Comparison of Multiple Sequence Alignment programs

Sellis Diamantis (dsellis@biol.uoa.gr), Charissi Anna (aharis@di.uoa.gr)
MSc Bioinformatics, National and Kapodistrian University of Athens

Keywords: MSA, progressive, iterative, exact, ClustalW, T-Coffee, Dialing-T

1. Introduction

1.1. What is a MSA

A multiple sequence alignment is an optimal alignment of more than two sequences. An example is shown in Figure 1.

Figure1: Multiple sequence alignment of animal α-hemoglobins. Conserved regions are shown in red blocks.

1.2 Why MSAs are important

Multiple sequence alignments are of great importance for biological research. Moreover, the rapid accumulation of DNA sequences during the last years made MSA a necessary tool for research.

A MSA can reveal conserved residues that enable the identification of possibly important sites. For example, conserved aminoacid residues are usually involved in protein function or are responsible for protein structural stability. In DNA sequences, conserved regions can represent a regulatory element. Besides of identifying conserved residues a more sophisticated approach is to use information form a MSA by using regions of residues with conserved properties to construct a statistical model such as a Position Specific Scoring Matrix or perhaps a Hidden Markov Model. These models are used to identify conserved regions in newly sequenced genomes, or they are used to construct databases such as PROSITE (Hulo et al., 2004) or PFAM (Bateman et al., 2004)).

Sequencing of a whole genome is a difficult task, especially when large eukaryotic genomes are considered. However, one of the main difficulties was raised by the use of MSAs. While the sequencing of a short stretch of DNA is a routine for many molecular labs, it is practically impossible for large sequences. The solution is to cut the sequence in random short stretches and sequence them. The reconstruction of the whole chromosome is done in silico, by aligning the stretches and finding their overlaps.

The construction of multiple sequence alignments is closely related to phylogenetic analysis. A phylogenetic tree can be inferred by a multiple sequence alignment as shown in Figure 2. The study of molecular evolution is an area where MSA is extensively used.

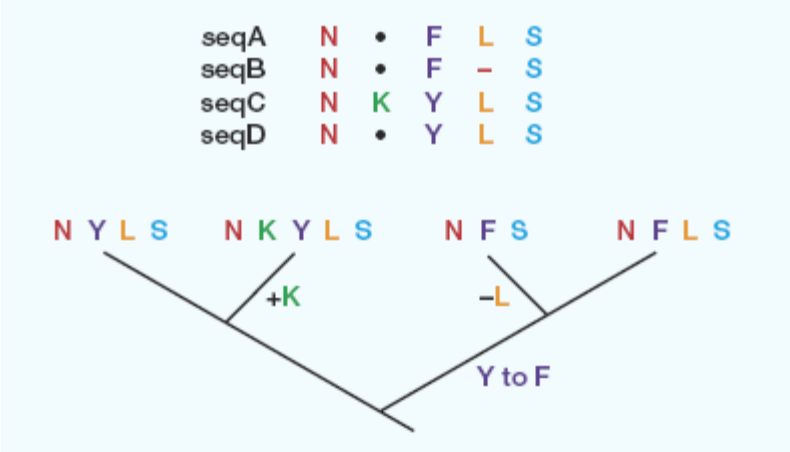

Figure 2: Multiple sequence alignment (top) of four sequences. There is no prior knowledge of which sequence is the ancestral, so in sequence C a K could be inserted or perhaps in the common ancestor the residue K existed was deleted in sequences A,B and D. The phylogenetic tree represents a possible evolutionary history that could generate the observed sequences from an ancestor sequence (Adapted from Mount, 2001).

Another very important application of MSAs is their incorporation in many methods of predicting the structure or function from sequence. These methods are a major contribution of bioinformatics to experimental research. The rate of known sequences is increasing, while other information such as their function is lagging behind. Many methods incorporate information from MSAs to improve their predictions. Such applications are pairwise sequence alignments using structural data (Marti-Renom et al., 2004), recent methods for the prediction of protein secondary structure (Predict Protein, PsiPred, jpred, Baxevanis & Ouellette, 2001) and gene prediction from comparison of sequenced genomes (Brent & Guigó, 2004).

1.3 The need of a benchmark study

The number of different MSA methodologies has greatly increased during the last years resulting to approximately 30 programs today. On the contrary only a few of them are used routinely by biologists. The reasons are many, but the main one is the lack of a consistent theoretical framework in sequence analysis (Notredame 2002). As a consequence "*programs with badly designed interfaces or poor portability have been discarded by natural selection, leaving their algorithms to be reinvented by later generations*" (Notredame 2002). Furthermore, most benchmarks are conducted from researchers that introduce a novel algorithm, and compare only the minority of most used programs. Therefore, a comparison study for the different MSA programs that are offered in the web is necessary, not only for the new scientists that enter the Bioinformatics area, but also for the biologists and bioinformaticians. A more wide and detailed knowledge of all the currently available methods helps scientists to use the proper software that interprets better their biological data and corresponds to their specific biological problem.

2. Methodology

In order to compare the Multiple Sequence Alignment Programs and have a full view of their capabilities, a two-stage comparison process was applied: the "high level" and the "low level" comparison. The former one includes comparisons of the interfaces, the portability, the functionalities and the parameterizations that each of the programs offers, all of them affecting the usability of the program and therefore, its popularity among the users. The latter one compares the "heart" of the programs, the algorithms, that defines the quality and the biological meaning of their results. Therefore, the researcher can choose the program that best fits his/her needs.

2.1 Comparison of functionality

The comparison in this stage focuses on both the front end of each program and its usability. The comparison criteria include the user-friendliness of the web interfaces, the existence of documentation and manual, the portability, the functionality and the parameterization.

Most of the MSA programs are offered through web interface, while many of them are also available for download. If there is a need for a batch mode processing, it is necessary to download and install the programs locally. This task most of the times is easy, even for non-computer specialists since they have only to follow the instructions described in a readme file. In rare cases, the batch processing may be time consuming since it may need extra effort from computer specialists to process the data automatically.

