Segment-based scores for pairwise and multiple sequence alignments

Burkhard Morgenstern[1,*], William R. Atchley[2], Klaus Hahn[1], and Andreas Dress[3]

[1]GSF – National Research Center for Environment and Health, Institute of Biomathematics and Biometry, Ingolstädter Landstr. 1, 85764 Neuherberg, Germany,
[2]Center for Quantitative Genetics, Department of Genetics, North Carolina State University, Raleigh NC 27695-7614, USA,
[3]Research Center for Interdisciplinary Studies on Structure Formation (FSPM), Universität Bielefeld, 33501 Bielefeld, Postfach 100131, Germany.

Abstract

In this paper, we discuss a novel scoring scheme for sequence alignments. The score of an alignment is defined as the sum of so-called *weights* of aligned segment pairs. A simple modification of the weight function used by the original version of the DIALIGN alignment program turns out to have a crucial advantage: it can be applied to both, global and local alignment problems without the need to specify a threshold parameter.

The alignment problem in computational biology

Sequence alignment is one of the most important tools of data analysis in molecular biology. Correspondingly, the problem of developing computer programs that are capable of automatically finding 'biologically correct' alignments, i.e. alignments reflecting the true biological relationships between sequences, is one of the great challenges in computational molecular biology.

It seems necessary to mention that this problem consists of two parts: First, an appropriate scoring scheme has to be defined by which the quality of different alignments of a given data set can be compared and evaluated. Then, given such a scoring scheme, the second part of the alignment problem is to find algorithms for the construction of optimal or at least reasonable suboptimal alignments according to that scheme. This paper is about the first part of the alignment problem.

The most popular scoring scheme for pairwise alignments was proposed in 1970 by Needleman and Wunsch (Needleman and Wunsch 1970). Given a *similarity matrix* consisting of similarity values for every possible pair of individual residues, they defined the overall similarity score of an alignment to be the sum of the similarity values of the aligned residues minus a penalty for every *gap* introduced into the alignment. Needleman and Wunsch also introduced a *dynamic programming* algorithm which finds an optimal alignment according to this criterion with reasonable computational costs. A local version of the Needleman-Wunsch method was proposed in 1981 by Smith and Waterman (Smith and Waterman 1981).

Since then, the alignment problem has generally been considered to be solved for pairwise alignments, and most efforts in the field of sequence alignment focussed on how to extend the Needleman-Wunsch method to multiple alignments and on how to choose the underlying parameters, especially the gap-penalty parameters (Fitch and Smith 1983, Vingron and Waterman 1994). There are several ways of generalizing the Needleman-Wunsch scoring scheme to the multiple alignment (Altschul and Lipman 1989, Gotoh 1986, Murata, Richardson, and Sussman 1985), and various methods have been proposed for finding optimal or suboptimal multiple alignments according to these criteria (Abdeddaïm 1997, Carrillo and Lipman 1988, Feng, Johnson, and Doolittle 1985, Thompson, Higgins, and Gibson 1994, Tönges *et al.* 1996, Vingron and Argos 1991, Stoye, Moulton, and Dress 1997).

If the sequences are closely related, an optimal alignment in the sense of Needleman and Wunsch generally approximates the 'biologically true' alignment, and alignment strategies based on this scoring scheme can therefore be applied successfully to construct biologically plausible alignments. However, if one considers distantly related sequences, there can be wide discrepancies between biologically meaningful alignments on the one hand and alignments with high Needleman-Wunsch scores on the other hand. Therefore, alignment strategies based on the Needleman-Wunsch optimization criterion often fail to produce acceptable alignments.

Segment-based alignment scores

Protein families are often characterized by a pattern of more or less conserved domains. This may be the result of functional and structural constraints during divergent evolution or of so-called *modular evolution*, for example by domain shuffling (Li 1997). Within these motifs, insertions and deletions are relatively rare

*To whom correspondence should be addressed (E-mail: morgenstern@gsf.de, FAX: ++49 89 3187 3127).

