Bioinformation by Biomedical Informatics Publishing Group *open access*
www.bioinformation.net **Hypothesis**

A comparison of MSA tools

Nadia Essoussi[1], Khaddouja Boujenfa[1,*] and Mohamed Limam[1]

[1]LARODEC, High Institute of Management, University of Tunis, Tunis, Tunisia; Khaddouja Boujenfa* - E-mail: Khaddouja.Boujenfa@isg.rnu.tn; * Corresponding author

received May 15, 2008; revised June 13, 2008; accepted July 06, 2008; published July 31, 2008

Abstract:
Multiple sequence alignment (MSA) is essential in phylogenetic, evolutionary and functional analysis. Several MSA tools are available in the literature. Here, we use several MSA tools such as ClustalX, Align-m, T-Coffee, SAGA, ProbCons, MAFFT, MUSCLE and DIALIGN to illustrate comparative phylogenetic trees analysis for two datasets. Results show that there is no single MSA tool that consistently outperforms the rest in producing reliable phylogenetic trees.

Keyword: multiple sequence alignment methods; phylogenetic trees; Robinson-Foulds distance; Neighbor-Joining method

Background:

Several multiple sequence alignment (MSA) methods are available in the literature. McClure and colleagues tested the ability of MSA methods to identify short motifs found in four datasets of homologous proteins **[1]**. Henikoff and Henikoff evaluated the ability of multiple alignments in identifying new family members in database search **[2]**. Thompson and colleagues presented a systematic analysis and comparison of several alignment programs using the BaliBASE reference alignments as test cases **[3]**. Despite these comparison studies, choosing an alignment method which produces the nearest phylogenetic test tree (TT) to the reference tree (RT) is still open for discussion. Multiple sequence alignment is a crucial step in phylogenetic analysis essentially for highly divergent data sets (<30% of sequence identity) that are difficult to align. Different methods produce non-identical alignments leading to variation in the constructed phylogenetic trees for a single dataset. Here, we perform the evaluation of eight alignment methods, namely ClustalX 1.81 **[4]**, Align-m 2.3 **[5]**, T-Coffee 3.93 **[6]**, SAGA 0.95 **[7]**, ProbCons 1.08 **[8]**, MAFFT 5.743 **[9]**, MUSCLE 3.6 **[10]** and DIALIGN 2.2.1 **[11]** to test their ability to generate similar phylogenetic test trees with respect to reference.

Methodology:

Datasets

We identified BaliBASE **[12]** and Homstrad **[13]** as reference alignment resources. The size of each reference alignment is fixed to more than four sequences.

Dataset #1: DS-BB

We selected 134 reference alignments from BaliBASE. The dataset is thereafter designated as DS-BB and divided into three categories according to the percent sequence identity within the reference alignment. Category 1: BB_10 contains 86 reference alignments at 0-10% sequence identity (ID) between each pair of sequences. Category 2: BB_20 contains 29 reference alignments at 10-20% ID. Category 3: BB_30 contains 19 reference alignments at 20-30% ID.

Dataset #2: DS-HOM

We downloaded 218 reference alignments from Homstrad. We created three different categories for this dataset similar to the first dataset. Category 1: HOM_10 contains 141 reference alignments. Category 2: HOM_20 contains 54 reference alignments. Category 3: HOM_30 contains 23 reference alignments. The dataset is thereafter designated as DS-HOM.

Comparison process

The eight alignment methods are run on DS-BB and DS-HOM datasets using default parameters. Tests were performed on a 1.6-GHz Intel Pentium M with 512 MB RAM. Each method generates a total of 352 test alignments: 134 (from DS-BB) + 218 (from DS-HOM). Thus, a total of 2816 (352*8) test alignments are obtained. The 352 test alignments of each method and the 352 reference alignments are given as input to the Neighbor Joining method described by Saitou and Nei, **[14]** to estimate phylogenetic TTs and RTs. Each 352 TTs of a given alignment method are compared to the 352 RTs.