The results of a query in the web interface may be viewed at once in the navigator window, or an e-mail with the results is sent to the user. In both cases the time needed to get the results depends on both the processing time and the traffic in the server. The requests from the users usually are served in a FIFO (First In – First Out) order.

Before using a MSA program, it is necessary to be aware of its functionality and its limitations. Many of the MSA programs can align not only multiple sequences, but also multiple alignments, or a sequence with a multiple alignment. The limitations of a MSA program include the maximum number of sequences that can accept in a query or also the maximum length of the sequences. So, in the case of large number of sequences (~100) an appropriate MSA program should be used. Another limitation that one has to take into consideration is the format of the input and output files of the MSA program. Nowadays, many of the programs accept the input or output file in many formats such as FASTA, MSF, ALN, UniProt, PIR. Some of the programs output is in a specific format that is used only by the specific program. As a result, a conversion to a more common format (e.g. MSF) is necessary for further processing of the data. There are available on the web programs that convert one format to another (e.g. readseq http://thr.cit.nih.gov/molbio/readseq/).

Another important factor for MSA programs is how many parameters the user is able to control. In most MSA programs the user can select the substitution matrices used (BLOSUM, PAM etc). The freedom the user to select other parameters (like the window sizes, the weights, the cost for gap open, the cost for gap end etc) is very important, because the user is able to customize the alignment to perform better for his biological data or to compare the effects of the various parameter to the alignment. For a newbie the

existence of many parameters may be a source of confusion, but the use of the default values is the best solution.

The comparison of computational time of the MSA programs was not included in this comparison study, because it is not a crucial factor for their popularity.

2.2 Comparison of algorithms

In order to compare the MSA programs in more depth there is a need to study the algorithms. The performance of the algorithms and the quality of the results is evaluated by comparing the results of a specific MSA program with a "correct" result. But, which multiple alignment is considered as correct? The subjectivity of the reliability of a correct alignment has led to the creation of reference datasets which contain multiple alignments that are generally accepted. And, how two multiple alignments are compared? The quantitative analysis of a comparison between a test multiple alignment and a reference one uses some scores that may are dependent on the alignment itself (e.g. Sum of Pairs, Total Column Score) or independent from it. In the next paragraphs there is a detailed description of the reference datasets, the algorithms and the evaluation scores used in MSA programs.

2.2.1 Reference Datasets

BAliBASE is the most known and general accepted database of aligned sequences. This database contains high quality, manually constructed multiple sequence alignments that are all based on three-dimensional structural superpositions. In each multiple alignment in the database a core block is annotated, which includes only the regions which can be reliably aligned. The first version provided sets of reference alignments dealing with the problems of high variability, unequal repartition, large N/C terminal extensions and internal insertions. The second version incorporates three new reference sets of alignments containing structural repeats, transmembrane sequences and circular permutations. An improved version has recently become available BAliBASE 3.0 with more reference alignments.

Another reference database is the Sequence Alignment Benchmark (SABmark) that contains alignments that cover the entire known fold space, as classified by SCOP. To limit the impact of highly abundant folds, each alignment contains at most 25 sequences. Two alignment sets are available, Twilight Zone and Super-families, which represent sequences with respectively very low to low, and low to intermediate similarity. These are based on subsets provided by the ASTRAL compendium and correspond roughly with 0-25% and 0-50% identical residues. Alignments of sequences with higher similarities are not provided, since the performance of most algorithms if already very good above 50% identities. Since many alignments are performed exactly to determine whether or not sequences are related, a second version of both the Twilight Zone and the Superfamilies set is given that addresses this issue: to each group of sequences to be aligned, the same number of 'false positive' sequences (sequences that belong to a different fold) is added.

Both BAliBASE and SABmark databases include mainly global multiple sequence alignments. In order to study the performance of local MSA programs there have been constructed artificial random sequences with implanted conserved motifs, such

as Lassmann and Sonnhammer's database (2002) and IRMbase (Subramanian et al., 2005)

2.2.2 Algorithms

Two sequences can be aligned either globally or locally, depending on the purpose of study. The global alignment tries to align the sequences at their entire length; therefore it is mostly used for sequences with high similarity in their whole length. Needleman-Wunsch algorithm uses dynamic programming to align globally two sequences allowing the insertion of gaps. Trying to align the sequence at the whole length, may lead to mismatch at local areas. For example, a global alignment of two proteins that share a common domain restricted in their N-terminal region will use information of the whole length of the proteins and perhaps the domains will not be aligned correctly. Smith-Waterman algorithm provides solution to this problem using dynamic programming with an extra condition that result to local alignments. Dynamic programming guarantees the mathematically optimal alignment for a specific score function that is needed to be maximized.

In the case of more than two sequences, dynamic programming algorithms can be extended to a multi-dimensional space; therefore the computational time and memory needed is prohibitive to use for more than three sequences. So, even though dynamic programming algorithms offer the optimal alignment, practically their use is not feasible with the provided current technology and computers. That is why new heuristic approaches have been developed using different strategies.

Exact Algorithms: These algorithms are high quality heuristics that find multiple alignments very close to the optimal. They are based on dynamic programming algorithms, but they exclude from the computation the portion of the multidimensional space that does not contribute to the solution. In this way the computational time and memory becomes less, offering a less optimal multiple alignment solution. A program called MSA implements this approach and manages to align up to ten closely related sequences in a reasonable computational time (Mount 2001). DCA (Stoye 1997) is a divide and conquer algorithm that uses MSA. DCA algorithm cuts the sequences in subsets of segments that are small enough to be fed to MSA. The critical issue is to cut the sequences at the right points so that the produced alignments remain as close as possible to optimal. DCA manages to align up to 20 – 30 closely related sequences. In the next years with the increase in the computational speed, all the above limitations may not be prohibitive for practical use of exact algorithms in everyday research.

