FEBS Letters 529 (2002) 126–130 FEBS 26432

Minireview

Quality assessment of multiple alignment programs

Timo Lassmann, Erik L.L. Sonnhammer*

Center for Genomics and Bioinformatics, Karolinska Institutet, SE-17177 Stockholm, Sweden

Received 24 July 2002; accepted 2 August 2002

First published online 14 August 2002

Edited by Gunnar von Heijne

Abstract A renewed interest in the multiple sequence alignment problem has given rise to several new algorithms. In contrast to traditional progressive methods, computationally expensive score optimization strategies are now predominantly employed. We systematically tested four methods (Poa, Dialign, T-Coffee and ClustalW) for the speed and quality of their alignments. As test sequences we used structurally derived alignments from BAliBASE and synthetic alignments generated by Rose. The tests included alignments of variable numbers of domains embedded in random spacer sequences. Overall, Dialign was the most accurate in cases with low sequence identity, while T-Coffee won in cases with high sequence identity. The fast Poa algorithm was almost as accurate, while ClustalW could compete only in strictly global cases with high sequence similarity.

Key words: Multiple sequence alignment; Alignment quality; Poa; Dialign; ClustalW; T-Coffee

1. Introduction

Multiple sequence algorithms play a crucial role in molecular biology today. The advent of large genome projects has led to an explosion of sequence data in public databases. Modern genome annotation and analysis tools rely heavily on accurate multiple alignments. The role of multiple sequence alignments in such systems has changed from simply transferring annotation from one sequence to another to a genome wide perspective. Analysis of protein families and their evolution, and detection of remote homologs are now the primary objectives.

To meet the new challenges, several new multiple sequence alignment algorithms have been developed. Overall, the trend goes from traditional progressive methods [1,2,3] to iterative and computationally expensive ones [4–7]. At the heart of iterative methods lies the optimization of a score function. The aspiration is that the score function reflects biological events such that optimization of the score leads to a biologically correct alignment. Secondly, almost all methods apply some heuristics to speed up the alignment procedure. The sole exception is the MSA algorithm [8], which, for that reason, can only be applied to very small data sets.

Multiple alignment programs can be divided into two main categories: methods aligning sequences over their entire length (global) and methods aligning regions of high similarity only (local). Traditionally, the focus has been on global methods, exemplified by ClustalW, as they perform well in cases when all sequences are of similar lengths. However, to be able to deal with the increasingly difficult alignments coming from genome projects in which often only parts are alignable, local methods have enjoyed an increased interest.

A comprehensive analysis of four methods was conducted with the following aims:

- In which situations do local or global methods perform best?
- How do computationally inexpensive methods compare against the new 'heavy weight' approaches?
- How do methods perform in difficult cases (complex domain architecture, low sequence identity)?

In contrast to other recent reviews [9,10], here the sheer number of test cases was increased by an order magnitude for a greater coverage. We also present the first comprehensive testing with multi-domain proteins, which is intended to mimic situations involving domain shuffling.

2. Materials and methods

2.1. Alignment programs

The four algorithms employed here were ClustalW, Dialign, T-Coffee and Poa [3,5,6,11]. The most prominent program, ClustalW, is a global progressive method. The algorithm works in two steps: initially, a guide tree based on sequence similarity is constructed, followed by successive pairwise alignments in the order given by the tree. Since its first appearance a wide range of improvements have been added [3].

Poa is a recent progressive algorithm, employing partially ordered graphs, as opposed to generalized profiles, to represent aligned sequences. Generalized profiles are only accurate when sequences are related solely due to a process of insertions, deletions and mutations. Partially ordered graphs can represent global cut-and-paste operations, and thus reflect the biological contents of multiple alignments more accurately [11]. Problems caused by the inherent loss of information in generalized profiles are therefore avoided. Interestingly, no evolutionary tree is used to guide the order in which sequences are aligned. The two most similar sequences are determined and aligned and all other sequences are added to this one profile in a stepwise fashion.

Dialign is a local algorithm which aligns whole segments rather than single residues. Initially all pairwise alignments are performed and all aligned ungapped regions picked up. The name 'Dialign' comes from these regions as they would appear as diagonals on a dotplot. A consistent set of diagonals is determined and iteratively added to the alignment.

