

Detecting Correct Structure of Protein Fragments in a Database of Motifs

Szymon Nowakowski,¹ Jerzy Tiuryn²

Keywords: structural motifs, mixture of PSSMs, mixture of Dirichlets, dependency

1 Introduction.

Fragment-based methods of protein structure prediction require identification of a local 3D structure for a fragment of a query sequence. We propose two techniques of identifying sequence-structure relationships between target protein and possible motifs. *Beta method* is almost as successful as *mixture of Dirichlets method*, but it is many times (up to 35 times) faster. *MAP estimation for two PSSMs* models dependencies between columns and is almost 1.5 times more successful than other methods, but it is very computationally expensive.

2 Methods.

We constructed our experiment by preparing a library of 335 motifs—structurally aligned fragment sets ([3]), taken from local neighborhoods from a core of proteins represented in ASTRAL 1.63 ([1]) and having less than 40% sequence identity to one another. There were 335 motifs, as we randomly chose only a subset of all (1753) motifs for our experiment.

Fragment sets were aligned structurally and then a sequential information was extracted from every alignment. We evaluated the sequence-structure matches by examining their compatibility in purely sequential 1-1 correspondence of the amino acid positions on a test set in a 5-fold cross validation test.

We used the following match scoring procedure: let's fix an alignment \mathcal{A} and a query sequence S . Let the query sequence be $S = s_1 s_2 s_3 \dots s_l$, let alignment \mathcal{A} have also length l . Suppose that \mathcal{A} is described by K PSSMs (profiles) and K weights, q_1, \dots, q_K . Every PSSM P^k is a matrix of size $20 \times l$: $P^k = (p_{ij}^k)$, where $k \leq K$ is a PSSM number. A match is scored as $\mathcal{M}(S, \mathcal{A}) = \sum_{k=1}^K q_k \cdot p_{s_1 1}^k \cdot p_{s_2 2}^k \cdot \dots \cdot p_{s_l l}^k$. When $K \geq 2$ dependencies are introduced into a scoring model. We evaluated the following profile estimation methods:

1. *Beta method*. Columns of multiple sequence alignments of fragment sets are first clustered by the similarity of their amino acid distributions. A profile correction method based on Bayesian statistics with prior distributions $Beta(a_x^i, b_x^i)$ for the frequency of each amino acid x is used if a column is in the i -th cluster. Let us fix an i -th column in a fragment set. For an observed frequency $\frac{n_x}{n}$ of amino acid x , an estimator of the probability of x appearing in the column is $\frac{n_x + a_x^i}{n + a_x^i + b_x^i}$. We use this estimator as the new frequency. After normalization we obtain a new profile for each alignment.
2. *Mixture of Dirichlets method*. A method described in ([2], [5]) is used. MPE (Mean Posterior Estimator) is used with a prior being a mixture of Dirichlet distributions.
3. *MAP estimation for two PSSMs*. Instead of MPE estimation, it is proposed in Nowakowski's PhD thesis ([4]) to use a MAP (Maximum a Posteriori) estimation in case of

¹Institute of Informatics, Warsaw University, Warszawa, Poland
e-mail: s.nowakowski@mimuw.edu.pl

²Institute of Informatics, Warsaw University, Warszawa, Poland

a PSSM mixture with a prior being a mixture of Dirichlets. The estimation is done with an extensive search over a probabilistic space followed by a gradient descent as a part of EM (Expectation Maximization) algorithm. As a result we get two PSSMs describing every fragment set, together with the weights q_1, q_2 of these PSSMs.

3 Results and Conclusions.

Our tests were conducted as follows: every fragment was tested exactly once against 335 fragment sets in one of 5 runs of the 5-fold cross validation procedure; in every run, for a given test fragment, each of the 335 fragment sets was scored according to a selected profile estimation method. Predictions consisted of m fragment sets with the highest score. Predictions were judged as successful if the correct fragment set (i.e. the one, from which the test fragment was removed in this run), was among the m chosen fragment sets.

| Estimation Method | K (no. of PSSMs) | Parameter m | | | |
|------------------------------|--------------------|---------------|----|----|----|
| | | 1 | 5 | 10 | 20 |
| Beta method | 1 | 20 | 42 | 54 | 66 |
| Mixture of Dirichlets method | 1 | 21 | 42 | 54 | 65 |
| MAP estimation for 2 PSSMs | 2 | 30 | 52 | 62 | 72 |

Table 1: Results of recognition tests: percentage of successful predictions.

Table 1 summarizes the results: it shows the percentage of successful predictions depending on the value of parameter m and on the profile estimation method used. The most successful is MAP estimation for two PSSMs: in case of $m = 1$ its success rate is almost 1.5 times greater than that of the most successful method without dependencies. As seen in Table 1, our method can identify correct motif in 30% of cases.

4 Acknowledgments.

We want to thank Krzysztof Fidelis and his group from BBRP in LLNL (Livermore, USA) for supplying us with the fragment sets and for valuable discussions.

The results were partly obtained with the use of computer resources of ICM UW, Warszawa, Poland. This work was supported by the Polish KBN grant 3 T11F 006 27.

References

- [1] Brenner, S.E., Koehl P. and Levitt M. 2000. The ASTRAL Compendium for Sequence and Structure Analysis. *Nucleic Acids Research* 28:254–256.
- [2] Brown M. P., Hughey, R., Krogh, A., Mian, I. S., Sjölander, K. and Haussler, D. 1993. Using Dirichlet Mixture Priors to Derive Hidden Markov Models for Protein Families. In Hunter, L., Searls, D., and Shavlik J. editors, *ISMB-93*, Menlo Park, CA: AAAI/MIT Press. pp. 47–55.
- [3] Kryshchak, A. and Fidelis, K. 2004. Local Descriptors of Protein Structure. Part I. General Approach and Classification of Local 3D Regions in Proteins (*in preparation*).
- [4] Nowakowski, S. Estimating Probability Distributions of a Sequence Occuring in an Alignment and Its Use in 3D Protein Structure Prediction (*in Polish*). PhD thesis (*in preparation*), Insitute of Informatics, Warsaw University, Warszawa, Poland.
- [5] Sjölander, K., Karplus, K., Brown, M., Hughey, R., Krogh, A., Mian, I. S. and Haussler, D. 1996. Dirichlet Mixtures: a Method for Improved Detection of Weak but Significant Protein Sequence Homology, *Computer Applications in Biosciences* 12:327–345.