Progressive Algorithms: These algorithms are the most widely used, since they can align multiple sequences in little time and with little memory. Their basic idea is that the final multiple alignment is the result of progressive building upon the alignment of two sequences (or multiple alignments). This means that a progressive assembly of the multiple alignment takes place where the sequences or the alignments are added one by one so that never more than two sequences (or multiple alignments) are simultaneously aligned using dynamic programming. The order of the sequences that are added to the alignment is indicated by a pre-computed tree, which is computed by aligning pair-wise all against all the sequences. To summarize the whole procedure:

- Align pair-wise all against all the sequences.
- Construction of distance matrix using the pair-wise alignment scores.

- Creation of a distance tree.
- Align the two closest sequences. To this alignment add the next closest sequence (or the next closest alignment) and align. Continue with progressive alignments where sequences are added to the multiple alignment according to the order indicated by the tree.

The most widely used progressive programs are ClustalW (Thompson et al., 1994) and T-Coffee (Notredame et al., 2000).

Iterative Algorithms: Iterative alignment methods depend on algorithms able to produce an alignment and to refine it through a series of cycles (iterations) until no more improvements can be made. Iterative methods can be deterministic or stochastic, depending on the strategy used to improve the alignment. The simplest iterative strategies are deterministic. They involve extracting sequences one by one from a multiple alignment and realigning them to the remaining sequences some of these methods can even be a mixture of progressive and iterative strategies. The procedure is terminated when no more improvement can be made (convergence). Stochastic iterative methods include HMM training and simulated annealing or genetic algorithms (Notredame 2002). Widely used iterative programs are Praline (Simossis et al., 2003), PRRP and SAGA (Notredame et al., 1996).

In this study ClustalW and TCoffee were tested as representative programs for global alignments, while DiAlign-T was tested as a representative of local alignment programs.

2.2.3 ClustalW

ClustalW (Thompson et al., 1994) follows a progressive approach to multiple sequence alignment and performs global alignments. The default procedure is the following:

Initially the program aligns pair-wise all-against-all the sequences using a heuristic approach (FastA algorithm). The scores of each alignment, which indicates similarity is transformed to a distance measure, and a distance matrix is constructed. The next step is to construct a tree from the distance matrix with the Neighbour-Joining method. This method constructs unrooted trees, so a root is placed in the middle of the largest branch. For each sequence a weight is computed in order to avoid the bias towards a group of very similar sequences. The weights depend on the distance of the sequence from the root, but sequences which have a common branch with other sequences share the weight derived from the shared branch. Using the tree as a guide, progressively the sequences are aligned with dynamic programming. Similar sequences are aligned first, while less similar are aligned later. During the progressive alignment a complicated function for gap penalties is used. The gap opening and gap extension penalties depend on the weight matrix used, on the similarity of the aligned sequences, the sequences lengths and local information such as conserved residues, the existence of gaps next to the position under consideration etc.

2.2.4 T-Coffee

T-Coffee (Notredame et al., 2000) is a progressive multiple sequence alignment program which combines information from global and local alignments to produce a global multiple sequence alignment. Initially two primary libraries of pair-wise

alignments are constructed. The alignments are local in one library and global in the other. Local alignments are constructed by Lalign, while global are constructed by ClustalW (for two sequences). Each residue pair in the primary library is assigned a weight based on the aligned sequences identities. The two primary libraries are combined and the weight of each residue pair is adjusted. The next step is a heuristic algorithm called library extension which incorporates the information of the whole library to conclude if each residue pair is consistent as part of the library. A progressive alignment with dynamic programming follows using the weights computed from the library extension.

2.2.5 Dialign-T

Dialign-T (Subramanian et al., 2005) is a segment based program for multiple sequence alignments. It is a re-implementation of Dialign2 with some improvements. Initially Dialign-T identifies very similar segment pairs that could not be caused by chance and assigns them a weight. Another score is computed for each segment pair from its consistency with the complete set of segment pairs. The multiple sequence alignment is assembled by discarding segment pairs which are not consistent with the rest. Some improvements which enhance the performance of Dialign-T are the exclusion of segment pairs that include low-scoring subfragments, the way inconsistent fragments are excluded and improvements in calculating the segment pair weight.

2.2.6 MSA assessment scores

The evaluation of the accuracy of a multiple alignment is performed in one of the following approaches (Raghava 2003):

(i) Dependent measures. These measures compare an alignment to a reference alignment. Their accuracy depends on the quality of the reference alignment. Dependent measures can assess the multiple alignment either as considering the alignment as a whole, or by examining the quality of each pairwise alignment within the multiple alignment. A complementary approach is to examine the accuracy only of conserved regions of the alignment.

(ii) Independent measures. These measures assess the score of a structural superposition implied by the multiple alignment.

(iii) Visualization tools that highlight differences between alignments and expert opinion.

In our study we used only dependent measures, since the reference alignments used are constructed using structural information (approach ii) and were refined by experts (approach iii) (Thompson et al., 1999a, Bahr et al., 2001).

Ideally a score of a multiple alignment should reflect its likelihood according to a given evolutionary model (Dured & Abdeddaim, 2000). The computation of this score is critical in determining how accurately the resulting alignment reproduces the evolutionary history of the sequences (Nicholas, et al., 2002).

2.2.7 Statistical analysis

Statistical analysis of MSA programs performance usually involves the comparison of scores achieved in a specific reference test by the programs, or by the

scores of a specific MSA program in different reference tests. In both cases there is no prior knowledge about the scores distribution, so it is necessary to use nonparametric statistical test, although there have been published comparisons using parametric tests (Katoh et al., 2002).

3. Materials and methods

3.1 Datasets

In the experiments the last version of the BAliBASE database, BAliBASE 3 (http://www-bio3d-igbmc.u-strasbg.fr/balibase/) was used. It consists of 218 reference alignments with more than 2100 sequences and is divided into 8 different reference sets. In this study References 1-5 were used.

Reference 1 (RV11 & RV12): Contains equi-distant sequences with 2 different levels of conservation. The dataset RV11 has the very divergent sequences (<20% identity) with 38 multiple alignment files, whereas the RV12 set has medium to divergent sequences (20-40% identity) with 44 multiple alignment files.