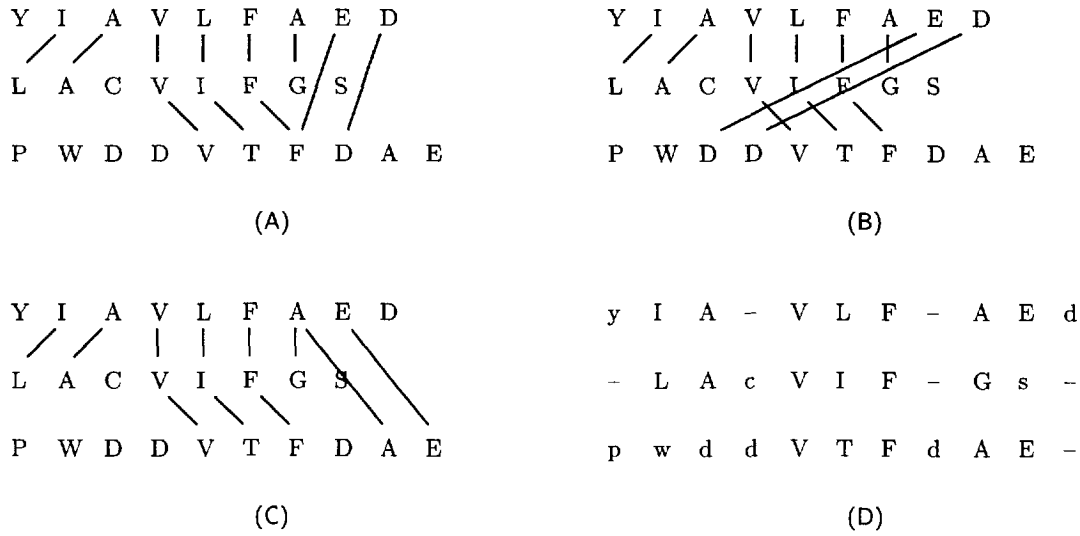

Figure 1: Non-consistent and consistent collections of *diagonals* (segment pairs). (A) and (B) represent *non-consistent* collections of diagonals: In (A) the 'F' in the third sequence is assigned simultaneously to two different residues of the first sequence. In (B) there is a 'cross over' assignment of residues. By contrast, (C) is a *consistent* collection of diagonals: It is possible to introduce *gaps* into the sequences such that residues connected by diagonals are in the same column of the resulting alignment (D). Residues not involved in any of the three diagonals are printed in lower-case letters. They are not considered to be aligned.

events. Therefore, it makes sense to employ alignment strategies based on comparing *gap-free segments* of the sequences in question rather than on comparing single residues.

Segment comparisons are successfully employed by data base search tools such as BLAST (Altschul *et al.* 1990) and FASTA (Pearson and Lipman 1988). Argos and Vingron have developed a pairwise alignment method where segment comparisons are used to reduce noise in the so-called *comparison matrix* (Argos and Vingron 1990). Waterman has proposed a method for multiple alignment based on *consensus words* of a given length (Waterman 1986). The MATCH-BOX program assembles alignments from *boxes* of segments (Depiereux and Feytmans 1992, Depiereux *et al.* 1997). The method allows for boxes of varying length. However, a pattern is reported only if it appears in *all* of the sequences, and all segments within a box must have the same length. The interactive program MACAW composes multiple alignments from sequence segments without these limitations (Schuler, Altschul, and Lipman 1991).

Recently, we proposed yet another alignment strategy relying on segment comparison (Morgenstern, Dress, and Werner 1996, Morgenstern *et al.* 1998). Alignments are composed of *gap-free pairs of sequence segments of equal length.* Such segment pairs are referred to as *diagonals* since they would appear as diagonals in the comparison matrix belonging to a pairwise sequence comparison.

A pairwise as well as a multiple alignment is considered to be represented by a collection of such diagonals meeting a certain *consistency* criterion. In short, a collection of diagonals is called *consistent* if there exists an alignment such that all segment pairs are matched (see Figure 1. For a precise mathematical definition of our notion of consistency, see Morgenstern, Dress, and Werner 1996). Diagonals may overlap if different pairs of sequences are involved. However, diagonals involving the same pair of sequences are not allowed to have any overlap (see Figure 1 and Figure 2).