Similar to Dialign, T-Coffee firstly performs all possible pairwise alignments within the set of sequences. However, T-Coffee performs

*Corresponding author.
E-mail address: erik.sonhammer@cgb.ki.se (E.L.L. Sonnhammer).

0014-5793/02/$22.00 © 2002 Published by Elsevier Science B.V. on behalf of the Federation of European Biochemical Societies.
PII: S0014-5793(02)03189-7

this step twice: once with ClustalW (global) and once with Lalign (local-Fasta package [12]). The results from both methods are combined into a primary library. A library extension step determines how residue pairs align with respect to other residues. Such triplets are used to assess how well sequences are aligned given the other sequences in the dataset, rather than looking at two sequences in isolation. The final alignment is then built progressively using the information in the library.

All alignment programs were used 'out of the box', i.e. using the default parameters.

2.2. Reference alignments

To test these programs two distinct sources of alignments were used: the BAliBASE [13] test set and artificially created alignments using Rose [14]. The BAliBASE database was constructed using both manual and computational methods. Only core blocks, which are known to be structurally aligned, are given in the reference alignments. There are five categories in BAliBASE, encompassing alignments of variable lengths, sequence identity, and alignments with N/C terminal extensions or internal insertions. As noted before [9], the absence of full-length sequences biases the test set towards global methods. Due to this limitation, Rose, a program simulating evolution of sequences using a probabilistic model, was used alongside BAliBASE. A tree guides the evolution of sequences from a common ancestor using insertions, deletions and substitutions. As all events in the history of the sequences are known, the 'true' multiple sequence alignment is created on the fly.

2.3. Evaluation function

A critical factor in the analysis on alignments is the quality measurement. A number of score functions exist for alignment optimization, e.g. weighted sum-of-pairs, maximum likelihood, minimum entropy, star, and consensus [15]. Some methods require a sequence tree to be estimated while others do not. The most popular score function is the weighted sum-of-pairs score (WSP). However in the alignments tested here, 63% of the alignments created by the four programs had a higher WSP than the correct alignments, indicating that there is poor correspondence between WSP score and alignment quality.

Thompson et al. [10] introduced two scores for comparing an alignment to a reference alignment: the column score and the sum-of-pairs score (SPS). The column score counts how many columns are identical between the reference and test alignment. Because this measurement reflects the ability of a program to align all sequences, one misaligned sequence reduces the score to zero. The sum-of-pairs score is the fraction of residue pairs that are aligned the same way as in the reference alignment. A pair of aligned residues occurring in both alignments gives a score of one. A modification to this scoring method was introduced by Karplus et al. [9], who assigned a weight of two to identically aligned pairs of residues and a weight of one to residues aligned to a gap in both alignments.

In this paper we use a slightly modified version of the sum-of-pairs score, called the overlap score. All pairs of aligned residue pairs in the reference and test alignments are stored in two sets. The original alignments can be reconstructed using these sets, hence residues aligned to gaps were omitted as they constitute redundant informa-

