Substitution Matrices for Contextual Alignment*

Anna Gambin†
Centre de Génétique Moléculaire, CNRS-Gif sur Yvette, France
and Institute of Informatics, Warsaw University, Poland

Jerzy Tyszkiewicz
Institute of Informatics, Warsaw University, Poland
{aniag|jty}@mimuw.edu.pl

Abstract

It has been observed that the role an amino acid plays at a site in a protein depends on its environment. To approximate and take advantage of this dependence of context for improving the sensitivity of alignments we propose in [4] a new model for comparison of biological sequences. In this paper we present a computational procedure to construct substitution matrices for contextual alignment model. Our method is a suitable adoption of approach of Henikoff & Henikoff [6] to construct a BLOSUM substitution matrix. The drawbacks of the proposed algorithm are discussed. Using this procedure, a number of contextual matrices are generated and some preliminary experiments testing theirs applicability are performed.

Motivation There are numerous known examples in biology, showing that indeed context affects the likelihood of changes in amino acids or DNA sequences. One of them is the elimination of adjacent pairs cytosine-guanine in DNA, caused by biochemical mechanisms of replication. Another one is observed in proteins: the relevance of the amino acid properties depends on their context. Each amino acid chain occurs in some local environment. These contextual factors have a significant influence on the rates of substitution between the amino acids. For example, the acceptability of various amino acids at a site has been observed to correlate with the polarity of contacting chemical group [8, 10, 12].

---

*Research supported by Polish Research Council KBN grant 7 T11F 016 21
†This work was done during the postdoctoral fellowship from the CNRS within the framework of the Centre Franco-Polonais de Biotechnologie des Plants.

1 The Problem

In [4] the model of contextual alignment is defined. It amounts to adopting the assumption that, while constructing alignment of biological sequences the score of a substitution of one residue by another one depends on the surrounding residues, which we call the context. The simplest version lets the context to be just two neighboring residues.
In that paper the theory of contextual alignment is developed, including the study of the structure of optimal alignments and an efficient algorithm for constructing them. Assuming that the sequences are of length m and n, respectively and that the insertion and deletion scores are affine, both the global and the local contextual alignment algorithms work in time $O(mn)$. However, the algorithm assumes, as an important part of its input data, a contextual scoring table, providing the score for every possible substitution in every possible context.
The aim of this paper is to present an algorithm for creating such tables for contextual alignment of proteins.

2 The Approach

The idea of the approach is to a large extent based on the Henikoff & Henikoff BLOSUM table [6].

- The main source of data are blocks of gap-free aligned sequences.
- The method to eliminate the influence of large number of highly similar sequences is by clustering such sequences together and subsequently weighing the contribution of each cluster equally, no matter how many sequences it contains.

However, we have goals substantially extending those faced by the Henikoffs. First of all, we need to make the substitution score dependent of the context. Next, we want to break up the symmetry of the tables.
The former is conceptually not difficult, however, it requires large amounts of data to yield statistically significant results. For the latter, we decided to adopt the convention that: *at each position, the relatively most frequent residue is the primitive one,* being aware, that it is a highly speculative one.

3 The Algorithm

Input data. The input data are blocks of many gap-free aligned protein sequences. In particular, all sequences in one block are of equal length.

Parameters. The following are the parameters of the algorithm, and their choice affects the resulting tables.

- The clustering constant $p \in (0, 1)$.
- The significance threshold ST.

Clustering. In order to compensate for the high influence of many highly similar sequences, we introduce clustering. We cluster together sequences which share more than a fraction of p of residues for the purpose of creating the statistics, exactly as it has been done while creating BLOSUM (op cit. [6]). All blocks and all positions are taken into account.
Clusters in a block are connected components of the graph, whose vertices are sequences in the block and whose edge relation $\sim$ is defined as follows:

$$(s \sim t) \iff \frac{|\{i : 1 \leq i \leq |s| \ \& \ s(i) = t(i)\}|}{|s|} \geq p.$$

Subsequently we assume that each block is clustered. The cluster of the sequence s in block B is denoted $[s]_B$.
If B' is a subset of block B, then the cluster of the sequence s in B' is $[s]_{B'} := [s]_B \cap B'$, even though the latter set need not be connected in the graph B' with edge relation $\sim$. We adopt this choice to avoid the large cost of recomputing the clusters over and over again.
Henceforth we assume all blocks to be clustered and all their subsets to inherit the clusters in the way described above.