Based on these ideas, we developed an algorithm which, given a set of two or more sequences, tries to find a suitable collection of diagonals representing an alignment of the input sequences. An implementation of this algorithm is distributed under the name DIALIGN 1 (for DIagonal ALIGNment).

The fundamental difference between the DIALIGN approach and other global or local alignment algorithms is the underlying optimization criterion: DIALIGN employs a so-called *weight function* assigning a *weight* to every possible diagonal. Given such a weight function, the *score* of an alignment is defined to be the *sum of the weights* of the incorporated diagonals. E.g., the score of the alignment in Figure 2 would be the sum of the weights of the seven diagonals it is composed of. (Note that in this definition of the score of an alignment, no *gap penalty* is involved.) Given this novel scoring

Sequences:

```
ASE-Fly       RRNARERNRVKQVNNGFALLREKIPEEVSEAFEAQGAGRGASKKLSKVETLRMAVEYIRSL
TFE3-Human    KKDNHNLIERRRRFNINDRIKELGTLIPKSSDPEMRWNKGTILKASVDYIRKL
MYC-Chicken   KRRTHNVLERQRRNELKLSFFALRDQIPEVANNEKAPKVVILKKATEYVLSI
```

Selected Diagonals:

```
 1   RRNARERNRVKQVNNGFALLREKIPEE   (ASE-Fly)
 4   NHNLIERRRRFNINDRIKELGTLIPKS   (TFE3-Human)

46   SKVETLRMAVEYIRSL   (ASE-Fly)
38   NKGTILKASVDYIRKL   (TFE3-Human)

 3   NARERNRVKQVNNGFALLREKIPE   (ASE-Fly)
 6   NVLERQRRNELKLSFFALRDQIPE   (MYC-Chicken)

42   SKKLSKVETLRMAVEYIRSL   (ASE-Fly)
33   NEKAPKVVILKKATEYVLSI   (MYC-Chicken)

 1   KKDNHNLIERRRR   (TFE3-Human)
 1   KRRTHNVLERQRR   (MYC-Chicken)

23   LGTLIPKSSDPE   (TFE3-Human)
23   LRDQIPEVANNE   (MYC-Chicken)

39   KGTILKASVDYIRKL   (TFE3-Human)
38   KVVILKKATEYVLSI   (MYC-Chicken)
```

Resulting Alignment:

```
ASE-Fly       ---RRNARERNRVKQVNNGFALLREKIPEEvseafeaqgagrgaSKKL-SKVETLRMAVEYIRSL
TFE3-Human    KKDNHNLIERRRRFNINDRIKELGTLIPKSSD-----------PEmrwNKGTILKASVDYIRKL
MYC-Chicke    KRRTHNVLERQRRNELKLSFFALRDQIPEVAN-----------NEKA-PKVVILKKATEYVLSI
```

Figure 2: Alignment of functional domains of three basic helix-loop-helix sequences as constructed by DIALIGN 2. The program has selected a *consistent* collection of seven segment pairs (so-called *diagonals*). Numbers on the left-hand side of the diagonals denote the first position of the respective segment. Residues involved in the selected segment pairs (*diagonals*) are shown in upper-case letters. Residues *not* belonging to any of these diagonals are shown in lower-case letters. They are not considered to be aligned.

scheme, the optimization task is to find a best scoring alignment – in other words: the task is to find a *consistent set of diagonals with maximal sum of weights.*

Weight functions for diagonals

It is obvious that the quality of alignments produced by this method depends first and foremost on the way the weights of diagonals are defined. The weight function employed by DIALIGN 1 is based on an idea proposed by Altschul and Erickson (Altschul and Erickson 1986): Given a diagonal D of length l_D, we denote by s_D the sum of the individual similarity values of residue pairs within this diagonal. If protein sequences are to be aligned, one of the usual substitution matrices, e.g. BLOSUM62 (Henikoff and Henikoff 1992), may be used.

Next, by $P(l_D, s_D)$ we denote the probability that a random diagonal of the same length l_D has at least the same sum s_D of similarity values. Mathematically, this probability is given as a sum of convolution products of the probability distribution of the individual similarity values. Then, the weight $w(D)$ of our diagonal D is defined to be $w(D) = -\log P(l_D, s_D)$ provided this value exceeds a certain user-defined threshold T, and is 0 otherwise. In addition, we require diagonals to have a minimum length of 7 residues.

