Journal of Biomedicine and Biotechnology • 2005:2 (2005) 124–131 • DOI: 10.1155/JBB.2005.124

RESEARCH ARTICLE

Contextual Multiple Sequence Alignment

Anna Gambin and Rafał Otto

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

Received 30 November 2003; revised 27 February 2004; accepted 12 March 2004

In a recently proposed contextual alignment model, efficient algorithms exist for global and local pairwise alignment of protein sequences. Preliminary results obtained for biological data are very promising. Our main motivation was to adopt the idea of context dependency to the multiple alignment setting. To this aim the relaxation of the model was developed (we call this new model *averaged contextual alignment*) and a new family of amino acids substitution matrices are constructed. In this paper we present a contextual multiple alignment algorithm and report the outcomes of experiments performed for the BAliBASE test set. The contextual approach turned out to give much better results for the set of sequences containing orphan genes.

INTRODUCTION

The multiple alignment of biological sequences has become an essential tool in molecular biology. It is used to find conserved regions and motifs in protein families, to detect the homology between new sequences and groups of sequences having an already known function and in a preliminary phase of protein structure prediction. Multiple alignment is also extensively used in molecular evolutionary analysis.

The various genome projects have provided the biologist with a great number of new protein sequences, and the rate of appearance of these data is steadily increasing. The development of an accurate and reliable multiple alignment program which is capable of handling many (often very divergent) sequences simultaneously is still of major importance.

The complexity of the problem does not allow to find the exact solution in a reasonable computational time [1]. Traditionally, the most popular heuristic approach has been the progressive alignment method [2].

In this paper we propose to explore new model for sequence alignment, in which the score for the substitution also depends on its neighborhood in the sequence. Such *contextual alignment model* has been proposed recently in [3] for the pairwise alignment problem. Preliminary results obtained for biological data by Gambin and Slonimski in [4] are very promising. To apply the contextual approach in the multiple alignment setting we have decided to relax slightly the model from [3]. However, we still need the family of contextual amino acid substitution matrices, for which a novel construction procedure is described. We present preliminary experimental results that illustrate the advantage of using a contextual approach in progressive alignment algorithm. It turned to be particularly useful in aligning the family of sequences containing several *orphans* (these are distantly related sequences, sometimes sharing the common fold).

It should be clear that the existence of orphan genes is unavoidable. Despite the accumulation of genetic information, newly sequenced genomes continue to reveal a high proportion (even to 50%) of uncharacterized genes. Among them there is a significant number of strictly orphan genes without any resemblance to previously determined protein sequences. Moreover, most genes found in databases have only been predicted by computer methods and have never been experimentally validated. Hence, for the alignment method it is important to tolerate orphans (some existing programs exclude the divergent orphans as unrelated or unalignable sequences) and to keep the stability of the family alignment when orphans are introduced into the sequence set.

The paper is organized as follows. We start with the description of an averaged contextual model. Then we present a construction method for contextual substitution tables. The next section proposes the progressive multiple alignment algorithm that takes context into account. The results of experimental analysis are presented in "results," which is followed by conclusions and discussion of further works.

Correspondence and reprint requests to Rafał Otto, Institute of Informatics, Warsaw University, Banacha 2, 02-097 Warsaw, Email: Poland; rotto@mimuw.edu.pl

AVERAGED CONTEXTUAL MODEL

The contextual alignment model considered in [3] cannot be directly applied to the problem of multiple alignment. In this model the score of an alignment depends on the order of operations (substitutions and indels) performed, as a substitution at one position can change the context for neighboring sites. The optimal alignment for the pair of sequences was defined as the alignment having the maximal score, when we maximize over all possible chronologies of evolutionary changes. More detailed study of the structure of optimal alignments and the description of efficient algorithms constructing them are included in [3].

To deal with several sequences simultaneously and also to keep the context dependency, we propose a relaxed contextual model. In this model we also penalize substitution considering two surrounding letters but we do not take care of the relative order of operations. In our algorithm the context independent and affine gap penalties are assumed.