Identifying contexts. For each block B and each pair of positions $i, i+2$ not exceeding the common length of the sequences in B, and for each choice of amino acids a and b we create the subblock $B^{a,b}_{i,i+2} = \{s \in B : s(i) = a \ \& \ s(i+2) = b\}$. Subsequently, the subblocks $B^{a,b}_{i,i+2}$ over all B and all i are used to calculate the frequencies and substitution rates in the context a_b.

Frequencies in the given context. For every triple a, b, c of amino acids and each cluster W in each subblock $B^{a,b}_{i,i+2}$ we calculate the frequency of c found between a and b in W, denoted $f^{a,b}_W(c)$ in W, by the following formula:

$$f^{a,b}_W(c) := \frac{|\{s \in W : s(i+1) = c\}|}{|W|}.$$

Now the global frequency $f_{a,b}(c)$ of c being found between a and b is defined by

$$f_{a,b}(c) := \text{average}\left\{ f_W^{a,b}(c) : \begin{array}{c} W \text{ is a cluster of } B^{a,b}_{i,i+2}, \\ B \text{ is a block, } i, i+2 \leq \text{the length of } B \end{array} \right\}.$$

Note that averaging over the *real* frequencies in clusters amounts exactly to imposing that the contribution of each cluster is equal, irrespectively of its cardinality.

The primitive amino acid. Now, for each subblock $B^{a,b}_{i,i+2}$ the amino acid c for which the ratio

$$\frac{f_{B^{a,b}_{i,i+2}}(c)}{f_{a,b}(c)}$$

is the highest, is assumed to be *primitive* at position $i+1$ in $B^{a,b}_{i,i+2}$. This particular c is denoted $c_{B^{a,b}_{i,i+2}}$.

Mutation rates. For each quadruple a, b, c, d of amino acids the observed mutation rate of c into d in the context a_b, denoted $m_{a,b}(c, d)$ is given by the formula

$$m_{a,b}(c,d) := \text{average}\left\{ f_W^{a,b}(d) \cdot f_W^{a,b}(c) : \begin{array}{c} B \text{ is a block,} \\ W \text{ is a cluster of } B^{a,b}_{i,i+2}, \\ i, i+2 \leq \text{the length of } B, \\ c = c_{B^{a,b}_{i,i+2}},\ |W| \geq ST \end{array} \right\}.$$

In this formula, again weighing each cluster equally (but excluding too small clusters, where the frequencies are statistically insignificant), we calculate the frequency of c being replaced by d, in the context a_b.
The expected (under the null hypothesis) mutation rate of c into d in the context a_b, denoted $M_{a,b}(c, d)$ is given by the formula

$$M_{a,b}(c,d) := f_{a,b}(d) \cdot f_{a,b}(c).$$

Note that the null hypothesis is symmetric.

Log-odds. The score of substituting c by d in the context a_b, i.e., the entry in the tables we are creating, is defined by the formula

$$\text{score}_{a,b}(c,d) := \log_2 \left(\frac{m_{a,b}(c,d)}{M_{a,b}(c,d)} \right).$$

This method defines the scores as log-odds of the observed and expected mutation rates. For non-contextual gap-free alignment it has been proved by Altschul [1] that this is essentially the only choice, because even tables constructed in another way effectively generate a model of substitution with mutation rates whose log-odds give back the values from the tables.

4 The problems

The above algorithm has been implemented by T. Gajewski [3]. It appears that even for moderate values of ST and p many values in the tables remain undetermined, because the base of blocks used does not provide enough data. For some substitutions in some contexts, even if they happen in the base, they are either removed by the principle of disregarding clusters of cardinality smaller than ST or the amino acid to be substituted is indeed never chosen as the primitive one.
Being aware of such a risk, we can propose several remedies. Which of them will turn out to give practically meaningful results remains to be seen.