Our experience has been that alignments optimized according to this scoring scheme are generally of high quality. (see Table 1, Table 2, Morgenstern, Dress, and Werner 1996, Morgenstern *et al.* 1998). Nevertheless, there is a general problem with the weight function as described above: It is absolutely necessary to specify either a certain minimum length for diagonals or a positive threshold T – otherwise even significant local similarities may get lost in the 'noise' of small random diagonals.

Generally, DIALIGN 1 tends to compose alignments from small diagonals. The length of the selected diagonals is often not much larger than the fixed minimum length of 7 residues. For example, if the data set shown in Figure 2 is aligned with DIALIGN 1, the alignment is composed from 16 short diagonals rather than from the 7 longer diagonals shown in Figure 2. It makes, of course, no difference if we include a long diagonal D into an alignment or if we split up D into several smaller diagonals $D_1, \ldots, D_n$, and then include all the diagonals $D_1, \ldots, D_n$ into the alignment – the resulting alignment is the same. For example, if we would split the first diagonal in Figure 2

```
RRNARERNRVKQVNNGFALLREKIPEE
NHNLIERRRRFNINDRIKELGTLIPKS
```

into three diagonals

```
RRNARERN    RVKQVNN    GFALLREKIPEE
NHNLIERR    RRFNIND    RIKELGTLIPKS
```

the resulting alignment would be exactly the same. One might therefore think that it made no difference if alignments are composed from many small or from few large diagonals.

The problem is, however, that if a weighting scheme for diagonals tends to assemble alignments from many shorter diagonals rather than from a few longer ones, it may easily happen that even significant 'biologically correct' diagonals are outweighed and displaced by 'biological wrong' random diagonals – especially if there is only local similarity among sequences.

Therefore, it is desirable to have a weight function defined on the set of all possible diagonals which would give relatively higher weights to longer diagonals of significant similarity. In other words, if a diagonal D with comparatively high and evenly distributed similarity between the paired individual residues is broken up into several smaller diagonals $D_1, \ldots, D_n$, the weight $w(D)$ should be significantly higher than the sum of weights $\sum_i w(D_i)$. Unfortunately, this is not the case with the weight function employed by DIALIGN 1 – which is exactly what necessitated the introduction of the threshold T and the minimum length for diagonals.

To overcome these shortcomings of the weight function employed by DIALIGN 1, we have introduced a new weight function defined on the set of all possible diagonals. Instead of considering the probability $P(l_D, s_D)$ of a *given* random diagonal to have a sum of individual similarity values of at least s_D, we considered the probability $P^*(l_D, s_D)$ to find *any* diagonal of length l_D whose sum of individual similarity values is at least as large as s_D somewhere within the comparison matrix of two random sequences of the same length as the original sequences. (Note that this probablility depends, of course, not only on the values l_D and s_D but also on the length of the sequences). We then defined the weight $w^*(D)$ of a diagonal D to be the negative logarithm of this probability.

Numerical values of the function P^* were calculated as follows: Probabilities $P^*(l, s) > 10^{-5}$ were determined based on random experiments. For smaller values, we used the simple approximation formula

$$P^*(l, s) \approx l_1 \cdot l_2 \cdot P(l, s)$$

where l_1 and l_2 are the lengths of the sequences. Putting $K := \log l_1 + \log l_2$, we obtain the following approximation formula:

$$w^*(D) = -\log P^*(l_D, s_D) \approx -\log[l_1 \cdot l_2 \cdot P(l_D, s_D)]$$

$$= -\log P(l_D, s_D) - \log l_1 - \log l_2 = w(D) - K,$$

i.e. we obtain the new weight function w^* by subtracting a constant K from the old weight function w (cf. Karlin and Altschul 1993). This is exactly the reason why the weight function w^* tends to assemble alignments from fewer longer diagonals rather than from many shorter ones: If a long diagonal D is replaced by n smaller diagonals $D_1, \ldots, D_n$, the constant K has to be subtracted n times instead of once from the respective values of w. We incorporated the new weight function w^* into our alignment algorithm, and we will refer to this new version of the program as DIALIGN 2.