The Robinson-Foulds distance (T_dRF) implemented in PAL **[15]** is used to compare a given phylogenetic TT to its corresponding RT. The T_dRF defines the distance between any two trees as the minimum number of transformations required to obtain the topology of one tree from the topology of the other. This is given by equation 1 in supplementary material. In order to evaluate the performance of each alignment method, we developed a score, namely the dRF(M), which considers only the identical TTs generated by each method compared to RTs. This is given by equation 2 under supplementary material. This score gives the average number of identical TTs produced by each method on each dataset category. High values of dRF(M) signify better performance by a method.

Alignment quality assessment

We used the sum-of-pairs score (SP) implemented in BaliBASE scoring scheme to estimate quality alignment for each method. The SP score determine the extent to which a method succeeds in aligning some or all sequences in the alignment. The aim here is to show if the alignment quality of a given method affects the reliability of its phylogenetic TT.

Discussion:

Biologists use MSA as a first step in phylogenetic analysis. A number of sequence alignment tools are available at the internet. However, the choice of a specific tool for a Biologist who is not an expert in the field of Bioinformatics is not trivial. Many comparison studies of multiple alignment methods are available **[1-3]**. These study lack arguments on phylogenetic analysis. Here, we evaluated eight MSA tools based on the comparison of their phylogenetic TTs. We use the Robinson-Foulds distance to compare the TTs of each alignment method with respect to the RTs. We derived the dRF(M) metric to estimate the percentage of identical TTs generated by each alignment method on each category of the two datasets used (DS-BB and DS-HOM). Figure 1 gives the variation of dRF(M) scores for all the eight methods used in the analysis. We notice that as sequence identity in each category of DS-BB and DS-HOM datasets is low as the percentage of identical TTs is low. All the methods show similar trends of dRF(M) scores. However, on categories BB_20 and BB_30 of DS-BB dataset, MUSCLE gives higher percentage of identical TTs than all the other methods. MUSCLE performs better on categories HOM_10 and HOM_30 in DS-HOM dataset.

We performed a Wilcoxon rank test for all pairs of methods (Table 1 under supplementary material) to assess the significance of the differences in the overall Robinson-Foulds distances (T_dRF) between all pairs of test and reference trees. Results suggest that the differences between methods are not statistically significant. Each method produces reliable phylogenetic TTs as those given by ProbCons, which is described by Do and colleagues **[16]** as the best performing method for generating accurate multiple alignment.

Figure 2 gives the variation of SP scores for all the methods on each category of DS-BB and DS-HOM datasets. It shows that ProbCons achieves the best performance on all the categories of each dataset. The significance in the difference for overall SP scores using the Wilcoxon rank test for all pairs of programs is given in Table 2 (supplementary material). The differences between methods are significant, with ProbCons showing the highest alignment quality. The results given in Table 1 and Table 2 (see supplementary material) suggest that quality alignment of the different methods do not heavily impact on the reliability of their phylogenetic TTs. It should be noted that all of them perform with good TTs as ProbCons.