4.1 Reduced context tables

A context reduction function is any mapping ϕ from the set of amino acids to another set D of reduced contexts. An example could be $D = \{H, P\}$ with

$$\phi(c) = \begin{cases} H & \text{if c is hydrophobic,} \\ P & \text{if c is polar.} \end{cases}$$

Then, in order to obtain an algorithm for reduced context tables, we modify the following fragments of the main algorithm.

Identifying contexts. For each block B and each pair of positions $i, i+2$ not exceeding the common length of the sequences in B, and for each choice of $a, b \in D$ we create the subblock $B^{a,b}_{i,i+2} = \{s \in B : \phi(s(i)) = a \,\&\, \phi(s(i+2)) = b\}$.
Subsequently, in the whole algorithm a and b range over elements of D, and, in particular, the contexts are pairs of elements of D.
The reduced context table scan be used as such, for reduced context alignment, or as a source of missing values in the general tables. For the latter purpose, one can substitute the undetermined values $\mathtt{score}_{a,b}(c, d)$ in the general tables by the values $\mathtt{score}_{\phi(a),\phi(b)}(c, d)$ for a suitable context reduction function ϕ.
Another, quite different purpose of reduced context tables is to use them as a limiting case, allowing one to compare the tables produced by our algorithm with their non-contextual inspiration, the BLOSUM tables.

In order to achieve that, one takes a *constant* context reduction function. This amounts to saying that all contexts are the same, i.e., the context does not play any role in the construction of the tables. The outcome tables can be then directly compared with the BLOSUM tables (keeping in mind that BLOSUM is symmetrical, while our tables are not). It turns out that they are indeed quite similar[3].

4.2 Symmetric contextual tables

A significant method for improving the quality of the tables would be to give up with the asymmetry of the tables. In this situation the following changes to the algorithm are necessary.

The primitive amino acid. This procedure is eliminated.

Mutation rates. For each quadruple a, b, c, d of amino acids the observed mutation rate of c into d and *vice-versa* in the context a_b, denoted $m_{a,b}(c, d)$ is given by the formula

$$m_{a,b}(c, d) := \text{average} \left\{ f_W^{a,b}(d) \cdot f_W^{a,b}(c) : \begin{array}{c} B \text{ is a block,} \\ W \text{ is a cluster of } B_{i,i+2}^{a,b}, \\ i, i+2 \leq \text{the length of } B, \\ |W| \geq ST \end{array} \right\}.$$

5 The Results

Using the presented method we have constructed several dozens of families of contextual substitution matrices. Each set of matrices is the outcome of our procedure for a fixed set of parameters' values. Among all the input parameters the most interesting are: the source of blocks and the context reduction function.

The source of data. Input data for our procedure was prepared based on two data-bases of biological sequences:

- BLOCKS Database [5, 7]: blocks are multiply aligned ungapped segments corresponding to the most highly conserved regions of proteins. The blocks are made automatically by looking for the most highly conserved regions in groups of proteins documented in the Prosite Database.

- COGs Database [11]: Clusters of Orthologous Groups of proteins (COGs) were delineated by comparing protein sequences encoded in 44 complete genomes, representing 30 major phylogenetic lineages. Each COG consists of individual proteins or groups of paralogs from at least 3 lineages and thus corresponds to an ancient conserved domain.

The database of COGs does not contain itself blocks of sequences which could be directly used by the procedure. However for each cluster it contains the multiple alignment of its sequences. These alignments were obtained using the ClustalW algorithm. For our purpose we extracted blocks from the multiple alignments cutting off the maximal gap-free fragments of the alignments.
The comparison of matrices built from these two datasets yields the following observation: Despite of quite different origins of molecular sequences (mammalian protein sequences in the case of Henikoff & Henikoff blocks vs. whole genomic sequences of several microorganisms in the case of COGs) the resulted matrices are not significantly different. The explanation for this phenomenon is very rigorous clustering which is unavoidable when identifying contexts.