Consider the following example of a short fragment of pairwise alignment:

$$\begin{aligned} &\ldots \text{HCA} \ldots \\ &\ldots \text{ADG} \ldots \end{aligned}$$

In the contextual model the score for substitution $C \mapsto D$ would depend on the order of operations. For instance, if the substitution $H \mapsto A$ has been performed after the substitution $C \mapsto D$ and the substitution $A \mapsto G$ has been performed before $C \mapsto D$, then the substitution $C \mapsto D$ would have the left context H and the right context G. In our simplified model we consider all 4 possible contexts for the middle substitution and take an average of 4 contextual scores. Notice that standard noncontextual, for example, Blosum matrix entry can be viewed as an average over all 400 possible pairs of contexts.

As a second example, consider the substitution surrounded by a deletion on the left and an insertion on the right:

$$\begin{aligned} &\ldots \text{HAHCC} - - - \text{A} \ldots \\ &\ldots \text{A} - - - \text{DDGAG} \ldots \end{aligned}$$

Now we have 9 possible contexts for the substitution $C \mapsto D$. On the left there are two different operations: the substitution $H \mapsto A$ and the deletion of AHC. If none of them has happened before the substitution $C \mapsto D$, then the left context is C. If AHC has been deleted before, then the left context is A or H depending on the relative order of these two substitutions. Analogous cases can be considered for the right context of the substitution $C \mapsto D$. As before we count the score as the average over 9 contextual scores.

CONTEXTUAL SUBSTITUTION TABLES

Methods

The contextual alignment algorithm, as an important part of its input data, takes a contextual scoring table, which provides the score for every possible substitution in every possible context.

The family of matrices proposed in [5] suffers from the fundamental difficulty that the amount of data necessary to construct a complete contextual substitution table exceeds the data presently available by an order of magnitude. To cope with this problem we present here a new approach to the construction of contextual substitution tables. The algorithm is in fact a contextual extension of the one that has been used to create Blosum tables [7].

As the input data we take the database of blocks (ungapped fragments of multiple alignments) and start by computing the observed frequency of substitutions. The extension from the existing method is the fact that we distinguish substitutions having different contexts. Let $f_{i,j}^{k,l}$ denote the number of observed substitutions $i \mapsto j$ in the context of k and l.

Each of the considered substitutions can have 4 different contexts so instead of increasing by 1 the entry $f_{i,j}$ we increase entry $f_{i,j}^{k,l}$ by 1/4 for all four possible pairs of (k, l).

Having computed frequency table we define the observed frequency for each substitution $i \mapsto j$ in the context (k, l) as

$$q_{i,j}^{k,l} = \frac{f_{i,j}^{k,l}}{\sum_{i,j} \sum_{k,l} f_{i,j}^{k,l}}. \tag{1}$$

Now, we can compute the observed frequency of the residue i in the context (k, l) as

$$p_i^{k,l} = q_{i,i}^{k,l} + \frac{1}{2} \sum_{i \neq j} q_{i,j}^{k,l}, \tag{2}$$

and the observed frequency of the context (k, l) as $u^{k,l} = \sum_i p_i^{k,l}$.

The expected frequency of the substitution $i \mapsto j$ in the context (k, l) is given by

$$e_{i,j}^{k,l} = \begin{cases} \dfrac{p_i^{k,l} p_j^{k,l}}{u^{k,l}} & \text{for } i = j, \\ \dfrac{2 p_i^{k,l} p_j^{k,l}}{u^{k,l}} & \text{for } i \neq j. \end{cases} \tag{3}$$

Finally the score for $i \mapsto j$ in the context (k, l) is

$$s_{i,j}^{k,l} = \log_2 \left(\frac{q_{i,j}^{k,l}}{e_{i,j}^{k,l}} \right). \tag{4}$$

To avoid the influence of highly similar sequences we adopt the idea of clustering inside blocks as it was done in the Blosum table.

Table 1. Characteristics of substitution scores.