Data set	Globins					Ribose			
Number of sequences	6					6			
Conserved domain	I	II	III	IV	V	I	II	III	IV
DIALIGN 1 (T=0)	4	5	6	4	6	6	6	5	5
DIALIGN 1 (T=10)	5	4	3,3	3,2	6	5	4	4	4
DIALIGN 2	5	4	3,3	3,3	6	6	5	4	4,2
CLUSTAL W	6	6	6	6	6	6	3	4	3,2
TWOALIGN	4	4	3,2	3,2	3,2	2,2	3	4	3
DCA	6	5	6	6	6	6	5	4	5
PIMA	5	4	3,2	3,2	3,2	2,2	5	5	2

Data set	Kinase								Protease		
Number of sequences	6								6		
Conserved domain	I	II	III	IV	V	VI	VII	IIX	I	II	III
DIALIGN 1 (T=0)	6	5	5	6	6	6	6	4	6	2	4
DIALIGN 1 (T=10)	6	5	6	6	6	6	6	4	5	0	3
DIALIGN 2	6	5	6	6	6	6	6	5	6	0	3
CLUSTAL W	6	6	6	6	6	6	6	4,2	5	3	4
TWOALIGN	6	2,2	4	6	6	5	6	3,2	5	0	2,2,2
DCA	6	6	5	6	6	6	6	6	4	4	2
PIMA	3,2	3,2	5	6	6	6	6	3	5	0	4

Table 1: Performance of different alignment methods applied to four different 'global' alignment problems. The table reports the ability of correctly aligning functional domains of the sequences. Entries in the table denote numbers of correctly aligned motifs. Multiple entries mean that a motif is correctly aligned *within* subgroups of sequences but not *between* these subgroups.

Results and discussion

To test the new version of our program systematically and to compare it to the original version as well as to other alignment programs, we used 7 different sets of protein sequences. Four of them are 'global' alignment problems, i.e. within each data set, sequences are globally related. The other 3 data sets are 'local' alignment problems where sequences share only isolated regions of local similarity.

The 'global' data sets are globin, ribose, kinase, and protease sequences. These sequences were used in (McClure, Vasi, and Fitch 1994) for a systematic comparison of alignment programs. Each data set contains 6 sequences. Within every data set, sequences have approximately the same length, and corresponding motifs are at similar positions within the sequences. Therefore, only relatively few gaps have to be inserted to correctly align these sequences.

The 'local' data sets are a set of 30 helix-turn-helix (HTH) proteins described in (Lawrence *et al.* 1993), a set of 16 acetyltransferases described in (Neuwald and Green 1994), and a set of 9 basic helix-loop-helix (bHLH) Proteins (Accession numbers P41894, Q02575, P17106, A55438, U10638, P13902, Q04635, U11444, A48085). Within each of these data sets, sequences differ considerably in length. Moreover, the conserved domains are at different positions within the sequences. A motif that occurs at the N-terminal of one sequence may occur at the C-terminal of another sequence.

We have tested the ability of different alignment programs to correctly align the conserved functional domains within these 7 data sets. In the bHLH sequences, the conserved domains are the first and the second α-helix as described in (Atchley and Fitch 1997). For all other data sets we have used the domains described in the quoted references.

We have tested the following programs: DIALIGN 1 (Morgenstern, Dress, and Werner 1996, Morgenstern *et al.* 1998), DIALIGN 2 (this study), CLUSTAL W (Thompson, Higgins, and Gibson 1994), TWOALIGN (Abdeddaïm 1997), Divide and Conquer (DCA) (Tönges *et al.* 1996, Stoye, Moulton, and Dress 1997), and PIMA (Smith and Smith 1992). All programs have been used with default parameters. In addition, we report the results of DIALIGN 1 with threshold $T = 10$ in order to study the influence of this parameter on the resulting alignments. The results of our program comparison are reported in Table 1 and Table 2.