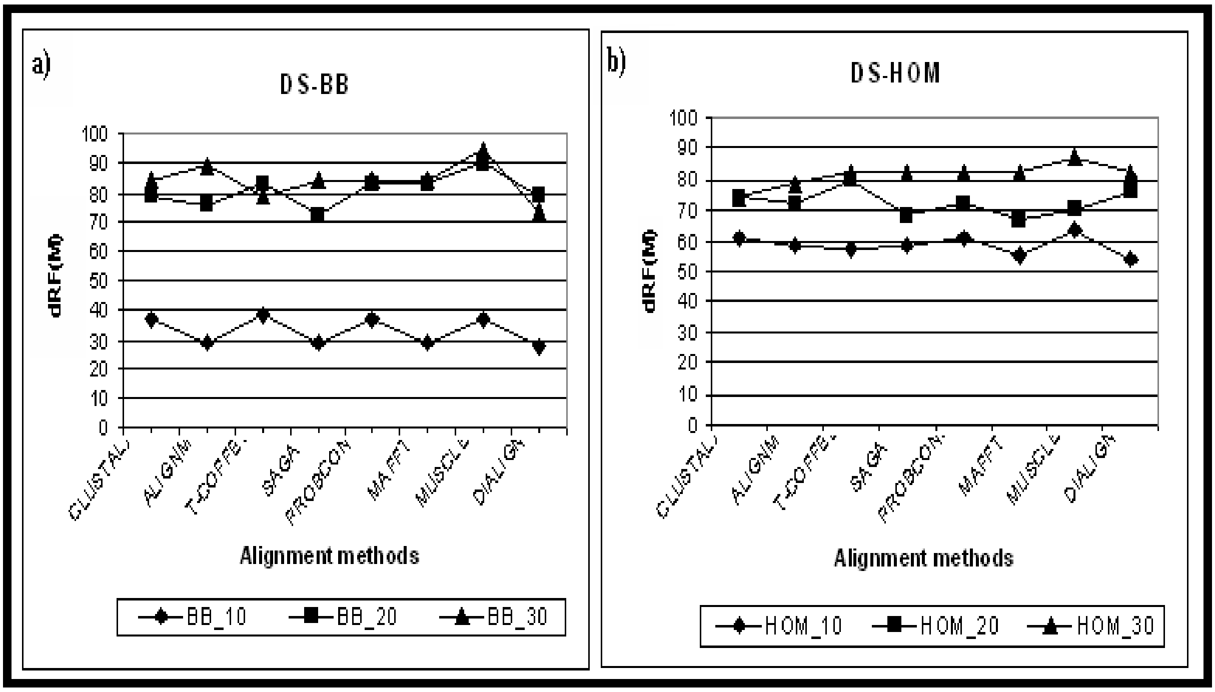

Figure 1: Performance of the eight alignment methods on dRF(M) scores for datasets DS-BB and DS-HOM is given. Line with markers indicates each data value. Values are given in percentage. All the methods show similar results. MUSLCE gives slightly higher performance on categories BB_20 and BB_30 of DS-BB dataset and on categories HOM_10 and HOM_30 of DS-HOM dataset.