Clustering %	NONCTX			CTX		
	Avg	StdDev	Entropy	Avg	StdDev	Entropy
100%	−0.5984	1.4561	1.0642	−0.4715	1.5498	1.0558
90%	−0.3363	1.2165	0.6515	−0.3124	1.2537	0.6620
80%	−0.2472	1.1086	0.5128	−0.2310	1.1316	0.5248
70%	−0.1658	0.9878	0.3839	−0.1590	0.9999	0.3970
60%	−0.0928	0.8449	0.2590	−0.0931	0.8523	0.2716
50%	−0.0429	0.6858	0.1519	−0.0500	0.7040	0.1622
40%	−0.0110	0.5607	0.0883	−0.0278	0.6013	0.1002

Table 2. The robustness of noncontextual tables. The range, median, and standard deviation for the number of examples drawn on per substitution score.

Table	No of pairs used	Min	Max	Med	StdDev
NONCTX100	910427386	201204	44246771	2089431	6447364
NONCTX90	397939179	85159	17134863	1154422	2226931
NONCTX80	228719630	52834	8703468	719781	1159503
NONCTX70	125188080	32428	4022674	429833	563942
NONCTX60	58669007	17718	1468427	218982	228104
NONCTX50	21889157	8121	424847	94091	70724
NONCTX40	7104342	3034	110252	32151	21160

Table 3. The robustness of contextual tables. The range, median, and standard deviation for the number of examples drawn on per substitution score.

Table	No of pairs used	Min	Max	Med	StdDev
CTX100	910427386	80	3033544	72832	105276
CTX90	397939179	52	1112640	38208	33440
CTX80	228719630	40	504100	21952	17628
CTX70	125188080	24	211316	14016	8620
CTX60	58669007	8	98096	1644	3560
CTX50	21889157	4	18736	700	1116
CTX40	7104342	1	7016	228	352

Results

As an input we have taken the BLOCKS+ database available at http://blocks.fhcrc.org (see [6]), which consists of 11 858 blocks representing 2608 groups. We have derived two kinds of tables: noncontextual (using the method in [7]) and contextual using the method described above. We have created tables with 7 different clustering percentages: 100%, 90%, 80%, 70%, 60%, 50%, and 40%. In Table 1 several characteristics of computed tables are summarized. The interesting observation is that the contextual tables have higher average score and higher entropy. Entropy also increases with clustering percentage as a normal consequence of reducing multiple contributions to amino acid pair frequencies from the most closely related sequences in the block. For the discussion of the notion of entropy in the context of substitution tables see [8].

The size of contextual tables (84 000 entries instead of 210 in case of noncontextual tables) implies a small amount of data that supports each table entry. If these statistics were too low, this could have direct impact on the quality of the score value. Tables 2 and 3 give a good view of these issues. Therefore we should discuss the robustness of proposed methods. The substitution table which was finally used in our experiments (CTX70) has an acceptable number of pairs impacting the average score; moreover there was no hole (blank entry) in this table.

For interested readers the matrices parameterized by different clustering constants can be found at http://www.mimuw.edu.pl/~aniag/TABLES.

MULTIPLE ALIGNMENT ALGORITHM

The averaged contextual model is proposed to enable computing multiple alignment in the contextual manner.

Ignoring the relative order of operations in the alignment simplifies the task of computing multiple alignment; however it is still not easy to keep the complexity on the reasonable level. Multiple alignment dynamic programming algorithm is extremely time consuming even in the case of the noncontextual model. A lot of heuristic approaches, which have been already developed, try to reach the optimal solution with the highest possible probability. The progressive alignment [1, 9] is one of the most popular alignment approaches.

Our contextual multiple alignment algorithm can be viewed as a contextual extension of popular ClustalW algorithm [9] or Feng-Doolittle algorithm [2], which belong to the family of progressive alignment algorithms. The main idea is to align pairs of sequences progressively and to deduce the multiple alignment from the set of pairwise alignments.

To this aim we have developed efficient averaged contextual pairwise alignment algorithms. These are appropriately modified standard dynamic programming procedures. We omit the details here; for interested readers all algorithms implemented in C++ can be found at http://www.cern.ch/rotto/Biology/Sources/ACM.

