

Introducing Dependencies into Alignment Analysis and Its Use for Local Structure Prediction in Proteins

Szymon Nowakowski^{1,2}, Krzysztof Fidelis³, and Jerzy Tiuryn¹

¹ Institute of Informatics, Warsaw University
Banacha 2, 02-097 Warszawa, Poland

² s.nowakowski@mimuw.edu.pl

³ Genome Center, University of California, Davis
Genome and Biomedical Sciences Facility

451 East Health Sciences Drive, Davis, CA 95616, USA

Abstract. In this paper we explore several techniques of analysing sequence alignments. Their main idea is to generalize an alignment by means of a probability distribution. The *Dirichlet mixture method* is used as a reference to assess new techniques. They are compared based on a cross validation test with both synthetic and real data: we use them to identify sequence-structure relationships between target protein and possible local motifs. We show that the *Beta method* is almost as successful as the reference method, but it is much faster (up to 17 times). *MAP (Maximum a Posteriori) estimation for two PSSMs (Position Specific Score Matrices)* introduces dependencies between columns of an alignment. It is shown in our experiments to be much more successful than the reference method, but it is very computationally expensive. To this end we developed its parallel implementation.

1 Introduction

Motif discovery in the case of DNA or protein sequences has a wide range of applications in modern molecular biology: from modeling mechanisms of transcriptional regulation [4, 11] to prediction of protein structure [2]. In this work we address the latter problem. We developed a probabilistic approach of generalizing sequence alignments representing structural motifs of local neighborhoods in proteins. We show that we can identify a correct motif of a given protein fragment accurately using only sequence of the query fragment.

This problem has been extensively studied during the last few years. The simplest estimator of seeing an amino acid in an alignment, called *Maximum Likelihood* estimator [8], which only counts occurrences of amino acids, is of no use for alignments consisting of too few sequences, as it may happen that an amino acid *is not seen* in the alignment at all, but it *would be seen*, if the number of aligned sequences were greater. To this end various methods introducing prior knowledge were proposed. These methods range from the *zero-offset*

method [10], through methods based on substitution matrices [3] or amino acid feature alphabets [13], to the most advanced *Dirichlet mixture method* [6, 12] for which a mixture of Dirichlet distributions is supplied as prior knowledge. The *pseudocount method* [14] is a special kind of the Dirichlet mixture method, where only one Dirichlet distribution is used. It is shown in [10] that, when the columns of an alignment are independent, the Dirichlet mixture method is close to the theoretical optimum.

Modeling dependencies between columns in DNA sequence alignments has been recently studied in [1, 4, 7]. In [7] dependencies are modeled only between adjacent columns with the use of Hidden Markov Models. Methods exist to model dependencies between columns not adjacent in an alignment. Authors of [4] analyse a number of such methods, one of them being the *mixture of PSSMs (Position Specific Score Matrices)* method. They use the pseudocount method to introduce prior knowledge (modeled by a single Dirichlet distribution). We propose a new method of estimating distributions from alignments that is more suitable for protein sequence alignments. It models column dependencies with a mixture of PSSMs and uses a mixture of many Dirichlet distributions as prior knowledge. We discuss advantages of this choice in Section 5.

We performed two experiments comparing different techniques of generalizing alignments. The first one was performed with synthetic data, generated from known probability distributions, the second with real data from a database of protein motifs. In these experiments we compared our two new techniques with the Dirichlet mixture method. First of the techniques we propose is *MAP (Maximum a Posteriori) estimation for two PSSMs*. This technique can model column dependencies. We show that it gives much better results than the reference method. Second, the *Beta method*, gives results comparable with the Dirichlet mixture method, but is up to 17 times faster.

2 Methods

Results of our experiments were obtained by a cross validation test done with both synthetic and real data. We performed 3-fold cross validation with synthetic data and 5-fold cross validation with real data. Let us describe the *F-fold cross validation test*, where F is a positive integer, usually between 3 and 10. The dataset for the test consists of a number of alignments. The test itself is performed as follows: sequences from every alignment are randomly divided into F subsets. Each of them is treated as a test set T in one of F runs of the cross validation test. At the same time the remaining $F - 1$ subsets are treated as a training set (profiles are estimated from the union of $F - 1$ subsets, according to a selected profile estimation method, see below for details).