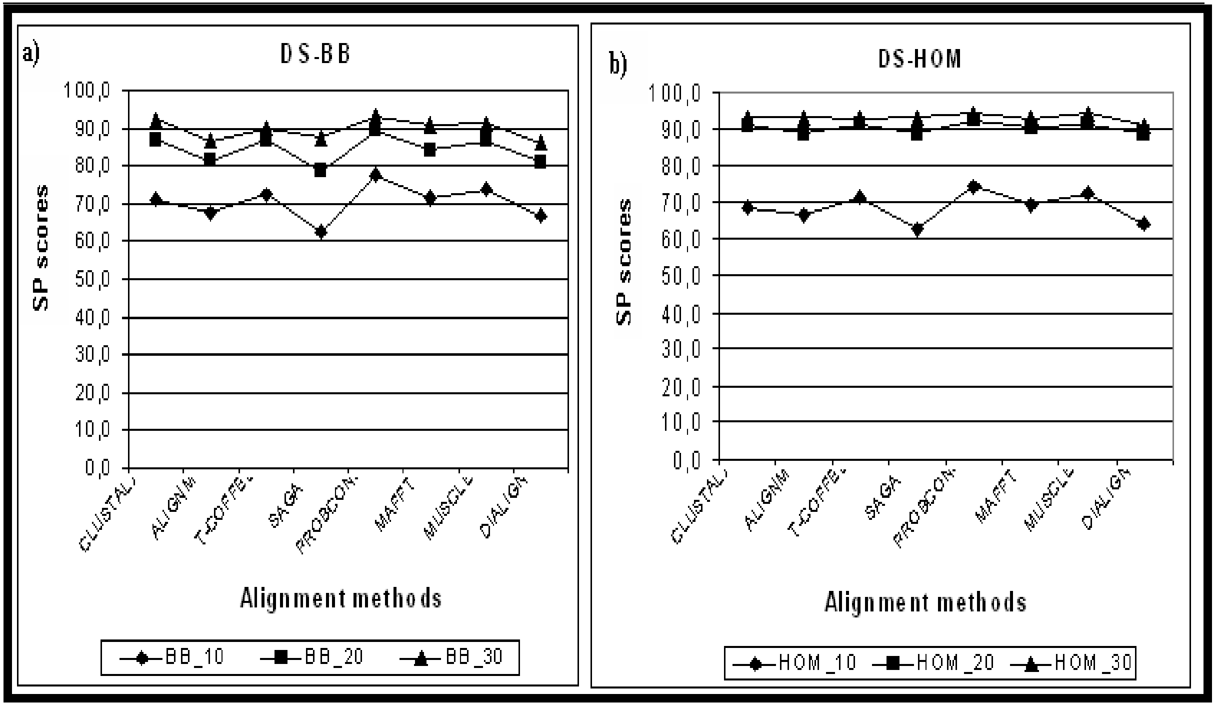

Figure 2: Performance of the eight alignment methods on SP scores for datasets DS-BB and DS-HOM is given. Line with markers indicates each data value. ProbCons shows higher SP scores than all the other methods on each category of DS-BB and DS-HOM datasets.

Conclusion:
A comparison of phylogenetic TTs of eight MSA for three categories of two sequence data sets is discussed. All methods perform equally well in producing reliable phylogenetic TTs. Despite the significant differences in alignments qualities produced by the different methods, the analysis shows that the statistical difference in phylogenetic TTs generated by each method is minimal. Several distances exist to compare trees, such as the Nearest-Neighbor interchange **[17]**. The application of the metric for large dataset would provide insights on MSA performances in divergent datasets.

References:
[01] M. A. McClure, *et al., Mol. Biol. Evol.*, 11: 4 (1994) [PMID: 8078398]
[02] S. Henikoff and J. G. Henikoff, *Protein. Sci.*, 6: 3 (1997) [PMID: 9070452]
[03] J. D. Thompson, *et al., Bioinformatics*, 15: 1 (1999) [PMID: 10068696]
[04] ftp://ftp-igbmc.u-strasbg.fr/pub/ClustalX/
[05] http://bioinformatics.vub.ac.be/
[06] http://www.tcoffee.org/Project_home_page/t_coffee_home_page.html
[07] http://www.tcoffee.org/Projects_home_page/saga_home_page.html
[08] http://probcons.stanford.edu
[09] http://www.biophys.kyoto-uac.jp/~katoh/programs
[10] http://www.drive5.com/muscle
[11] http://bibiserv.techfak.uni-bielefeld.de/dialign/
[12] ftp://ftp-igbmc.u-strasbg.fr/pub/BAliBASE2/
[13] ftp://dalek.nibio.go.jp/homstrad/
[14] N. Saitou and M. Nei, *Mol. Biol. Evol.*, 4: 4 (1987) [PMID: 3447015]
[15] A. Drummond and K. Strimmer, *Bioinformatics*, 17: 7 (2001) [PMID: 11448888]
[16] C. B. Do, *et al, Genome Res.*, 15: 2 (2005) [PMID: 15687296]
[17] M. S. Waterman and T. F. Smith, *J. Theor. Biol.*, 73: 4 (1978) [PMID: 703348]

Edited by P. Kangueane
Citation: Essoussi *et al.*, Bioinformation 2(10): 452-455 (2008)

Supplementary material

Equations

$$T\_dRF\left(TT_{ij}, RT_{ij}\right) = 0, \text{ if the two trees are identical} \quad \rightarrow \quad (1)$$
$$\neq 0, \text{ otherwise,}$$

where TT_{ij} and RT_{ij} are, respectively, the test tree j and the reference tree j inside the category i (i = 1 to 3) of each of DS-BB and DS-HOM datasets. If the TT is identical to the RT, the T_dRF is equal to 0.

$$dRF(M) = \sum_{j=1}^{k} \left(T\_dRF\left(TT_{ij}, RT_{ij}\right) = 0 \right) \Big/ k, \quad \rightarrow \quad (2)$$

where M is a given alignment method and k is the number of RTs in the category i of each dataset.