Partitions of the set of amino acids. The appropriate choice of context reduction function is crucial for the applications. It seems reasonable to consider the partition of the set of amino acids which reflects theirs chemical properties and which is suitable for studying the molecular evolution.
We have made some preliminary experiments using three different context reduction functions:

- full context, i.e. we consider 400 different pairs of contexts. The resulting matrices have some number of non-determined entries, which should be filled with reduced context values.
- 6 groups of context: These groups are based on accepted point mutation data [2]. The molecular sizes and shapes are very similar within each group. This is a crucial factor in determining which amino acid interchanges are acceptable to natural selection.

GROUP NAME	AMINO ACID RESIDUE
Small Aliphatic	Alanine, Proline, Glycine
Acid amide	Glutamine, Aspargine, Glutamic Acid, Aspartic Acid
Hydroxyl & Sulfhydryl	Serine,Threonine, Cysteine
Aliphatic	Valine, Isoleucine, Methionine, Leucine
Basic	Lysine, Arginine, Histidine
Aromatic	Phenylalanine, Tyrosine, Tryptophan

- HP context; we consider the partition of amino acids into two sets: polar ($\{C, D, E, G, H, K, N, Q, R, S, T, Y\}$) and hydrophobic ($\{A, F, I, L, M, P, V, W\}$).

To estimate the influence of the context we have performed the following experiment. We consider the set of sequences from a single cluster of orthologous groups (COG). All pairwise comparisons are done using both contextual and standard local alignment algorithms. Interesting results are also obtained when global alignment algorithm [9] is used instead of Smith-Waterman method and its contextual counterpart.

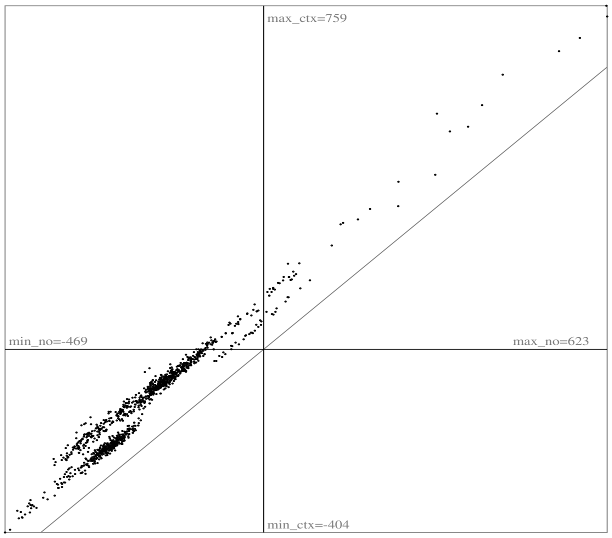


Figure 1: The pairwise global alignments of sequences from $COG0575$, the influence of context in the twilight zone.

For each pair of sequences, two scores (X**-axis** for the **non-contextual** score, Y**-axis** for the **contextual** score) represent one point on the Figure 1 and 2. One could observe that choosing more context groups results in better differentiating of pairwise comparison scores. The most promising fact is that the contextual approach better distinguishes pairs located in the so called twilight zone.

Context sensitivity. To determine the pairs of amino acids, substitution of which are context sensitive, we consider the contextual substitution matrix in the case of 6 context groups. For each substitution we calculate the minimal and maximal

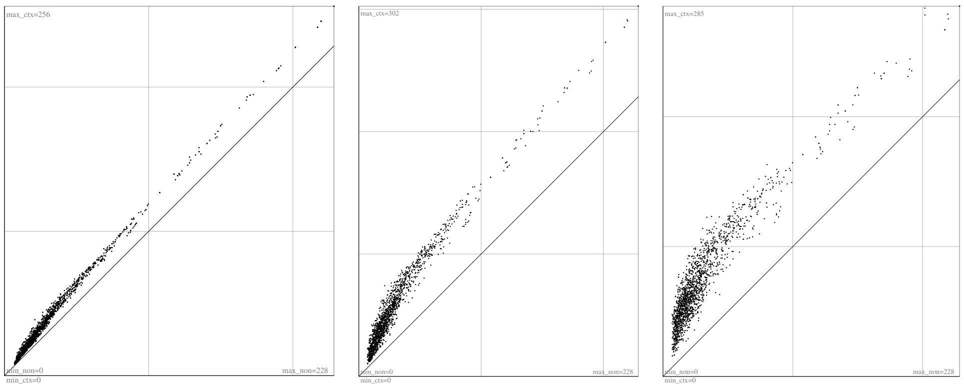


Figure 2: All pairwise optimal local alignments score for proteins from COG0089; three context reduction functions are used: HP, 6 groups, 20 groups.