Reference 2 (RV20): Contains the families aligned with a highly divergent "orphan" sequence and has 41 multiple alignment files.

Reference 3 (RV30): Contains subgroups with <25% residue identity between groups and has 30 multiple alignment files.

Reference 4 (RV40): Contains sequences with N/C-terminal extensions and has 49 multiple alignment files.

Reference 5 (RV50): Contains sequences with internal insertions and has 16 multiple alignment files.

References 6, 7 & 8: Contain alignments with structural repeats, transmembrane sequences and circular permutations.

3.2 Software

For the comparison of the functionality, 15 out of the approximately 30 MSA programs were used (see Table 1). In order to compare the algorithms, ClustalW and TCoffee were tested as representative programs for global alignments. These two programs were chosen because they are the most widely used at the time of the comparison study and they appear to have the best performance in various review papers. As local alignment program DiAlign-T was tested. Both programs were tested using their default parameters, since these parameters are calibrated from their creators to give the optimal performance.

ClustalW version 1.82 was tested, which uses the below default values for its parameters:

- DNA Gap Open Penalty = 15.0
- DNA Gap Extension Penalty = 6.66
- DNA Matrix = Identity
- Protein Gap Open Penalty = 10.0
- Protein Gap Extension Penalty = 0.2
- Protein matrix = Gonnet
- Protein/DNA ENDGAP = -1

- Protein/DNA GAPDIST = 4

Tcoffee version 2.66 was tested, which uses the default values of ClustalW parameters.

The default parameters for Dialign-T are:

- Length of a low-scoring region = 4
- Maximum fragment length that is allowed to contain regions of low quality =40

3.3 MSA assessment scores

The performance of MSA programs was assessed using the sum-of-pairs score and the column score as implemented in the BaliScore program. (Thompson et al., 1999b):

Sum-of-pairs score (SPS). Suppose we have a test alignment of N sequence consisting of M columns. We can designate the ith column in the alignment by $A_{i1,}$ A_{i2},…A_{iN}. For each pair of residues A_{ij} and A_{ik}, p_{ijk} is defined such that $p_{ijk} = 1$ if residues A_{ij} and A_{ik} are aligned with each other in the reference alignment, otherwise $p_{ijk} = 0$. The score S_i for the ith column is defined as:

$$S_i = \sum_{j=1,\, j\neq k}^{N} \sum_{k=1}^{N} p_{ijk}$$

The SPS for the alignment is:

$$SPS = \frac{\sum_{i=1}^{M} S_i}{\sum_{i=1}^{Mr} S_{ri}}$$

Where Mr is the number of columns in the reference alignment and S_{ri} is the score S_i for the ith column in the reference alignment.

Column score (CS): Using the same symbols as above, the score C_i of the ith column is equal to 1 if all the residues in the column are aligned in the reference alignment, otherwise is equal to 0. Therefore the column score is:

$$CS = \frac{\sum_{i=1}^{M} C_i}{M}$$

The assessment scores were computed only for the core blocks of the reference set alignments since these regions are reliably aligned (Thompson et al, 1999a). According to Thompson et al. (1999a) the SPS score is preferable for reference sets 1 and 2 while CS for the rest.

3.4 Statistical analysis

For the comparison each methods scores for a specific reference set Friedman test is used (Edgar, 2004, Katoh et al., 2005. Thompson et al., 1999b), which does not make any assumption for the underlying distribution and the only condition is that the samples should have the same size (Lioki-Leivada & Asimakopoulos, 2002).

4. Results

4.1 Comparison of functionality

Table 1 summarizes the values of the comparison criteria as described in section 2.1. The programs are listed in alphabetical order.

4.2 Comparison of algorithms

In the following tables (Tables 2 & 3) summarizes the results of the Friedman test computed for the resulting data. The values of the tables are represented in bar chars (Figure3 & 4).

SPS	ClustalW	T-coffee	Dialign-T
RV11	2,71	2,16	1,13
RV12	2,07	2,89	1,05
RV20	2,06	2,88	1,06
RV30	2,64	2,00	1,36
RV40	2,74	1,96	1,30
RV50	2,81	2,13	1,06

Table 2: Frieman test indicates a statistically significant difference ($p<0,05$) in the sum-of pairs score. The numbers indicate mean rank.

CS	ClustalW	T-coffee	Dialign-T
RV11	2,37	2,09	1,54
RV12	2,26	2,53	1,20
RV20	2,44	1,68	1,68
RV30	2,61	2,18	1,21
RV40	2,46	1,99	1,55
RV50	2,47	2,00	1,53

Table 3: Frieman test indicates a statistically significant difference ($p<0,05$) in the column score. The numbers indicate mean rank.