The log-probability is computed for every sequence from T and the alignment from which that sequence was removed. For each alignment we then compute its mean log-probability value for a given estimation method, averaged over all sequences in all runs of the cross validation procedure. Let us fix the alignment \mathcal{A} . To address the statistical significance of the difference in mean log-probability

values for \mathcal{A} between two methods, the paired t-test is performed. Following [4], we call one method *better* than the other on \mathcal{A} when the difference in mean log-probability values for \mathcal{A} is positive, and *significantly better* when the associated paired t-test p-value is below the 0.05 threshold. It is called *worse* or *significantly worse* when the other method is, respectively, better or significantly better.

Additionally, in the case of synthetic data, which consists of alignments with the same number of columns, during the cross validation procedure the prediction test is performed as follows: for a given sequence from T , every alignment is scored as described below. The prediction is judged as successful if the correct alignment (i.e. the one, from which the test sequence was removed) is the one with the highest score.

Let us define a notion of PSSM. By PSSM we understand a sequence profile. Formally, *PSSM* of length l is a matrix (p_{ij}) of size $20 \times l$. We estimate PSSMs from alignments using one of the techniques described below. All those techniques accept a mixture of Dirichlet distributions as prior knowledge.

1. *Dirichlet mixture method.* This method, described in detail in [6, 12], is used as a reference method. In short, PME (Posterior Mean Estimator) is used with a prior being a mixture of Dirichlet distributions.
2. *Beta method.* Columns of all sequence alignments are first clustered by the similarity of their amino acid distributions. The clusters are built around center points computed from the supplied prior. After construction of the clusters the prior is discarded and new and much simpler priors are estimated for every cluster. Let us refer to a profile obtained from a column with the use of the Maximum Likelihood estimator as an ML-profile. For every amino acid x in the i -th cluster the new prior is the $Beta(a_x^i, b_x^i)$ distribution. The values of a_x^i, b_x^i are estimated from all ML-profiles in the i -th cluster.

Let us fix a column c in an alignment. Suppose that c is in the i -th cluster. Let us consider any amino acid x . Suppose that x has the observed frequency $\frac{n_x}{n}$ in the ML-profile of c . A posterior estimator of the probability of x appearing in c is $\frac{n_x + a_x^i}{n + a_x^i + b_x^i}$. We repeat that for all 20 amino acids. After normalization by $N = \sum_x \frac{n_x + a_x^i}{n + a_x^i + b_x^i}$ we obtain a new profile for c .

3. *MAP estimation for two PSSMs.* Instead of PME estimation (as in the case of the Dirichlet mixture method) we use MAP estimation to obtain a mixture of two PSSMs together with the weights q_1, q_2 of these PSSMs. This representation takes into account column dependencies.

The implementation is parallel since the estimation is computationally very demanding. MPI environment is used. We use a master-slave paradigm in which one of the processes distributes the tasks and collects the results, while the rest of the processes perform calculations. The calculations are done in two phases. In the first, called the *search phase*, an extensive sampling of the probability space is done to find good starting points for the second phase, which is called the *maximizing phase*. In the second phase a gradient descent as a part of EM (Expectation Maximization) algorithm is performed to locally maximize our goal function.

We use the following scoring procedure. Let us fix an alignment \mathcal{A} of length l and a query sequence $S = s_1 s_2 s_3 \dots s_l$. \mathcal{A} is described by $P^{(1)}, \dots, P^{(K)}$, PSSMs estimated according to one of the techniques described above, and by K positive weights of these PSSMs, q_1, \dots, q_K , such that $\sum_{i=1}^K q_i = 1$. The value of K depends on the estimation technique and it equals 1 or 2 in this work. Every PSSM $P^{(k)} = (p_{ij}^{(k)})$ is a matrix of size $20 \times l$. The score of \mathcal{A} is $\mathcal{M}(S, \mathcal{A}) = \sum_{k=1}^K q_k \cdot p_{s_1 1}^{(k)} \cdot p_{s_2 2}^{(k)} \cdot \dots \cdot p_{s_l l}^{(k)}$. Logarithm of this score is used as the log-probability value for the alignment \mathcal{A} in the cross validation procedure.