value of the substitution score over all possible contexts, the mean and the standard deviation. The outcomes are summarized in the following table.

amino acids	mean	standard deviation	min	max
His Pro	-1.487	6.703	-12.498	0.276
Val Asp	-4.521	4.794	-11.918	-2.577
Asp Met	-3.21	4.335	-9.762	-0.524
Gly Met	-3.427	3.988	-13.401	-1.393
Pro Met	-1.86	3.646	-12.339	-0.349
His Asp	-0.108	3.603	-10.915	0.946
Ala Asp	-3.606	3.357	-10.643	-1.2
Arg Cys	-2.426	3.053	-10.9	0.106
Glu Cys	-3.351	3.044	-11.488	-0.325
Thr Cys	-1.602	2.974	-10.658	0.403
Cys Lys	-0.857	2.954	-10.085	0.535
Ile Asn	-3.174	2.933	-11.675	-1.024
Tyr Asp	-2.088	2.807	-10.506	-0.136
Leu Cys	-3.053	2.618	-9.492	-0.44
Met His	-0.555	2.438	-7.35	1.347

Asymmetry in the tables. The asymmetry in matrices implies that the score of an optimal alignment depends on the order of compared sequences. In some cases this difference can be significant, but in general the small amount of input data used for construction of matrices does not let to explore this property of amino acid substitutions.

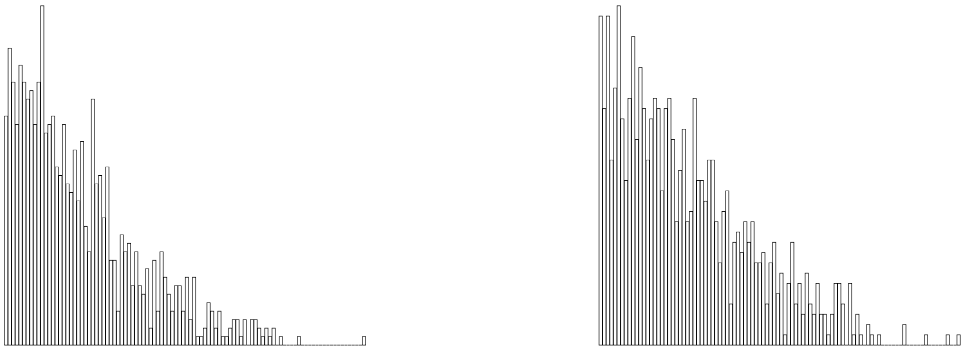
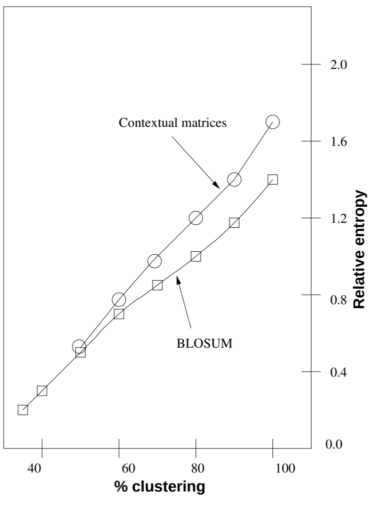


Figure 3: The distribution of differences resulting from asymmetric scores (contextual on the left and noncontextual on the right histogram) for all pairwise comparison of sequences from COG0013. The maximal difference is 169 compared with the maximal score 2051.

The relative entropy. The relative entropy is defined as a weighted average of all substitution scores, where weights are the observed frequencies of substitutions [1]. In out setting we define the entropy as follows:

$$H = \sum_{a,b} \sum_{c,d} m_{a,b}(c,d) score_{a,b}(s,d) = \sum_{a,b} \sum_{c,d} m_{a,b}(c,d) \log_2 \left(\frac{m_{a,b}(c,d)}{M_{a,b}(c,d)} \right).$$



In information theoretic terms, H is the relative entropy of the target and background distributions and intuitively it measures the average information available per position to distinguish the alignment from chance. The figure above illustrates the relationship between percentage clustering and relative entropy for contextual tables compared with the results for the BLOSUM family.