Program	Web	Stand Alone	Results (I/E)[1]	Max # Sequences	Max $# Sequences[2]	Scoring Matrices	Input Format	Output Format	Alignment /Algorithm[3]	URL
ClustalW	yes	yes	I/E	500/10MB		BLOSUM PAM GONNET ID	NBRF/PIR EMBL / UniProt FASTA,GDE,GCG/MSF ALN/ClustalW, RSF	ALN, GCG, PHYLIP, PIR, GDE	Global /Progressive	http://www.ebi.ac.uk/clustalw/
T-Coffee	yes	yes	I/E	50	10,000	BLOSUM PAM GONNET ID	FASTA	CLUSTAL PIR, GCG FASTA, MSF PHYLIP	Global /Progressive	http://igs-server.cnrs-mrs.fr/~cnotred/Projects_home_page/t_coffee_home_page.html
MatchBox	yes	(yes)	E	50	2000	BLOSUM PAM GONNET	FASTA MSF HSSP	FASTA MSF HSSP	Local /Progressive	http://www.sciences.fundp.ac.be/biologie/bms/matchbox_submit.shtml
Praline	yes	(yes)	E	500	2000	BLOSUM PAM,GONNET	FASTA	MSF FASTA	Global /Progressive	http://ibivu.cs.vu.nl/programs/pralinewww/
DiAlign-T	yes	yes	I/E				FASTA	DiAlign FASTA	Global /Iterative	http://dialign-t.gobics.de/
Pima	yes	yes	I		20,000		NBRF/PIR EMBL / UniProt FASTA,GDE,GCG/MSF ALN/ClustalW, RSF	NBRF/PIR EMBL / UniProt FASTA,GDE,GCG/MSF ALN/ClustalW, ,RSF	Local /Progressive	http://searchlauncher.bcm.tmc.edu/multi-align/Options/pima.html
DCA	yes	yes	I/E	500MB		BLOSUM PAM, GONNET	GDE FASTA	FASTA NEXUS	Global /Exact	http://bibiserv.techfak.uni-bielefeld.de/dca/submission.html
POA	[yes]	-					FASTA	POA	Global	http://www.hgmp.mrc.ac.uk/Registered/Option/poa.html
MAFFT	yes	yes	I/E	2000		BLOSUM JTT	FASTA	FASTA	Global /Iterative	http://www.biophys.kyoto-u.ac.jp/%7Ekatoh/programs/align/mafft/
BALI	-	yes	I			BLOSUM PAM, GONNET	FASTA,GDE ALN/ClustalW GCG/MSF,RSF	FASTA,GDE ALN/ClustalW GCG/MSF,RSF	Local /Iterative	http://xray.bmc.uu.se/dennis/
MultiAlign	yes	yes	I				SwissProt, Accession Number, Sequences with or without gaps	FASTA MSF	Global /Iterative	http://cbrg.inf.ethz.ch/Server/MultAlign.html

Table 1: Summarizes the values of the comparison criteria, as described in section 2.1
1: This field indicates whether the results of a query are obtained instantly (I) through the web interface, or are sent via e-mail to the user (E)
2: This field indicates the maximum length of the query sequences
3: This field indicates whether the alignment is Global or Local, and whether the algorithm is progressive, iterative or exact

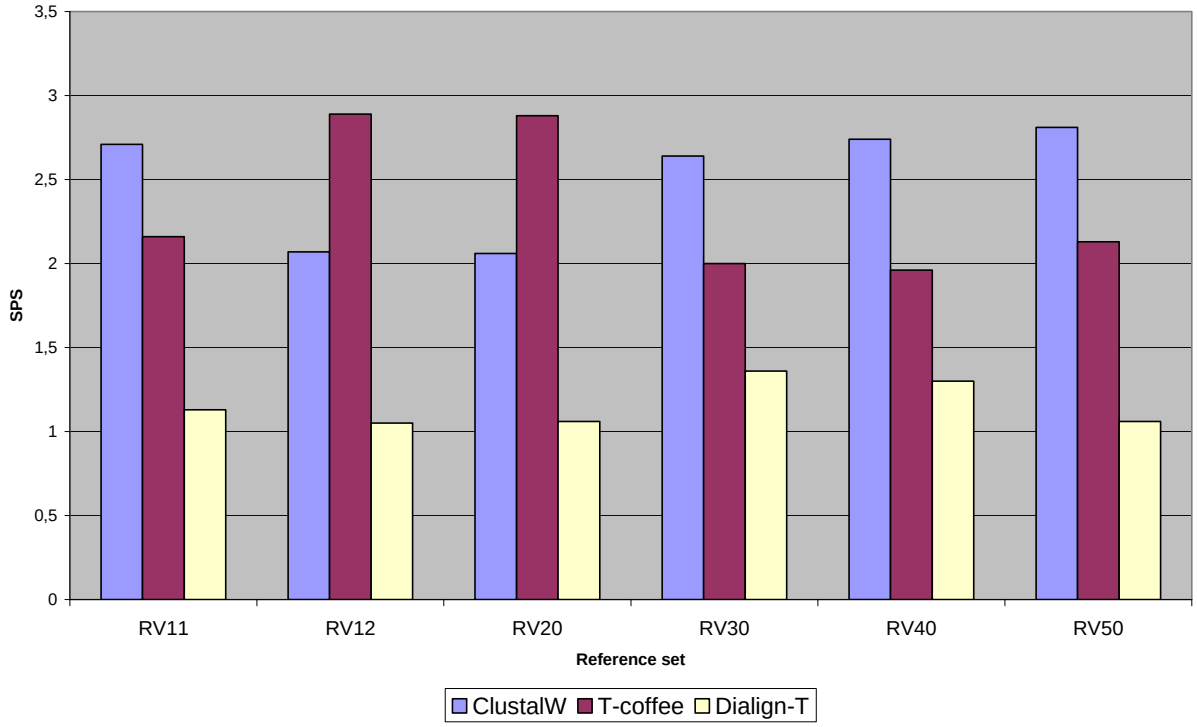

Figure 3: Bar chart of the mean rank computed by the Friedman test on the SPS for each reference test (data from Table 2)