Tables

	ClustalX	Align-m	T-Coffee	SAGA	ProbCons	MAFFT	MUSCLE	DIALIGN
ClustalX		(0.18)	(0.74)	(0.11)	(0.8)	(0.26)	(0.91)	(0.18)
Align-m	(0.69)		(0.32)	(0.77)	(0.29)	(0.82)	(0.17)	(0.95)
T-Coffee	(0.99)	(0.69)		(0.21)	(0.93)	(0.44)	(0.68)	(0.34)
SAGA	(0.61)	(0.9)	(0.61)		(0.19)	(0.65)	(0.09)	(0.77)
ProbCons	(0.88)	(0.58)	(0.88)	(0.51)		(0.4)	(0.73)	(0.29)
MAFFT	(0.37)	(0.61)	(0.37)	(0.7)	(0.29)		(0.24)	(0.86)
MUSCLE	(0.64)	(0.39)	(0.63)	(0.33)	(0.74)	(0.17)		(0.17)
DIALIGN	(0.51)	(0.79)	(0.51)	(0.88)	(0.42)	(0.81)	(0.26)	

Table 1: Statistical significance on DS-BB and DS-HOM datasets is shown. Each entry in the table contains the *P*-value assigned by a Wilcoxon rank test to the difference between a pair of methods. The upper-right corner of the matrix is obtained from T_dRF distances on DS-BB, the lower-left corner from T_dRF distances on DS-HOM. All the *P*-value are >0.05. The differences are not considered significant and are shown in parentheses.

	ClustalX	Align-m	T-Coffee	SAGA	ProbCons	MAFFT	MUSCLE	DIALIGN
ClustalX		(0.3)	(0.3)	+0.01	-0.031	(0.65)	(0.37)	(0.07)
Align-m	(0.4)		-0.045	(0.13)	-1.2×10^{-3}	(0.59)	(0.061)	(0.39)
T-Coffee	(0.56)	(0.15)		$+1.3 \times 10^{-3}$	(0.24)	(0.13)	(0.9)	$+5.4 \times 10^{-3}$
SAGA	(0.052)	(0.28)	-0.0128		-7×10^{-6}	-0.029	-1.1×10^{-3}	(0.37)
ProbCons	+0.048	+0.0048	(0.14)	$+1.3 \times 10^{-4}$		$+5 \times 10^{-3}$	(0.16)	$+4.8 \times 10^{-5}$
MAFFT	(0.71)	(0.64)	(0.32)	(0.1)	-0.014		(0.15)	(0.15)
MUSCLE	(0.3)	(0.054)	(0.59)	$+2.9 \times 10^{-3}$	(0.34)	(0.126)		$+6.5 \times 10^{-3}$
DIALIGN	-0.038	-0.007	-0.007	(0.98)	-4.1×10^{-5}	(0.078)	-1.6×10^{-3}	

Table 2: Ranks and statistical significance on DS-BB and DS-HOM datasets is shown. Each entry in the table contains the *P*-value assigned by a Wilcoxon rank test to the difference between a pair of methods. The upper-right corner of the matrix is obtained from SP scores on DS-BB, the lower-left corner from SP scores on DS-HOM. If the method to the left is ranked higher than the method above, the *P*-value is preceded by +. If the method to the left is ranked lower, the *P*-value is preceded by -. If the *P*-value is >0.05, the difference is not considered significant and is shown in parentheses.