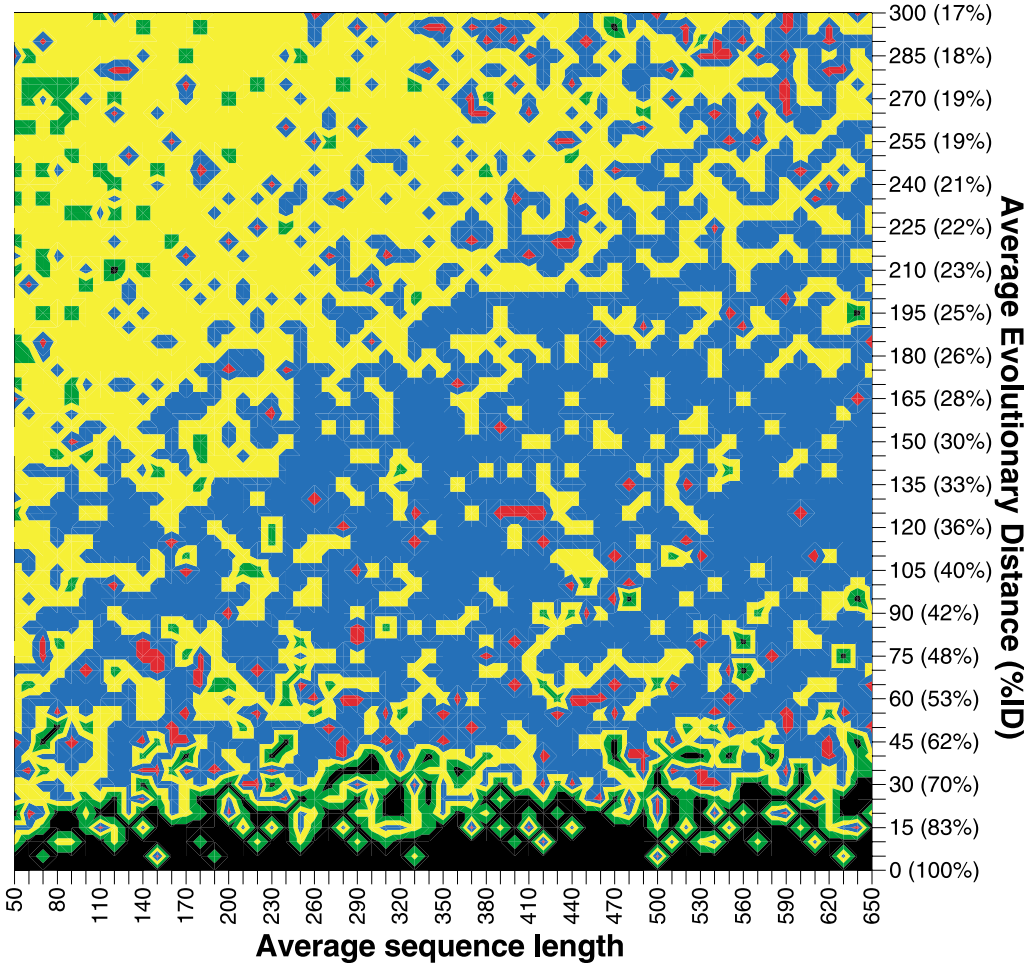

Fig. 1. Color coded matrix showing which method performed best for each pair-combination of conditions: average sequence length (x-axis) and average evolutionary distance (y-axis). The methods are Poa (green), Dialign (yellow), T-Coffee (blue) and ClustalW (red).