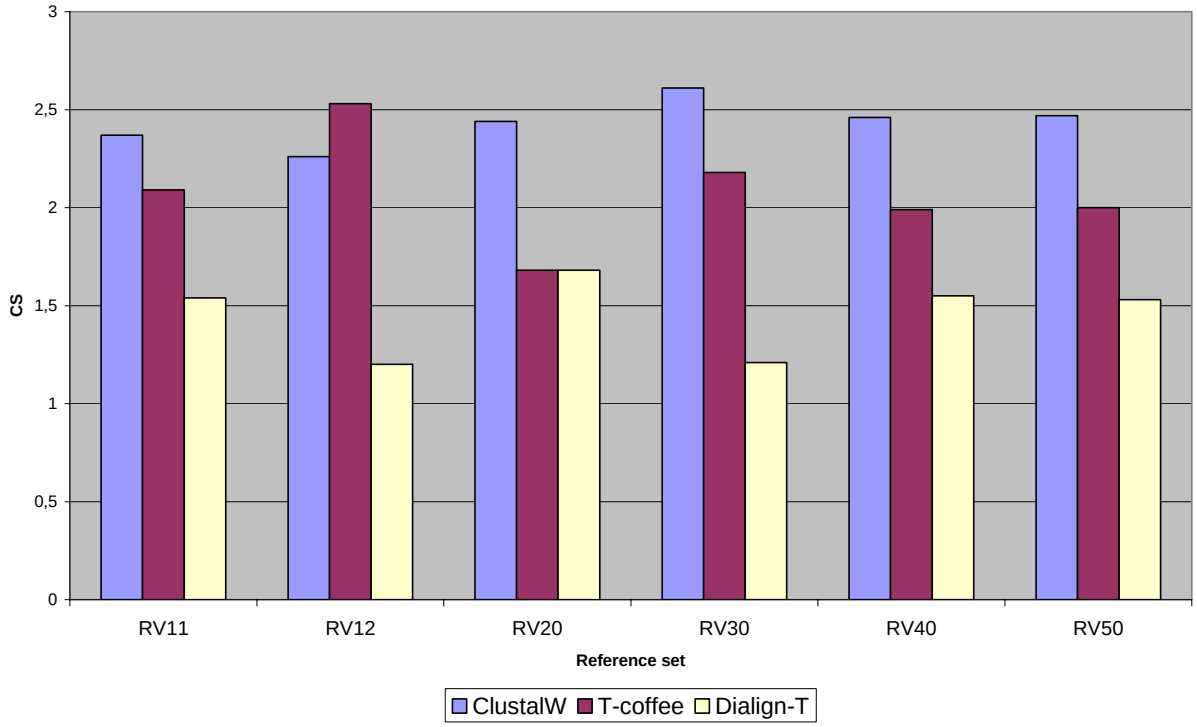

Figure 4: Bar chart of the mean rank computed by the Friedman test on the CS for each reference test (data from Table 3)

5. Discussion

5.1 Comparison of functionality

For the purposes of this comparison study, many of the available MSA programs were carefully examined. We attempted to download most of the programs to use for batch processing. Some of them were downloadable and easy to install such as ClustalW and T-Coffee. In many cases the programs were accessible for download only after request from the creators. But request is not always sufficient to obtain the program. In some cases such as POA two days are needed to gain access for using the web interface. Generally, many MSA programs are accessed via web, and a small portion of them is available for download for various operating systems (Unix, Linux, Window, MacOs).

Newest versions and programs offer a greater flexibility to the user allowing for different I/O file formats, various scoring matrices etc.

5.2 Comparison of Algorithms

As shown in Figures 3 & 4, Dalign-T achieves the lowest scores. This is explained by the bias of BAliBASE towards globally related sequences. The multiple sequence alignments included are well known protein sequences with more than 50% of their residues reliable aligned. Furthermore, some of the sequences are truncated keeping only the aligned regions.

Considering the SPS for reference sets RV11, RV12 and RV20 and the CS for the rest, T-Coffee seems to make better alignments for reference sets RV12 and RV20, while ClustalW achieves better scores for the rest.

In reference set RV11 the aligned sequences are diverged (<20% identities) while RV12 the sequences share a greater similarity (20%-40% identities). From the figures it is obvious that ClustalW outperforms T-Coffee in the difficult test set (RV11). This could probably caused by the incorporation of information from local alignments in T-Coffee, which misleads the global alignment, since BAliBASES is biased towards global alignments.

In the reference set RV12 T-Coffee achieves higher SPS compared to ClustalW. Perhaps the refinement of the alignments made by T-Coffee is better when sequences with great similarity are used.

The reference set RV20 includes alignments of a group of similar sequences with an outlier sequence. Observing the SPS T-Coffee achieves higher scores compared with ClustalW, while the opposite holds for CS. This could be explained because T-Coffee aligns well the group of highly similar sequences, but not the outlier, while ClustalW aligns correctly more columns.

In reference sets RV30, RV40 and RV50, ClustalW achieves higher SPS and CS than T-Coffee. Generally ClustalW seems to perform better with difficult alignments.

Recent Benchmark studies comparing the studied programs use different comparison schemes. A comparison of our results with benchmarks included in the publications of Dialign-T algorithm (Subramanian et al., 2005) and T-Coffee (Notredame et al., 2000) is shown in figures 5, 6 & 7 respectively.

The Benchmark studies used previous versions of BAliBASE: version 1 (1999) was used by the T-Coffee benchmark and version 2 (2001) was used by the Dialign-T benchmark. Our study uses the newest version of BAliBASE (version 3, 2005).

Moreover, the T-Coffee benchmark did not include Dialign-T but its predecessor Dalign2.

As is shown in figure 7, in the T-Coffee benchmark T-Coffee achieves the higher score in all reference sets. In figures 5 and 6 (Dialign-T benchmark) Dialign-T achieves the highest score in reference set RV40 while the scores of the rest reference sets are comparable with the scores of ClustalW and T-Coffee. These differences could be caused by differences in the reference data sets, differences in estimating assessment scores, or the use of different versions of the programs.

In conclusion, these inconsistencies in benchmark studies reveal the necessity of a carefully designed benchmark using truly representative reference sets and a variety of scoring schemes.