In the 'global' alignment problems, alignments produced by DIALIGN 1 are comparable with alignments produced by standard global methods as DCA and CLUSTAL W. In these situations, DIALIGN 1 performed best *without* a threshold T, i.e. with the default value $T = 0$. In contrast, if sequences are only locally related, DIALIGN 1 was superior to standard alignment programs. However, in these 'local' alignment problems, it was necessary to specify a positive threshold (T=10) in order to obtain optimal results.

Data set	HTH	Transferase		bHLH	
Number of sequences	30	16		9	
Conserved domain		I	II	I	II
DIALIGN 1 (T=0)	6,6,3,2,2	12,2	9	7	3,2,2
DIALIGN 1 (T=10)	19,2,2	16	13,2	9	9
DIALIGN 2	24,2	16	14,2	9	9
CLUSTAL W	5,3,2,2,2	13	12	3,2	3,2
TWOALIGN	10, 6, 3, 3	13,2	6,5,2	9	5,2,2
DCA	5,3,2,2,2,2	11	11	0	2
PIMA	5,4,3,3,2,2	10,3,2	8,3,2	2	2

Table 2: Performance of different alignment methods applied to three different 'local' alignment problems. The table reports the ability of correctly aligning functional domains of the sequences. Entries in the table denote numbers of correctly aligned motifs. Multiple entries mean that a motif is correctly aligned *within* subgroups of sequences but not *between* these subgroups. A motif is considered to be correctly aligned if at least 75 % of the residues are correctly aligned.

The necessity of specifying a threshold T – depending on what kind of sequences are to be aligned – is a major drawback of DIALIGN 1.

Alignments produced by DIALIGN 2 seem to be comparable to the results of DIALIGN 1 and of standard global alignment methods as DCA and CLUSTAL W if sequences are globally related. However, if sequences are only locally related, DIALIGN 2 seems to be clearly superior to other methods. DIALIGN 2 yielded fully satisfactory alignments in both, 'global' *and* 'local' test examples *without* the necessity of specifying any parameter. Therefore, we think that, in view of a general applicability of the program, the new version of DIALIGN should be prefered to the old version.

Availability

An online version of DIALIGN 2 is available at

`http://bibiserv.TechFak.Uni-Bielefeld.DE/dialign/`

Acknowledgements

This work was done while B.M. was a visiting scientist at North Carolina State University. We would like to thank the GSF Research Center for their generous support of his stay at NCSU. We would like to thank Jens Stoye, Folker Meyer, Dirk Evers, and Alexander Sczyrba for helping us to establish the DIALIGN 2 WWW interface and Kornelie Frech and Kerstin Quandt for continuous support during program developement. Kurt Wollenberg has critically read the manuscript. The comments of three unknown reviewers proved to be very helpful.

References

Abdeddaïm, S. 1997, Incremental Computaton of Transitive Closure and Greedy Alignment. *Proc. of 8-th Annual Symposium on Combinatorial Pattern Matching.* Lecture Notes in Computer Science 1264, 167 - 179.

Altschul, S.F. and Erickson, B.W. 1986. A nonlinear measure of subalignment similarity and its significance levels. *Bull. Math. Biol.* 48, 617 - 632.

Altschul, S.F. and Lipman, D.J. 1989. Trees, Stars and Multiple Biological Sequence Alignment. *SIAM J. Appl. Math.* 49, 197 - 209.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D. 1990. Basic Local Alignment Search Tool. *J. Mol Biol.* 215, 403 - 410.

Argos, P. and Vingron, M. 1990. Sensitivity Comparison of Protein Amino Acid Sequences. *Methods in Enzymology* 183, 352 - 365.

Atchley, W.R. and Fitch, W.M. 1997. A natural classification of the basic helix-loop-helix class of transcription factors. *Proc. Natl. Acad. Sci.* 94, 5172 - 5176.

Carrillo, H. and Lipman, D. 1988. The multiple sequence alignment problem in biology. *SIAM J. Appl. Math.* 48, 1073 - 1082.

Depiereux, E. and Feytmans, E. 1992. Match-Box: a fundamentally new algorithm for the simultaneous alignment of several protein sequences. *CABIOS* 8, 501 - 509.