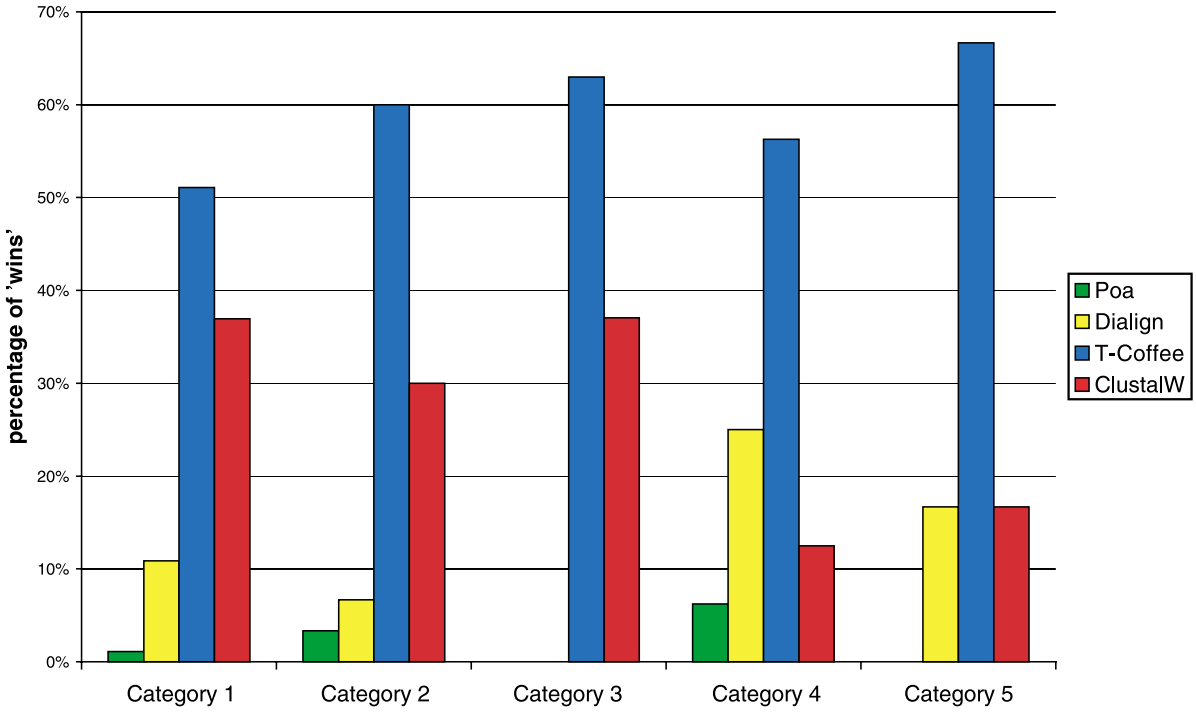

Fig. 2. Results of BAliBASE testing, showing the fraction that each program had the best accuracy (SPS) in each of the five BAliBASE categories (see text).

tion. The intersection of the two sets, i.e. the residue pairs found in both sets, is divided by the total number of pairs in both sets divided by two to yield the overlap score.

3. Results

3.1. Evolutionary distance and sequence length

In this test the effect of sequence length and evolutionary distance on alignment quality was tested using 3720 Rose alignments. All four algorithms were run on this test set and overlap scores to the reference alignment were calculated. As expected, all methods performed increasingly poorly with increasing evolutionary distance. Conversely, increased sequence length had a positive effect on the alignment quality in all cases. This corresponds to previous findings by Thompson et al. [10]. Only small differences between the four methods were observed (5% on average). Alignments created by Poa and ClustalW were on average marginally poorer than the alignments of Dialign and T-Coffee. To determine if certain conditions favor one particular method, a 2D matrix representing the 'winner' given one particular pair of conditions was created (Fig. 1). In black areas, two or more methods achieved the same optimal score. Overall, T-Coffee dominated at alignments with low to moderate evolutionary distances while Dialign performed best in alignments with high evolutionary distances. Both ClustalW and Poa only rarely perform better than the other two methods.

3.2. BAliBASE

The BAliBASE test sets were used in a similar fashion. The four alignment programs were run on all BAliBASE alignments and their overlap scores calculated. Again, a 'winner takes all' approach was taken to highlight the differences in performance. The results are summarized in Fig. 2. T-Coffee and ClustalW are the top two programs in categories one, two and three (global alignment with different % identity and orphan sequences). In categories four (N/C terminal extensions) and five (long internal insertions), T-Coffee dominates, while Dialign performs similarly to ClustalW. Poa only rarely produced the best alignment in all categories.

3.3. Multi-domain-proteins

To evaluate the effect of domain organization on the quality of alignments, three tests were carried out. Test sets were constructed by inserting one, two or three Rose alignments, each representing a unique domain, into randomly generated sequences. Only alignments that had an overlap score exceeding 95% with all four programs were used. The length of the random sequences was varied to make it increasingly difficult for the methods to accurately find and correctly align the domains. The average length of the domains was 50 in the single-domain case, 50 and 60 in the two-domain case, and 50, 60, and 100 in the three-domain case. Initial results suggested that changing domain order has a drastically negative effect on all alignment methods. To focus on the differences between the methods, the domain order in all three tests was kept the same. However, in the third test set the occurrences of domains was varied. One sequence could contain all three domains 'ABC', while another could contain only domains 'BC'. The alignment quality in all cases was calculated for each domain separately, using the initial Rose alignments as references. In cases two and three, the average score of all domains present is given. Fig. 3 shows the results.

Not surprisingly, the two local methods Poa and Dialign perform very well in all three tests. Although T-Coffee performs well in the single-domain case, it cannot compete with Poa and Dialign in the multi-domain alignments. ClustalW performs poorly in the single-domain case, in which the three