The important remark here is that we do not (not yet) intend to concur with the existing algorithms. Our goal is to demonstrate the usefulness of the contextual approach. We want to design algorithms that can be applied to the contextual model as well as to the noncontextual one. Then, we are able to compute alignments in both models and finally compare the results. In fact the ClustalW algorithm is equipped with the huge number of additional nontrivial heuristics (such as sequences weighting, substitution matrices varied at different alignment stages, residue-specific gap penalties, etc) which are not applied in our algorithm.

An overview of our algorithm is as follows:

(1) Calculate a distance matrix from pairwise scores for a given group of sequences.

(2) Construct a *guide tree* from the distance matrix using the neighbor-joining clustering algorithm [10].

(3) Progressively align the sequences in order of decreasing similarity. Three kinds of alignments are considered here:

  (i) pairwise alignment of two sequences,

  (ii) alignment of a sequence with an alignment,

  (iii) alignment of two alignments.

Calculating distance matrix

Several methods to derive the pairwise evolutionary distance (sometimes called difference score) from alignment scores are proposed (see, eg, [2]). Being aware of the drawbacks of all these approaches (see Gonnet and Korostensky, *Optimal scoring matrices for estimating distances between aligned sequences* available at http://www.inf.ethz.ch/personal/gonnet/papers/Distance/Distance.html for a detailed discussion) we decided to use the method proposed by Feng and Doolittle [2]. It works for global and local alignments. Assuming that $S(V, W)$ is the local similarity score between the sequences V and W, then their distance is defined via

$$D(V, W) = -\ln\left(\frac{S(V, W) - S_{\text{rand}}}{S_{\text{iden}} - S_{\text{rand}}}\right), \tag{5}$$

where S_{iden} is the average of the two scores for the two sequences compared with themselves and S_{rand} is the expected score of two random sequences with the same amino acids composition as V and W.

Constructing guide tree

The next step in progressive multiple alignment is building the guide tree. Here we use the neighbor-joining method [10] and two alternative methods to find the root of the tree. The first one is by adding an outgroup sequence to the given sequence group [1] and the second is by finding the middle point of the tree.

Aligning

The last step is to progressively align sequences according to the order given by the guide tree. It means that for each internal node of the tree we align sequences already aligned from the left child of the node with sequences already aligned from the right child of the node. In the simplest case this alignment is pairwise, but closer to the root of the tree we have to align two alignments. We decided to solve that problem using the method of Feng and Doolittle [2]. First, in two given alignments we replace gap letter with *neutral letter* X having at the end two groups of sequences over the extended alphabet $\Sigma \bigcup \{X\}$ (where Σ is an alphabet of amino acids). Also, we extend substitution matrix by adding scores for $a \mapsto X$ equal to 0 for all a. Then for each pair V, W of sequences where V is a sequence from the first group and W is a sequence from the second group, we compute pairwise alignment. The alignment with the maximal score is chosen and according to it we align two groups of sequences. Finally, in all sequences we replace *neutral letter* X back with a gap letter. In that way we obtain multiple alignment while reaching the root of the tree.

RESULTS

BAliBASE: multiple alignment test set

BAliBASE (Benchmark Alignments dataBASE) is a database of manually refined multiple sequence alignments available at http://www-igbmc.u-strasbg.fr/BioInfo/BAliBASE/.

It is specifically designed for the evaluation and comparison of multiple sequence alignment programs. The sequences included in the database are selected from alignments in structural databases (such as FSSP and

HOMSTRAD) or from manually constructed structural alignments taken from the literature. In our experimental analysis we have used the test sequences from 4 (out of 8) parts of the database, the so-called reference sets.

(i) *Reference 1.* It consists of families of equidistant protein sequences of similar length. Sequences are divided into 6 groups depending on theirs lengths and percent residue identity (% ID).

(ii) *Reference 2.* Each set here contains the group of closely related sequences (more than 25% ID) and up to three orphan genes (ie, genes sharing the common fold with the family, but having weak sequence similarity).

(iii) *Reference 3.* It includes groups consisting of several divergent protein families of equidistant sequences. The reference alignments consist of up to 4 families, with less than 25% ID between any two sequences from different families.

(iv) *Reference 6.* This includes the protein families containing repeats of different residue similarity.