3 Experiment with Synthetic Data

We performed two tests with synthetic data: in the first one, alignments consisted of 200 sequences and in the second test of 600 sequences. This allowed us to assess the impact of the number of sequences in an alignment on the estimation accuracy.

For both tests we used the same prior distribution as in the case of the real data (i.e. a mixture of 54 Dirichlet distributions). We believe that it made our synthetic experiment more realistic. We generated 300 alignments of length 30 for both tests. The procedure to generate a single alignment was as follows: first, generate three PSSMs with 30 columns distributed according to the prior; then, from this mixture of PSSMs, generate 200 or 600 (depending on the test) sequences of length 30. This procedure was repeated 300 times to generate 300 alignments.

3-fold cross validation was performed as it was described in Section 2. Thus the estimation was done in fact on $\frac{2}{3}$ of the alignment. For the estimation we used the same prior that had been used to generate alignments.

In this experiment we tested the MAP estimation for two PSSMs and the Beta method. Their comparison to the reference method, the Dirichlet mixture method, presented in Table 1, shows the number of alignments, for which each method had higher mean log-probability value than that of the reference method. Table 1 also shows the percentage of successful predictions (evaluated as described in Section 2) depending on the profile estimation method used.

As seen in Table 1, for alignments comprising relatively small number of sequences the difference in prediction accuracy between the Beta method and the Dirichlet mixture method was larger. This is due to a much simpler character of the Beta method. The Beta method also had significantly worse mean log-probability values in almost all cases. In addition we can see that in the case of insufficient data the MAP estimation for two PSSMs performed very poorly. It had lower mean log-probability values in all cases and in the prediction accuracy test it was also much worse than both methods without dependencies. This is caused by a need to estimate a larger number of parameters.

However, for alignments with greater number of aligned sequences the Beta method performed almost as well as the Dirichlet mixture method (when we consider prediction accuracy) and MAP estimation for two PSSMs proved to be much more successful, having both much better prediction accuracy and higher

Table 1. Results of tests with synthetic data: the number of alignments with higher mean log-probability value than that of the reference method (i.e. better results), the number of significantly better results and the number of significantly worse results. The reference method row is presented in bold face. Column 6 shows the percentage of successful predictions. Columns 7 and 8 show the number of sequences in each alignment and an average number of sequences in the training set used for estimation in the 3-fold cross validation procedure

Estimation method	PSSMs	Better	Sig. better	Sig. worse	Succ. pred.	Seqs	Tr. seqs
Dirichlet mixture	1	0	0	0	18%	200	133
Beta	1	0	0	293	16%	200	133
MAP estimation	2	0	0	300	11%	200	133
Dirichlet mixture	1	0	0	0	22%	600	400
Beta	1	63	5	128	22%	600	400
MAP estimation	2	285	274	7	29%	600	400

mean log-probability values in almost all cases, most of them being significantly better.

We can see that it is enough to estimate *two* PSSMs instead of one to greatly increase success rate, in spite of the fact, that data was generated from a mixture of *three* PSSMs.

4 Experiment with Real Protein Motifs

Our experiment was performed on a list of 117 sets of structurally aligned protein fragments. They were taken from proteins represented in ASTRAL 1.63 [5] and having less than 40% sequence identity to one another. The fragments were composed of several contiguous subfragments forming a local 3D neighborhood in a protein core, as described in [9].

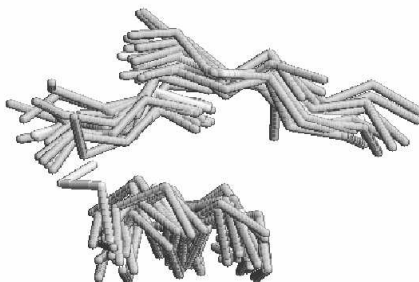


Fig. 1. Example of a set of structurally aligned fragments. There are 3 contiguous subfragments in every fragment in this set

Fragments were structurally aligned based on their 3D structure similarity measured by RMSD (Root Mean Square Deviation). After the structural alignment had been constructed, a corresponding sequence alignment was obtained for each set of fragments. Figure 1 shows an example of such a structurally aligned set. The sets (and consequently alignments) contained from 508 to 6996 fragments (sequences), as we rejected alignments with less than 500 sequences (too few sequences to perform a reliable MAP estimation for two PSSMs).

