$ where Δ is a loss function, and this exactly the place where inference algorithm is needed for learning. Since $H\left(y,y'\right)$ can be understood as a unary part of our optimized functional, then our inference algorithm can be applied straightforwardly. On the

Table 1: Experimental results

Loss function	$\mathbf{H}_\phi$ °C	$\mathbf{H}_\psi$ °C	$\mathbf{MDA}, \%$	$\mathbf{C}$
$\mathbf{H}^{\mathrm{discr}}$	22.7	47.9	55.0	64.0
$\mathbf{MDA}^{\mathrm{discr}}$	22.8	48.3	56.5	4.0
HMMSTR[4]	-	-	57.1	-
PHD[17]	-	-	48.0	-

contrary, $\mathbf{MDA}$ has a very "global" structure. Therefore, instead of maximizing the previous sum, we used a heuristic that reduces to excluding the loss part.

The sample set was divided into three subsets: a training sample, a holdout sample (for fitting some parameters) and a test sample. The training sample was used to train parameters with fixed C, the holdout sample was used to choose C and the resulting prediction accuracy was accessed on the test set. The results of experiments are given in Table 1. Rows show results for fixed loss functions and columns show exact attained values for measures of accuracy on a test set.

In the work of Bystroff & al. [4], whose method has a lot of common with ours, $\mathbf{MDA}$ value attained 57% on the training set consisting of 618 randomly selected proteins. Such an accuracy became possible to achieve by using a special library of structural motives, called I-sites, that was generated by clustering of all motives. Also, instead of using amino acid symbols, each position in a protein sequence was associated with the amino acids profile, which significantly improved overall accuracy. The highest MDA achieved by the first generation protein structure annotations predictors is approximately 60%. In the second generation of predictors 64% accuracy was reported. The third generation of predictors, based on Deep Learning algorithms, predict secondary structure at over 70% accuracy [23]. Thus, so far our method reproduces the accuracy of the first-generation software.

Our algorithm can be improved by adding structural information for amino acids, like solvent accessibility or secondary structure. We generated a set of patterns by a very simple procedure, and two factors were crucial in defining a set of patterns: we wanted to choose a set that would map all local conformational structures with good scale, at the same time the cardinality of this set could not be too high to overfit or to become computationally intractable. Both these problems could be addressed by introducing long "superpatterns", by which we mean clusters of patterns that have the same weights in the energy model. In such an approach, the problem of clusterization of pattern set appears that should be addressed before learning of weights starts. It is also important to note that our inference algorithm can easily adapt to formulation with long superpatterns by just considering a superpattern as a set of patterns for which one weight is attributed. If elements of superpatterns will map all patterns with fixed length from a discretized database, then the number of such patterns will grow proportionally to NLP, where N is the number of proteins in the database, L is the average length of a protein, and P is a pattern length.

6 Conclusions and future work

The goal of our paper was to show perspectives of the pattern-based conditional random field in bioinformatics applications. Hidden Markov Models, currently very popular in protein folding prediction, gene prediction, and other problems with sequence (tagging) labeling flavor, can be considered as a precursor of pattern-based CRFs. By definition, the current state of HMM is probabilistically determined by the previous state. When there are high distance interactions between labels on the sequence, a notion of HMM state should include a lot of context structural information. Instead of complicating the notion of a state, we propose to consider long label sequences as analogs of states. But for our formulation, there is a lack of research dedicated to problems of learning and inferencing in pattern-based models. We need more efficient algorithms for the inference part of the problem since in some applications, the length of patterns could be high (for example, the typical helix region of a protein can be 20 amino acids long, then only ω, ϕ, ψ angles sequence of such regions can be of length 60 or longer, not including other structural information) and the number of patterns could be very large (in superpatterns approach it could be comparable with the size of the learning database). We developed an efficient algorithm for this problem based on dynamic programming with a preprocessing step to tune the parameters of an algorithm for concrete patterns set. On the learning part, we used the structural SVM technique.

An important issue to be addressed in future work is how our approach could be developed in the context of bioinformatics applications, such as protein folding and gene prediction. Learning weights is the hardest part of such problems, which requires a thorough analysis of maximum likelihood and structured risk minimization parameter learning. Besides, our exact definition of pattern-based CRF can be easily generalized both in the direction of including superpatterns and in the direction of defining various areas of dependence for weights to be learned.

Acknowledgement

This work was supported from the Nazarbayev University collaborative research grant project OPCRP2022002.

References

[1] Yasemin Altun, Ioannis Tsochantaridis, and Thomas Hofmann. Hidden markov support vector machines. In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, page 3–10, 2003.

[2] C B Anfinsen. The formation and stabilization of protein structure. *Biochemical Journal*, 128(4):737–749, 07 1972. https://doi.org/10.1042/bj1280737.

[3] Zhenisbek Assylbekov and Rustem Takhanov. Reusing weights in subword-aware neural language models. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1413–1423, New Orleans, Louisiana, June 2018. https://doi.org/10.18653/v1/N18-1128.

[4] Christopher Bystroff, Vesteinn Thorsson, and David Baker. Hmmstr: a hidden markov model for local sequence-structure correlations in proteins11edited by j. thornton. *Journal of Molecular Biology*, 301(1):173 – 190, 2000. https://doi.org/10.1006/jmbi.2000.3837.

[5] Peter Y. Chou and Gerald D. Fasman. Prediction of protein conformation. *Biochemistry*, 13(2):222–245, 1974. PMID: 4358940. https://doi.org/10.1021/bi00699a002.