Depiereux, E., Baudoux, G., Briffeuil, P., Reginster, I., De Boll, X., Vinals, C., and Feytmans, E. 1997. Match-Box_server: a multiple sequence alignment tool placing emphasis on reliability. *CABIOS* 13, 249 - 256.

Feng, D.F., Johnson, M.S., and Doolittle, R.F. 1985. Aligning Amino Acid Sequences: Comparison of Commonly Used Methods. *J. Mol. Evol.* 21, 112 - 125.

Fitch, W.M. and Smith, T.F. 1983. Optimal sequence alignments. *Proc. Natl. Acad. Sci.* 80, 1382 - 1386.

Gotoh, O. 1986. Alignment of Three Biological Sequences with an Efficient Traceback Procedure. *J. Theor. Biol.* 121, 327 - 337.

Henikoff, S. and Henikoff, J.G. 1992. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci.* 89, 10915 - 10919.

Karlin, S. and Altschul, S.F. 1993. Applications and statistics for multiple high-scoring segments in molecular sequences. *Proc. Natl. Acad. Sci.* 90, 5873 - 5877.

Lawrence, C.E., Altschul, S.F., Boguski, M.S., Liu, J.S., Neuwald, A.F., and Wootton, J.C. 1993. Detecting subtle Sequence Signals: A Gibbs Sampling Strategy for Multiple Alignment. *Science* 262, 208 - 213.

Li, W.-H. 1997. *Molecular Evolution*, Sinauer Associates, Inc.

McClure, M.A., Vasi, T.K., and Fitch, W.M. 1994. Comparative Analysis of Multiple Protein-Sequence Alignment Methods. *Mol. Biol. Evol.* 11, 571 - 592.

Morgenstern, B., Dress, A., and Werner, T. 1996. Multiple DNA and protein sequence alignment based on segment-to-segment comparison. *Proc. Natl. Acad. Sci.* 93, 12098 - 12103.

Morgenstern, B. Frech, K., Dress, A., Werner, T. 1998. DIALIGN: Finding local similarities by multiple sequence alignment. *Bioinformatics* 14, in press.

Murata, M., Richardson, J.S., and Sussman, J.L. 1985. Simultaneous Comparison of three protein sequences. *Proc. Natl. Acad. Sci.* 82, 3073 - 3077.

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

Neuwald, A.F. and Green, P. 1994. Detecting Patterns in Protein Sequences. *J. Mol. Biol.* 239, 698 - 712.

Pearson, W.R. and Lipman, D.J. 1988. Improved tools for biological sequence comparison *Proc. Natl. Acad. Sci.* 85, 2444 - 2448.

Schuler, G.D., Altschul, S.F., and Lipman, D.J. 1991. A Workbench for Multiple Alignment Construction and Analysis. *PROTEINS: Structure, Function and Genetics* 9, 180 - 190.

Smith, T. and Waterman, M. 1981. Comparison of Biosequences. *Advances in Applied Mathematics* 2, 482 - 489.

Smith, R.F. and Smith, T.F. 1992. Pattern-induced multi-sequence alignment (PIMA) algorithm employing secundary structure-dependent gap penalties for use in comparative protein modeling. *Protein Eng.* 5, 35 - 41.

Stoye, J., Moulton, V. and Dress, A. 1997. DCA: An Efficient Implementation of the Divide-and-Conquer Approach to Simultaneous Multiple Sequence Alignment. *CABIOS* 13, 625 - 626.

Thompson, J.D., Higgins, D.G., and Gibson, T.J. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research* 22, 4673 - 4680.

Tönges, U., Perrey, S.W., Stoye, J., and Dress, A. 1996. A general method for fast multiple sequence alignment. *Gene* 172, GC33 - GC41.

Vingron, M. and Argos, P. 1991. Motif Recognition and Alignment for many Sequences by Comparison of Dot-matrices. *J. Mol. Biol.* 218, 33 - 43.

Vingron, M. and Waterman, M.S. 1994. Sequence Alignment and Penalty Choice. Review of Concepts, Case Studies and Implications. *J. Mol. Biol.* 235, 1 - 12.

Waterman, M.S. 1986. Multiple sequence alignment by consensus. *Nucleic acid research* 14, 9095 - 9102.