Methodology

In the first stage of our experiment the multiple alignments were calculated for all reference sets in two settings: contextual and noncontextual. Then, the results obtained were compared with the reference alignments from the database. For this comparison the following measure (*sum-of-pairs score* [11]) was used. Let A_1 and A_2 be two multiple alignments of N sequences. Denote by M_1 and M_2 the lengths of these multiple alignments. Let $A(i, j)$ stand for the ith residue in the jth sequence of A. Define for two residues a and b $\delta(a, b) = 1$ if and only if $a = b$ and $\delta(a, b) = 0$ if and only if $a \neq b$. Now, for one column from the multiple alignment A we define

$$S(A, i) = \sum_{j=1}^{N} \sum_{k=1, k\neq j}^{N} \delta\big(A(i, j), A(i, k)\big). \tag{6}$$

And, finally

$$\mathrm{SPS}(A_1, A_2) = \frac{\sum_{i=1}^{M_1} S(A_1, i)}{\sum_{i=1}^{M_2} S(A_2, i)}. \tag{7}$$

SPS is the frequency of properly aligned pairs of residues with respect to the reference alignment.

In the second phase of the analysis we examine the robustness of the alignment to the introduction of orphans. To this aim we use the alignments from Reference 2, which contains related families with divergent, orphan sequences. Denote by $\mathcal{G}$ the set of all sequences from the considered group. Let $\phi(\mathcal{G})$ be the subset of $\mathcal{G}$ consisting only of the family of highly related sequences. Firstly, the multiple alignments $A_{\mathcal{G}}$ were calculated for all groups $\mathcal{G}$; then the multiple alignments for reduced groups (without orphans) $A_{\phi(\mathcal{G})}$. Let $\Phi(A_{\mathcal{G}})$ be the operation of cutting out from $A_{\mathcal{G}}$ the rows which correspond to the orphans. Define the following measure:

$$\mathrm{SPS}' = \mathrm{SPS}\big(A_{\phi(\mathcal{G})}, \Phi(A_{\mathcal{G}})\big). \tag{8}$$

It tests the ability of a model to align divergent sequences and also the degree to which the alignment of the family is disrupted by the introduction of the orphans. We have performed the experiments with various substitution tables and gap penalties. The best scores are obtained for NONCTX70 and CTX70 with gap open penalty 5 and gap extension penalty 1.

Results of the first experiment are presented in Table 4. The entries are the SPS measures averaged over the groups of sequences. Clearly, the contextual approach yields much better in case of sequence families from Reference 2 set, which contains families with orphan genes (especially in case of families of short sequences).

Table 5 summarizes the outcomes of the second experiment. The entries here are SPS′ values for investigated groups of sequences. Results of this experiment confirm the observation taken in the previous one.

The advantage of the contextual approach in aligning families containing orphan genes shown above is quite clear. Here we present some statistics taken on the whole set of experimental data to show that our method is performing a little better than other existing methods also in general case. Figure 1a proves that the results of the contextual model fit those given by the noncontextual one, but Figure 1b shows that contextual scores are more uniform. The fact that the contextual approach improves alignment more significantly in case of small values of noncontextual score is presented in Figure 1c.

Figure 1, Tables 4 and 5 then show the contextual model performs slightly better than the noncontextual one. However, there are some examples when the contextual approach yields much better results. Among them we have the families listed in Table 6.

The challenging task here is to explain the biological phenomena that stand behind such an excellent behavior of the contextual approach in all of these examples. Answering this question could help to discover a better alignment algorithm that profits from contextual information. Probably such an algorithm could be very efficient for the sequences belonging to some special class of sequences. The characterization of this class remains an interesting open problem.

CONCLUSIONS AND FURTHER WORKS

It is clear that the experimental analysis described in this work is just a beginning and cannot be treated as a definitive proof. Various improvements and other experiments can be envisaged.

Table 4. Summarized score for contextual versus noncontextual model. Score here corresponds to the frequency of properly aligned pairs of residues.