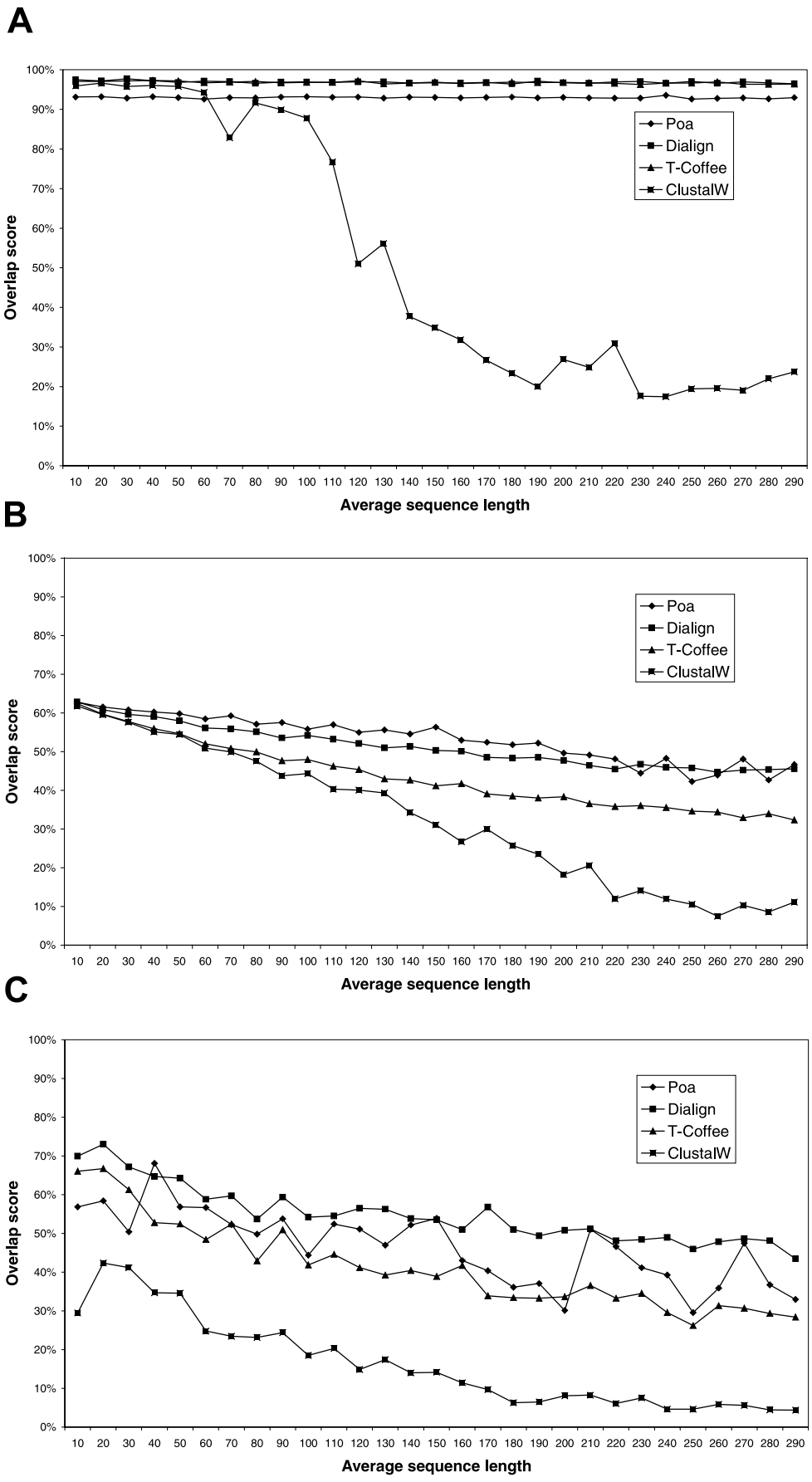

Fig. 3. Plots of the overlap scores achieved by Poa, Dialign, T-Coffee and ClustalW in synthetic cases generated by Rose. (a) One domain, (b) two domains and (c) three domains (one or two domains may be deleted).