6 Final remarks

There are many ways to extent the approach described in this paper. First our definition of context was both limited and static. While the perspectives of considering wider context are rather pessimistic (the huge amount of data blocks is needed), taking into account more distant context are possible.
Below we sketch an alternative approach to construction of matrices. The main motivation in this approach is to take advantage of the whole data set; in this case the algorithm becomes very simple and can be easily parametrized for considering more distant context, e.g. context for amino acid i being the pair $i-2$, $i+2$ (motivated by the secondary structure of β sheets).

An alternative approach. The scores are defined as log-odds of the observed and expected mutation rates. The main difference is to eliminate clustering the sequences into subblocks $B^{a,b}_{i,i+2}$. Let $f_{c,d}$ be the total number of all pairs (substitutions) between amino acids c and d observed in the input data. Our aim is to group the set of these substitutions by context. To this end we investigate each substitution: it has at most 4 possible difrent contexts. We cannot distinguish among them hence we place the substitution with weight $\frac{1}{4}$ into four groups. In such a way we calculate the observed mutation rate of all substitution in all context $f^{a,b}_{c,d}$. Now the observed mutation rate is defined as:

$$m_{a,b}(c,d) = \frac{f^{a,b}_{c,d}}{\sum_{a,b}\sum_{c,d} f^{a,b}_{c,d}}.$$

The expected mutation rate $M_{a,b}(c,d)$ we calculate exactly as in [6] (for each context a_b independently):

$$M_{a,b}(c,d) = 2p^{a,b}_c p^{a,b}_d \quad \text{where} \quad p^{a,b}_c = m_{a,b}(c,c) + \frac{1}{2}\sum_{i \neq j} m_{a,b}(c,d).$$

We hope that this approach, in spite of being quite simple, may lead to the significant improvement of the quality of tables.

7 Acknowledgments

We thank Piotr Slonimski, Claude Thermes and Sławek Lasota for the helpful discussions.

References

[1] S. F. Altschul. Amino acid substitution matrices from an information theoretic perspective. *Journal of Molecular Biology*, 219:555–565, 1991.

[2] M. Dayhoff, R. Schwartz, and B. Orcutt. A model of evolutionary change in proteins. In M. Dayhoff, editor, *Atlas of Protein Sequence and Structure*, volume 5, suplement 3, pages 345–352. National Biomedical Research Foundation, 1978.

[3] T. Gajewski. Tablice substytucyjne dla modelu przyrównania kontestowego, 2001. M.S. Thesis (in Polish).

[4] A. Gambin, S. Lasota, R. Szklarczyk, J. Tiuryn, and J. Tyszkiewicz. Contextual alignment of biological sequences. Poster submission to RECOMB, 2002.

[5] J. Henikoff, E. Greene, S. Pietrokovski, and S. Henikoff. Increased coverage of protein families with the blocks database servers. *Nucl. Acids Res.*, 28:228–230, 2000.

[6] S. Henikoff and J. Henikoff. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA*, 89:10915–10919, 1992.

[7] S. Henikoff, J. Henikoff, and S. Pietrokovski. Blocks+: A non-redundant database of protein alignment blocks dervied from multiple compilations. *Bioinformatics*, 15(6):471–479, 1999.

[8] T. Ioerger. The context-dependence of amino acid properties. In *Intelligent Systems in Molecular Biology*, pages 157–166. AAAI Press, 1997.

[9] S. Needleman and Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, 48:443–453, 1970.

[10] J. Overington, D. Donnelly, J. Johnson, A. Sali, and T. Blundell. Environment-specific amino acid substitution tables: Tertiary templates and prediction of protein folds. *Protein Science*, 1:216–226, 1992.

[11] R. Tatusov, E. Koonin, and D. Lipman. Genomic perspective on protein families. *Science*, 278:631–637, 1997.

[12] P. Warme and R. Morgan. A survey of amino acid side-chains interactions in 21 proteins. *Journal of Molecular Biology*, 118:205–218, 1978.