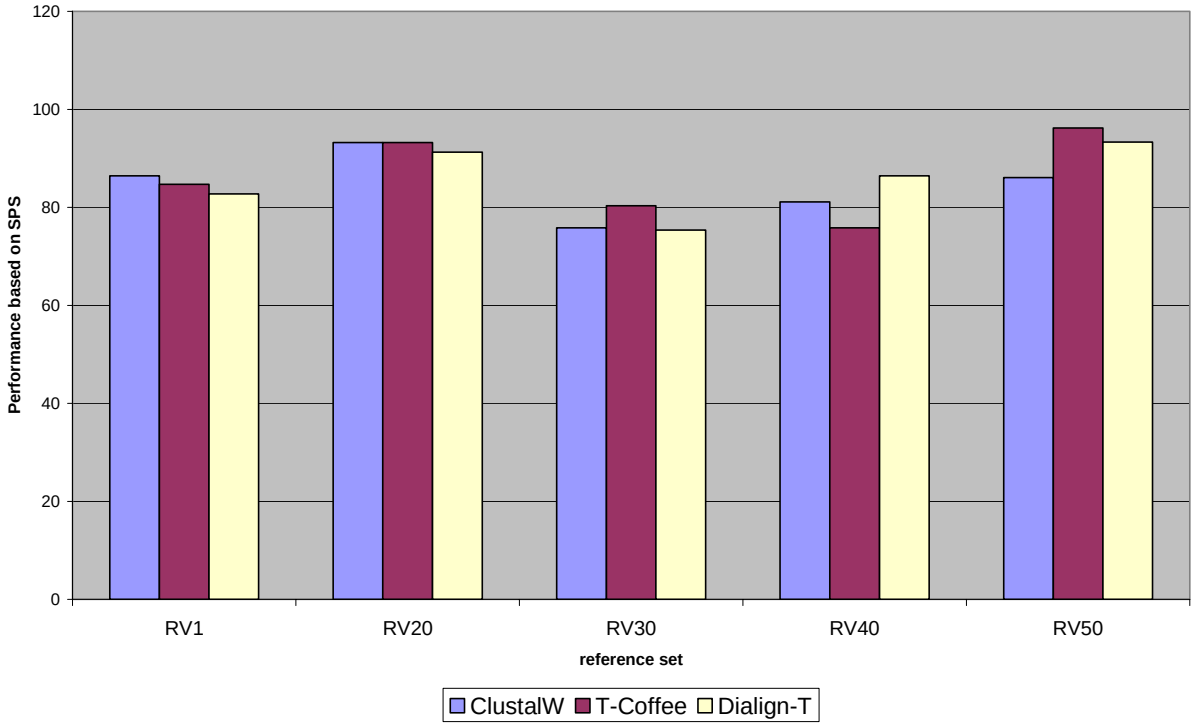

Figure 5: Bar chart of performance based on SPS, constructed from the benchmark data published with the algorithm Dialign-T (Subramanian et al., 2005)

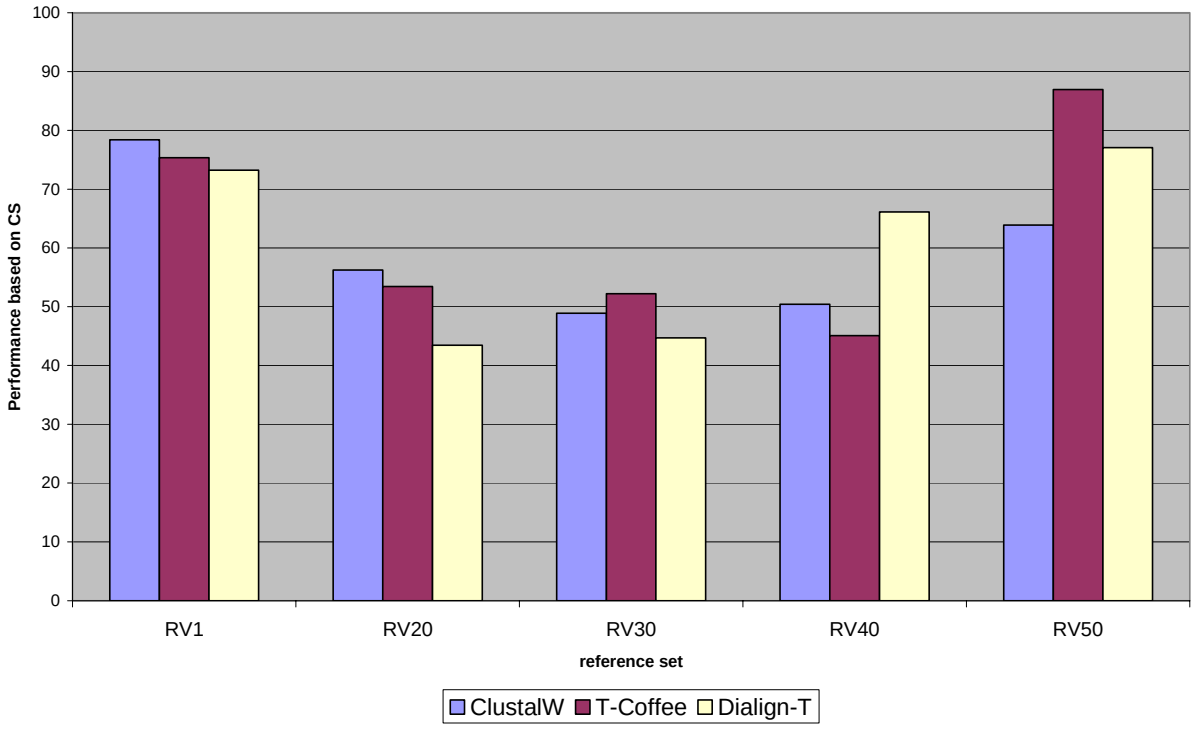

Figure 6: Bar chart of performance based on CS, constructed from the benchmark data published with the algorithm Dialign-T (Subramanian et al., 2005)