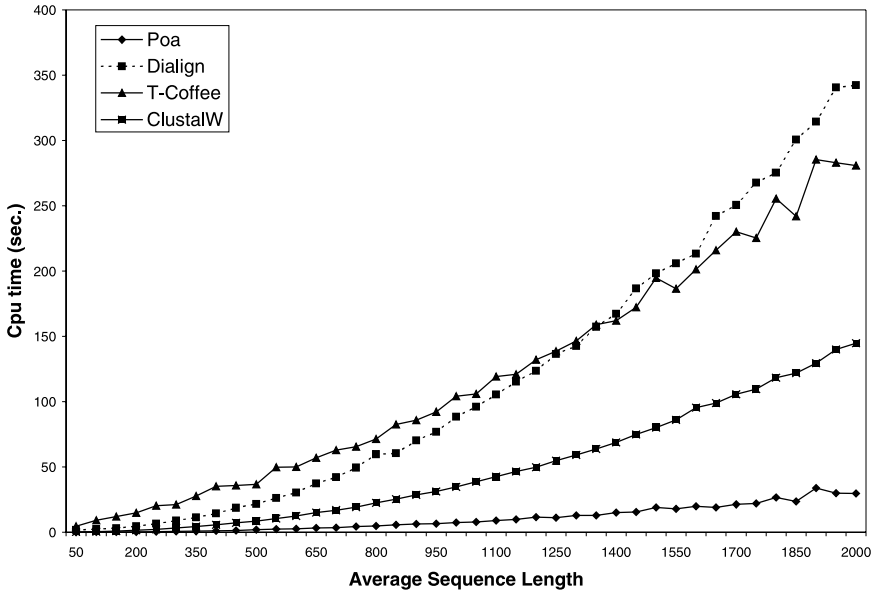

Fig. 4. CPU time consumed by each program to align sets of increasingly long sequences.

other methods surprisingly manage to maintain a constantly high accuracy as the lengths of the random 'base' sequences are increased.

The speed of all four methods was tested using a test set of Rose alignments with increasingly long sequences (Fig. 4). Poa is the quickest method, followed by ClustalW. Although Dialign is initially faster than T-Coffee, it has a higher time complexity, causing it to be the slowest method in large alignments.

4. Discussion

The quality of alignments produced by the four methods is clearly dependent on the initial data. A diverse set of sequences poses a problem to all methods. More importantly, in such cases the quality of resulting alignments is poor no matter what program was used. An exception to this rule is when input sequences share either predominantly local or global similarity. In such cases the difference in quality between local or global methods becomes significant.

Overall, the two computationally expensive methods T-Coffee and Dialign performed better Poa and ClustalW. However, in most cases the differences were only marginal. In automated annotation strategies, where computational power becomes prohibitive, one might therefore choose to opt for the quicker methods.

Perhaps the most important issue regarding multiple sequence alignments is the quality measurement. Although scoring functions employed in algorithms generally give reasonable alignments, a high score does not necessarily imply a good, i.e. biologically correct, alignment. Today, the only approach to assess alignment quality without a reference alignment is by comparing the outputs of several programs.

The arguably best alignment program to date is Poa. It produces good alignments in a fraction of the time taken by other methods. Especially in difficult multi-domain alignments, Poa stands out in terms of speed and quality.

References

[1] Feng, D.F. (1987) J. Mol. Evol. 25, 351–360.
[2] Higgins, D.G. and Sharp, P.M. (1989) Comput. Appl. Biosci. 5, 151–153.
[3] Higgins, D.G., Thompson, J.D. and Gibson, T.J. (1994) Nucleic Acids Res. 22, 4673–4680.
[4] Gotoh, O. (1996) J. Mol. Biol. 264, 823–838.
[5] Morgenstern, B. (1999) Bioinformatics 15, 211–218.
[6] Notredame, C., Higgins, D. and Heringa, J. (2000) J. Mol. Biol. 302, 205–217.
[7] Stoye, J. (1998) Gene 211, GC45–GC56.
[8] Lipman, D.J., Altschul, S.F. and Kececioglu, J.D. (1989) Proc. Natl. Acad. Sci. USA 86, 4412–4415.
[9] Karplus, K. and Hu, B. (2001) Bioinfomatics 17, 713–720.
[10] Thompson, J.D., Plewniak, F. and Poch, O. (1999) Nucleic Acids Res. 27, 2682–2690.
[11] Lee, C., Grasso, C. and Sharlow, M. (2002) Bioinformatics 18, 452–464.
[12] Pearson, W.R. and Lipman, D.J. (1988) Proc. Natl. Acad. Sci. USA 85, 2444–2448.
[13] Thompson, J.D., Plewniak, F. and Poch, O. (1999) Bioinformatics 15, 87–88.
[14] Stoye, J., Evers, D. and Meyer, F. (1998) Bioinformatics 14, 157–163.
[15] Gotoh, O. (1999) Adv. Biophys. 39, 159–206.