[6] A.G. de Brevern, C. Etchebest, and S. Hazout. Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. *Proteins: Structure, Function, and Bioinformatics*, 41(3):271–287, 2000. https://doi.org/10.1002/1097-0134(20001115)41:3<271::AID-PROT10>3.0.CO;2-Z.

[7] R. Fletcher. *Newton-Like Methods*, chapter 3, pages 44–79. John Wiley and Sons, Ltd, 2000. https://doi.org/10.1002/9781118723203.ch3.

[8] J. Garnier, D.J. Osguthorpe, and B. Robson. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *Journal of Molecular Biology*, 120(1):97 – 120, 1978. https://doi.org/10.1016/0022-2836(78)90297-8.

[9] Blaise Gassend, Charles O'Donnell, William Thies, Andrew Lee, Marten van Dijk, and Srinivas Devadas. Learning biophysically-motivated parameters for alpha helix prediction. *BMC bioinformatics*, 8 Suppl 5:S3, 02 2007. https://doi.org/10.1186/1471-2105-8-S5-S3.

[10] Misha Gromov. Crystals, proteins, stability and isoperimetry. *Bulletin of the American Mathematical Society*, 48(2):229–257, 2011. https://doi.org/10.1090/S0273-0979-2010-01319-7.

[11] DT Jones. Protein secondary structure prediction based on position-specific scoring matrices. *Journal of molecular biology*, 292(2):195—202, September 1999. https://doi.org/10.1006/jmbi.1999.3091.

[12] Vladimir Kolmogorov, Michal Rolínek, and Rustem Takhanov. Effectiveness of structural restrictions for hybrid csps. In *Algorithms and Computation - 26th International Symposium, ISAAC 2015, Proceedings*, LNCS, pages 566–577, Germany, January 2015. Springer Verlag. https://doi.org/10.1007/978-3-662-48971-0_48.

[13] Petros Kountouris, Petros Kountouris, and Jonathan D. Hirst. Prediction of backbone dihedral angles and protein secondary structure using support vector machines. *BMC Bioinformatics*, 10(2):437, 2009. https://doi.org/10.1186/1471-2105-10-437.

[14] Jooyoung Lee, Sitao Wu, and Yang Zhang. *Ab Initio Protein Structure Prediction*, pages 3–25. Springer Netherlands, Dordrecht, 2009. https://doi.org/10.1007/978-1-4020-9058-5_1.

[15] Sebastian Nowozin and Christoph H. Lampert. Structured learning and prediction in computer vision. *Foundations and Trends® in Computer Graphics and Vision*, 6(3–4):185–365, 2011. https://doi.org/10.1561/0600000033.

[16] Xian Qian, Xiaoqian Jiang, Qi Zhang, Xuanjing Huang, and Lide Wu. Sparse higher order conditional random fields for improved sequence labeling. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 849–856, New York, NY, USA, 2009. Association for Computing Machinery. https://doi.org/10.1145/1553374.1553483.

[17] Burkhard Rost. Better 1d predictions by experts with machines. *Proteins: Structure, Function, and Bioinformatics*, 29(S1):192–197, 1997. https://doi.org/10.1002/(SICI)1097-0134(1997)1+<192::AID-PROT25>3.0.CO;2-I.

[18] Rustem Takhanov. Hybrid vcsps with crisp and valued conservative templates. In *28th International Symposium on Algorithms and Computation, ISAAC 2017*, volume 92, Germany, December 2017. https://doi.org/10.4230/LIPIcs.ISAAC.2017.65.

[19] Rustem Takhanov. Searching for an algebra on csp solutions, 2017. arXiv:1708.08292.

[20] Rustem Takhanov and Zhenisbek Assylbekov. Patterns versus characters in subword-aware neural language modeling. In *Neural Information Processing*, pages 157–166, Cham, 2017. https://doi.org/10.1007/978-3-319-70096-0_17.

[21] Rustem Takhanov and Vladimir Kolmogorov. Inference algorithms for pattern-based crfs on sequence data. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 145–153, Atlanta, Georgia, USA, 17–19 Jun 2013. URL: https://proceedings.mlr.press/v28/takhanov13.html.

[22] Rustem Takhanov and Vladimir Kolmogorov. Combining pattern-based crfs and weighted context-free grammars. *Intell. Data Anal.*, 26(1):257–272, 2022. https://doi.org/10.3233/IDA-205623.

[23] Mirko Torrisi, Gianluca Pollastri, and Quan Le. Deep learning methods in protein structure prediction. *Computational and Structural Biotechnology Journal*, 18:1301–1310, 2020. https://doi.org/10.1016/j.csbj.2019.12.011.

[24] Ioannis Tsochantaridis, Thomas Hofmann, Thorsten Joachims, and Yasemin Altun. Support vector machine learning for interdependent and structured output spaces. In *Proceedings of the Twenty-First International Conference on Machine Learning*, page 104, New York, NY, USA, 2004. https://doi.org/10.1145/1015330.1015341.

[25] Yuedong Yang, Jianzhao Gao, Jihua Wang, Rhys Heffernan, Jack Hanson, Kuldip Paliwal, and Yaoqi Zhou. Sixty-five years of the long march in protein secondary structure prediction: the final stretch? *Briefings in Bioinformatics*, 19(3):482–494, 12 2016. https://doi.org/10.1093/bib/bbw129.

[26] Nan Ye, Wee Lee, Hai Chieu, and Dan Wu. Conditional random fields with high-order features for sequence labeling. In *Advances in Neural Information Processing Systems*, volume 22, 2009. URL: https://proceedings.neurips.cc/paper/2009/file/94f6d7e04a4d452035300f18b984988c-Paper.pdf.