Reference	Protein families	Context	Noncontext	% of improvement
Ref 1	Short (< 25%)	0.5619	0.5260	6.83
	Short (20%–40%)	0.7323	0.7309	0.19
	Short (> 35%)	0.9004	0.8964	0.45
	Medium (< 25%)	0.4034	0.4091	−1.39
	Medium (20%–40%)	0.7951	0.7879	0.90
	Medium (> 35%)	0.9202	0.9198	0.04
	AVG	0.7379	0.7318	0.83
Ref 2	Short	0.6868	0.6633	3.52
	Medium	0.6580	0.6561	0.30
	AVG	0.6742	0.6602	2.12
Ref 3	Short	0.4008	0.4263	−5.49
	Medium	0.5880	0.5790	1.55
	AVG	0.4810	0.4917	−2.17
Ref 6	AVG	0.45	0.442	1.81
AVG	—	0.6674	0.6610	0.96

Table 5. The influence of orphans on the quality of the alignment.

Protein family	Context	Noncontext	% of improvement
Short	0.8918	0.8461	5.4
Medium	0.8593	0.8807	−2.4
AVG	0.8776	0.8613	1.89

Table 6. Families for which the contextual model gives much better alignments.

Protein family	No of sequences	Context	Noncontext	% of improvement
1ycc: cytochrome e	4	0.765	0.665	15.04
2trx: thioredoxin	4	0.671	0.468	43.38
1aboA: sh3	15	0.683	0.580	17.76
1uky: uridyl kin	24	0.541	0.464	20.91
sh3-2-ref6: sh3	6	0.553	0.454	21.81
sh3-3-ref6: sh3	5	0.430	0.214	100.93
AVG	—	0.606	0.474	29.11

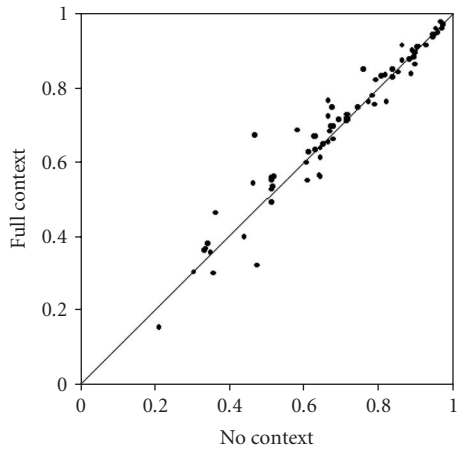
Wider and more distant contexts

In this paper we consider the simplest contextual model. The main idea of the context is to reflect the neighborhood of a given residue in the 3*D* protein structure. It suggests several possible extensions. One possibility is to consider a wider context, for example, two amino acids on each side. This approach is however limited by the huge size of substitution table ($20^6 = 64\,000\,000$ entries). The solution here is to consider a reduced context (ie, 20 amino acids can be divided into a small number of groups having similar biochemical properties (cf [5])).

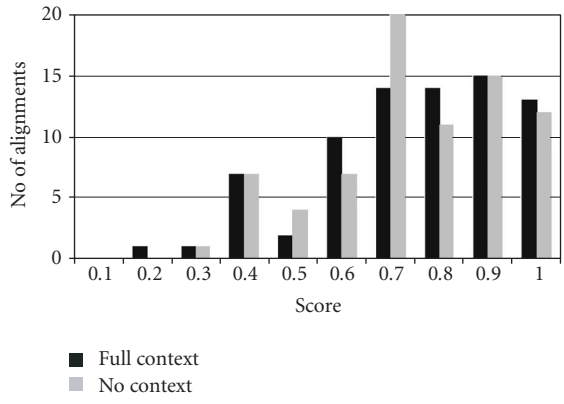
Another approach is to consider a more distant context. As an example look at the alignment

$$\ldots \text{CHCAD} \ldots$$
$$\ldots \text{HADGC} \ldots$$

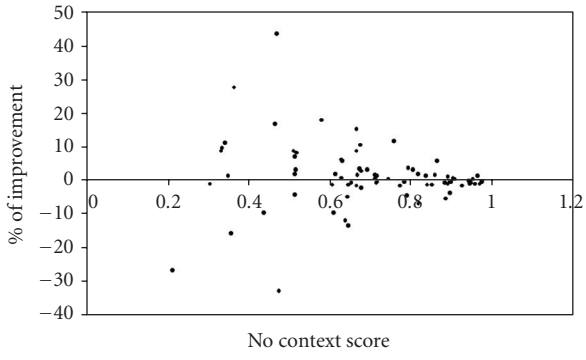
In the model we have presented, as a context we have taken two amino acids surrounding given substitution, that is, the left context of the substitution $C \mapsto D$ consists of amino acid H or A, and the right consists of A or G. It is however biologically motivated (by a secondary structure) to consider the contexts which are separated by