To gather more sequential information for estimation methods, sequences of the subfragments were extended by 4 amino acids on both ends. Extensions in the set need not share common structure, as is the case for the original fragments.

Table 2 summarizes the results: it shows the number of alignments for which the considered method was better (significantly better, significantly worse) than the reference method (evaluated as described in Section 2) in the 5-fold cross validation procedure with a mixture of 54 Dirichlet distributions used as a prior.

Table 2. Results of the test: the number of alignments with higher mean log-probability value as compared to the reference method (i.e. better results), the number of significantly better results and the number of significantly worse results. The reference method row is presented in bold face. Last column shows computation time on a single processor on all data (without cross validation, averaged over 3 runs)

Estimation method	PSSMs	Better	Sig. better	Sig. worse	Comp. time
Dirichlet mixture	1	0	0	0	34 sec.
Beta	1	64	32	20	2 sec.
MAP estimation	2	83	77	26	no data

As seen in Table 2, the Beta method was more successful than the Dirichlet mixture method in 64 of 117 cases (but not significantly in half of them), while the estimation time in the former method was much shorter than in the case of the latter. To assess the speed of estimation, we performed an additional test (repeated 3 times) in which no cross validation was performed and both methods were used to estimate profiles from the data. The test was performed on a computer with Pentium4 2.80GHz processor and 512MB RAM. The estimation took 2 sec., 2 sec. and 2 sec. for the Beta method and 34 sec., 32 sec. and 37 sec. for the Dirichlet mixture method; on average the Beta method was 17 times faster.

Including dependencies in our method with two PSSMs greatly increased accuracy of estimation, making it better in 83 of 117 cases, most of them being significantly better. It was significantly worse only in 26 cases.

MAP estimation, although the most successful, is also computationally most demanding: the test described took about a week of computation time on a cluster of 16 two-processor computers (AMD Opteron 2GHz, 2GB RAM). To assess the impact of the number of processes on the computation time, additional test was performed on that cluster. Table 3 summarizes the results of running the

Table 3. Results of the scaling test for the MAP estimations for two PSSMs: the computation time as a function of the number of processes. The second column presents the number of processes doing the calculations. Columns 3 and 4 present the real computation time while two last columns present the computation time predicted by dividing the times from the row in bold face by the number of processes performing the calculations. The values agree very well

Processes	Calc. processes	Computation time		Predicted time	
		Search phase	Max. phase	Search phase	Max. phase
2	1	88704 sec.	961640 sec.	-	-
4	3	29568 sec.	332843 sec.	29568 sec.	320547 sec.
8	7	12833 sec.	130843 sec.	12672 sec.	137377 sec.
16	15	5989 sec.	63409 sec.	5914 sec.	64109 sec.
32	31	2905 sec.	34648 sec.	2861 sec.	31021 sec.

computations without cross validation on a subset of alignments which consisted of less than 1000 sequences (48 such alignments were present among all 117 alignments under consideration). The computations were run with 2, 4, 8, 16 and 32 processes. One of these processes, as described in Section 2, distributed the tasks and collected the results, the rest of the processes performed the calculations. There were, respectively, 1, 3, 7, 15 and 31 processes doing the calculations. As seen in Table 3, the computations scale very well. The computation time in each case can be well predicted by dividing the time taken by one process doing the calculations by the number of calculating processes.