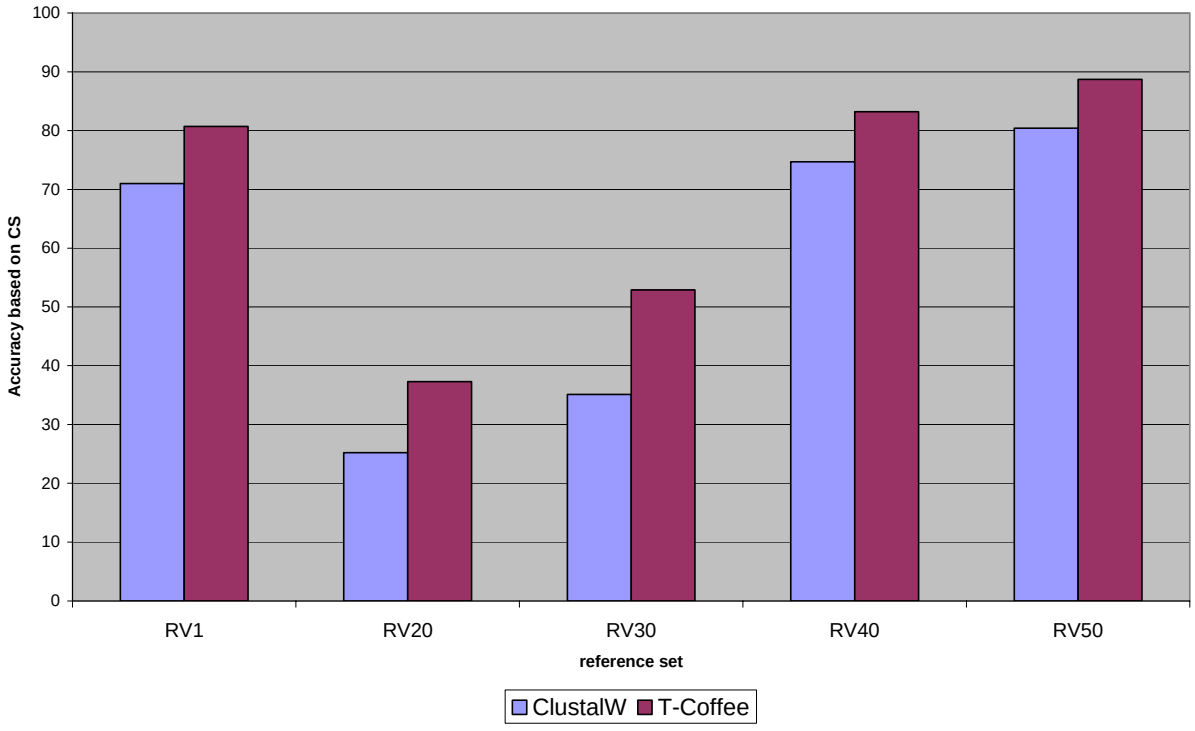

Figure 7: Bar chart of accuracy based on CS, constructed from the benchmark data published with the algorithm T-Coffee (Notredame et al., 2000)

6. Bibliography

Bahr, A., Thompson, J.D., Thierry, J-C., Poch, O. 2001. BAliBASE (Benchmark Alignment data BASE): enhancements for repeats, transembrane sequences and circular permutations. Nucleic Acids Research, 29(1):323-326.

Baxevanis, A., Ouellette, B.F.F. 2001. Bioinformatics : A practical guide to the analysis of genes and proteins (2nd ed). John Wiley & Sons, Inc.

Brent, M.R., Guigó, R. 2004. Recent advances in gene structure prediction. Current Opinion in Structural Biology, 14:264-272

Edgar, R. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research, 32(5):1792-1797.

Katoh, K., Kuma, K., Toh, H., Miyata, T. 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Research, 33(2):511-518.

Katoh, K., Misawa, K., Kuma, K,. Miyata, T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Research, 30(14):3059-3066.

Lioki-Leivada, I, Asimakopoulos, D.N. 2002. Introduction to applied statistics. V.1. University of Athens.

Marti-Renom, M.A., Madhusudhan, M.S., Sali, A. 2004. Alignment of protein sequences by their profiles. Protein Science 13:1071-1087

Mount, D.W., 2001. Bioinformatics, Sequence and Genome Analysis. Cold Spring Harbor Laboratory Press.

Nicholas, H.B., Alexander, J.R., Deerfield II, D.W. 2002. Strategies for Multiple Sequence Alignment. BioTechniques, 32(3):572-591.

Notredame, C., 2002. Recent progress in multiple sequence alignment: a survey. Pharmacogenomics 3(1):131-144.

Raghava, G.P.S., Searle, M.J., Audley, P.C., Barber, J.D., Barton, G.J. 2003. OXBench: A benchmark for evaluation of protein multiple sequence alignment accuracy. BMC Bioinformatics 4:47.

Thompson, J.D., Plewniak, F., Poch, O. 1999a. BAliBASE: a benchmark alignment database for the evaluation of multiple alignment programs. Bioinformatics. 15(1):87-88

Thompson, J.D., Plewniak, F., Poch, O. 1999b. A comprehensive comparison of multiple sequence alignment programs. Nucleic Acids Research, 27(13):2682-2690.

Hulo, N., Sigrist, C.J.A., Le Saux, V. Langendijk-Genevaux, P.S, Bordoli, L., Gattiker, A., De Castro, E., Bucher, P, Bairoch, A. 2004. Recent improvements to the PROSITE database. Nucleic Acids Research, 32 (Supplement 1):134-137

Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E.L.L, Studholme, D.J., Yeats, C., Eddy, S. 2004. The Pfam protein families database. Nucleic Acids Research. 32 (Suppplement 1):138-141

Duret, L., Abdeddaim, .S. 2000. Multiple alignment for structural, functional or phylogenetic analyses of homologous sequences. In: Higgins, D., Taylor, W., (eds) Bioinformatics, sequences, structures and databanks. Oxford University Press.

Lassmann T, Sonnhammer EL: Quality assessment of multiple alignment programs. FEBS Letters 2002, 529:126-130.

Subramanian, A.R., Weyer-Menkhoff, J., Kaufmann, M., Morgrenstern, B., 2005. DIALIGN-T: An improved algorithm for segment-based multiple sequence alignment. BMC Bioinformatics, 6:66.

Stoye, J. 1997. Divide-and-Conquer Multiple Sequence Alignment. Dissertation Thesis. University of Bielefeld, Forschungsbericht der Technischen Fakultat, Abteilung Informationstechnik

Simossis, V.A., Heringa, J. 2003. The PRALINE online server: optimising progressive multiple alignment on the web. Computational Biology and Chemistry 27 (2003) 511–519.

Notredame, C., Higgins, D.G. 1996. SAGA: sequence alignment by genetic algorithm. Nucleic Acids Research 24(8): 1515–1524.

Thompson, J.D., Higgins, D.G., 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(22):4673-4680.

Notredame, C., Higgins, D.G, Heringa, J. 2000. T-Coffee: a novel method for fast and accurate multiple sequence alignment. J. Mol. Biol. 302:205-217.