(a)



(b)



(c)

Figure 1. Comparison of Contextual and noncontextual scores. (a) SPS score comparison, (b) SPS score distributions, and (c) Improvement versus noncontextual score.

one position from the given residue, that is, the left context for the substitution $C \mapsto D$ is amino acid C or H, and the right context consists of D or C.

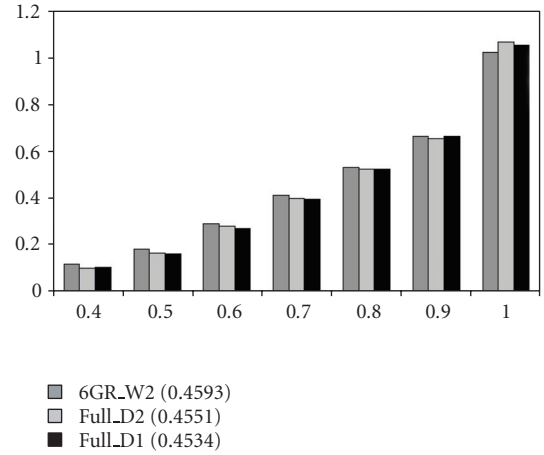


Figure 2. The entropy for substitution tables with a standard context (full_D1), a wider but grouped context (6GR_W2), and a more distant context (full_D2).

Preliminary results (see Figure 2) for the entropy of such defined substitution tables are very promising and encourage further research in this direction (more on entropy can be found in [8]).

Improvements for contextual multiple alignment algorithm

The algorithm presented in this paper follows the *progressive alignment* approach. The most popular algorithm and one of the most effective algorithms of this kind in the standard noncontextual model is ClustalW [9]. It contains a lot of additional heuristics. The challenging task is to design analogous improvements for the contextual model.

ACKNOWLEDGMENTS

This work was partially supported by the KBN Grant no 7 T11 F016 21. This work was also supported by the Open Society Institution (OSI) Grant no 18527.

REFERENCES

[1] Durbin R, Eddy S, Krogh A, Mitchison G. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids.* Cambridge, UK: Cambridge University Press; 1998.

[2] Feng DF, Doolittle RF. Progressive alignment of amino acid sequences and construction of phylogenetic trees from them. *Methods Enzymol.* 1996;266:368–382.

[3] Gambin A, Lasota S, Szklarczyk R, Tiuryn J, Tyszkiewicz J. Contextual alignment of biological sequences. *Bioinformatics.* 2002; 18 (suppl 2): S116–27.

[4] Gambin A, Slonimski P. Hierarchical clustering based upon contextual alignment of proteins: a different way to approach phylogeny. *C R Biol.* 2005;328(1):11–22.
[5] Gambin A, Tyszkiewicz J. Substitution tables for contextual alignment. In: Proceedings of Journees Ouvertes Biologie Informatique Mathematique (JOBIM 2002) ; 2002 San Malo, France.
[6] Henikoff S, Henikoff JG, Pietrokovski S. Blocks+: a non-redundant database of protein alignment blocks derived from multiple compilations. *Bioinformatics.* 1999;15(6):471–479.
[7] Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci USA.* 1992;89(2):10915–10919.
[8] Altschul SF. Amino acid substitution matrices from an information theoretic perspective. *J Mol Biol.* 1991;219(3):555–565.
[9] Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 1994;22(22):4673–4680.
[10] Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol.* 1987;4(4):406–425.
[11] Thompson JD, Plewniak F, Poch O. A comprehensive comparison of multiple sequence alignment programs. *Nucleic Acids Res.* 1999;27(13):2682–2690.