5 Conclusions

The success of MAP estimation for two PSSMs not only in the experiment with synthetic data, but also with real data, is caused, we believe, by three factors:

1. The way to include column dependencies, we introduce in our model (a mixture of PSSMs), has a strong biological background: it models the fact, that the structure of a protein motif can be stabilized in more than one way.
2. The dependent positions, which make the stabilization possible, are located on different subfragments, not adjacent in an alignment. When the models with dependencies between nonadjacent columns are considered, mixtures of PSSMs have relatively few parameters [4], which makes estimation more reliable.
3. The prior knowledge we use in our model (a mixture of many Dirichlet distributions) makes it possible to model many amino acid contexts in columns, in contrast to the pseudocount method used in [4], which models only one context: the background. Thus we can model hydrophobic columns, columns with big amino acids, columns with positive amino acids, etc.

Comparison of the results on synthetic and real data shows that when performing estimation with dependencies it is very important to include only align-

ments with enough sequences to make the estimation reliable. The results can be very poor when the dependencies are considered but not enough examples are provided for the estimation procedure.

6 Acknowledgments

This work was supported by Polish KBN grant 3 T11F 006 27. Our results were obtained with the use of computer resources of ICM Warsaw University, Poland.

References

1. Agarwal, P., Bafna, V.: Detecting Non-adjointing Correlations with Signals in DNA. in RECOMB'98 (1998) 2–8
2. Aloy, P., Stark, A., Hadley, C., Russell, R.B.: Predictions Without Templates: New Folds, Secondary Structure, and Contacts in CASP5. *Proteins: Struct. Funct. Genet.* **53** (2003) 436–456
3. Altschul, S.F.: Amino Acid Substitution Matrices from an Information Theoretic Perspective. *JMB* **219** (1991) 555–565
4. Barash, Y., Elidan, G., Friedman, N, Kaplan, T,: Modeling Dependencies in Protein-DNA Binding Sites, in RECOMB'03 (2003) 28–37
5. Brenner, S.E., Koehl P., Levitt M.: The ASTRAL Compendium for Sequence and Structure Analysis. *Nucleic Acids Research* **28** (2000) 254–256
6. Brown M. P., Hughey, R., Krogh, A., Mian, I. S., Sjölander, K., Haussler, D.: Using Dirichlet Mixture Priors to Derive Hidden Markov Models for Protein Families. In: Hunter, L., Searls, D., Shavlik J. (eds.) *ISMB-93*, Menlo Park, CA: AAAI/MIT Press. (1993) 47–55
7. Bulyk, M.L., Johnson, P.L., Church, G.M.: Nucleotides of Transcription Factor Binding Sites Exert Interdependent Effects On the Binding Affinities of Transcription Factors, *Nuc. Acids Res.*, **30** (2002) 1255–1261
8. Durbin, R., Eddy, S., Krogh, A, Mitchison, G.: Biological Sequence Analysis. Cambridge University Press (1998)
9. Hvidsten, R.H., Kryshtafovych, A., Komorowski, J., Fidelis, K.: A Novel Approach to Fold Recognition Using Sequence-Derived Properties From Sets of Structurally Similar Local Fragments of Proteins, *Bioinformatics*, **19** (2003) 81–91
10. Karplus, K.: Regularizers for Estimating Distributions of Amino Acids from Small Samples. Technical Report UCSC-CRL-95-11, University of California, Santa Cruz, CA, USA (1995), <ftp://ftp.cse.ucsc.edu/pub/tr/ucsc-crl-95-11.ps.Z>
11. Liu, X., Brutlag, D.L., Liu, J.S.: Bioprospector: Discovering Conserved DNA Motifs in Upstream Regulatory Regions of Co-expressed Genes. In *PSB'01* (2001)
12. Sjölander, K., Karplus, K., Brown, M., Hughey, R., Krogh, A., Mian, I. S., Haussler, D.: Dirichlet Mixtures: a Method for Improved Detection of Weak but Significant Protein Sequence Homology, *Computer Applications in Biosciences* **12** (1996) 327–345
13. Smith, R.F., Smith, T.F.: Automatic Generation of Primary Sequence Patterns from Sets of Related Protein Sequences. *PNAS* **87** (1990) 118–122
14. Tatusov, R.L., Altschul, S.F., Koonin, E.V.: Detection of Conserved Segments in Proteins: Iterative Scanning of Sequence Databases with Alignment Blocks. *PNAS* **91** (1994) 12091–12095