International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:2, No:8, 2008

Comparison of Phylogenetic Trees of Multiple Protein Sequence Alignment Methods

Khaddouja Boujenfa, Nadia Essoussi, and Mohamed Limam

***Abstract*—**Multiple sequence alignment is a fundamental part in many bioinformatics applications such as phylogenetic analysis. Many alignment methods have been proposed. Each method gives a different result for the same data set, and consequently generates a different phylogenetic tree. Hence, the chosen alignment method affects the resulting tree. However in the literature, there is no evaluation of multiple alignment methods based on the comparison of their phylogenetic trees. This work evaluates the following eight aligners: ClustalX, T-Coffee, SAGA, MUSCLE, MAFFT, DIALIGN, ProbCons and Align-m, based on their phylogenetic trees (test trees) produced on a given data set. The Neighbor-Joining method is used to estimate trees. Three criteria, namely, the dNNI, the dRF and the Id_Tree are established to test the ability of different alignment methods to produce closer test tree compared to the reference one (true tree). Results show that the method which produces the most accurate alignment gives the nearest test tree to the reference tree. MUSCLE outperforms all aligners with respect to the three criteria and for all datasets, performing particularly better when sequence identities are within 10-20%. It is followed by T-Coffee at lower sequence identity (<10%), Align-m at 20-30% identity, and ClustalX and ProbCons at 30-50% identity. Also, it is noticed that when sequence identities are higher (>30%), trees scores of all methods become similar.

***Keywords*—**Multiple alignment methods, phylogenetic trees, Neighbor-Joining method, Robinson-Foulds distance.

I. Introduction

MULTIPLE sequence alignment is a crucial first step in phylogenetic analysis. Several alignment methods have been proposed, such as ClustalX [1] T-Coffee [2], MUSCLE [3], MAFFT [4], ProbCons [5], Align-m [6], DIALIGN [7] and SAGA [8].

In the literature, many comparisons of multiple alignment methods have been conducted. Reference [9] stressed the ability of the methods to identify correctly short motifs found in four sets of homologous proteins. Reference [10] evaluated the ability of multiple alignments in identifying new family members in databases searches. Reference [11] predicts the reliability of seven multiple alignment servers in order to allow users to select the most suitable technique according to their requirements in terms of selectivity and sensitivity. Reference [12] presents a systematic analysis and comparison of several alignment programs, using the BaliBASE reference alignments as test cases. Despite these comparison studies, for a given data set, choosing an alignment method which produces the nearest phylogenetic test tree to the reference tree is still open to discussion.

Each alignment method gives a different alignment result for the same input data, and consequently produces a different tree. Hence, the resulting tree depends on the underlying alignment method. The evaluation of these methods, based on their generated phylogenetic tree has not yet been done. Therefore, for each alignment method, a comparison of test trees with respect to reference trees is proposed.

This comparison is based on three metrics, namely the Nearest Neighbor Interchange distance (dNNI) [13], the Robinson-Foulds distance (dRF) [14], and the identity tree (Id_Tree). The BaliBase [12] and OXBench [15] benchmark alignment data sets are used to generate the reference trees.

II. Methods

A. Benchmarking

BaliBase 2.0 is a benchmark alignment database, consisting of a collection of 141 reference protein alignments, dedicated to the evaluation of multiple alignment programs.

The database addresses all problems that can be encountered when aligning complete sequences. It provides five reference alignment sets. Reference 1 (with 82 reference alignments) consists of a few equidistant sequences with various levels of conservation. Reference 2 (with 23 reference alignments) contains families of closely related sequences with up to three distant 'orphan' sequences. Reference 3 (with 12 reference alignments) is made of divergent families. Reference 4 and 5 (with 12 reference alignments, respectively) contain sequences with large N/C terminal extensions and internal insertions, respectively.

OXBench is a data set of reference alignments and software tools for benchmarking pair-wise and multiple alignment methods. OXBench includes three reference alignment data sets. The Master data set, (with 672 reference alignments) is used to assess alignments accuracies and not to optimize method's parameters. The Full data set is the full-length sequences of the domains contained in the master data set (with 605 reference alignments). It is used to test if a method

Manuscript received August 15, 2008.

K. Boujenfa is with the Laboratory of Operational Research Decision and Process Control (LARODEC), Bouchoucha, Le Bardo 2000 Tunisia (corresponding author to provide phone: 216-99161858; e-mail: khaddouja.Boujenfa@ isg.rnu.tn).

N. ESSOUSSI is with the Computer Science Department, High Institute of Management, Tunis, Tunisia (e-mail: Nadia.Essoussi@isg.rnu.tn).

Professor Limam is with LARODEC laboratory, University of Tunis (e-mail: Mohamed.Limam@isg.rnu.tn).

can correctly align a domain that is contained within a longer protein sequence. The Extended data set (with 672 reference alignments) is the master set of domains augmented by sequences of unknown structures. It is used to test the effect of having more sequences in an alignment on the alignment quality. Only the Master data set is considered in this study.

BaliBase and OXBench are categorized according to the percent sequence identities within the reference alignment (0-10%, 10-20%, 20-30%, 30-50%), and labeled accordingly. The names and details of each category are described in Table I. The size of each data set is restricted to test cases of four or more sequences.

B. Alignment Programs

Eight multiple alignment programs are used to generate alignments as inputs for phylogenetic analysis. Corresponding URLs are listed in Table II. The programs used are ClustalX 1.81, SAGA 0.95, ProbCons 1.08, T-Coffee 3.93, DIALIGN2.2.1, MAFFT 5.743, MUSCLE 3.6 and Align-m 2.3. ClustalX is a windows interface for the widely-used progressive multiple sequence alignment program CLUSTALW [16]. SAGA uses a genetic algorithm to optimize a multiple sequence alignment given an objective function. ProbCons uses a consistency-based objective function to align sequences. It maximizes the score between the final multiple alignment and a library of pair-wise hidden Markov models. As ProbCons, T-Coffee uses a consistency-based objective function and a library of local and global alignments. DIALIGN uses a local greedy alignment algorithm to construct a global alignment. It finds local similarities by segment comparisons. MAFFT uses a progressive approach to generate an alignment. In order to rapidly identify homologous regions, MAFFT uses the fast Fourier transform. MUSCLE is also based on a progressive alignment algorithm but uses the Log Expectation scoring function to align two profiles. Align-m uses the consistency of pair-wise alignments to construct the final multiple alignment. All these methods are run with default settings on the eight test cases given in Table I. Tests were performed on a 1.6-GHz Intel Pentium M with 512 MB RAM.

C. Estimation of Phylogenetic Trees

The Neighbor Joining method [17] is used to estimate all the trees. Based on distances between sequences, this method uses a greedy algorithm which predicts an evolutionary tree based on progressively adding the next most-alike sequence, or set of sequences, as an additional branch to an existing tree. The 363 multiple alignments generated from each aligner are given as input to the Neighbor Joining method. Thus, a total of 2904 (363*8) test trees (TT) is generated. Each 363 test trees are compared to the 363 reference trees (RT) estimated from the reference alignments of BaliBase and OXBench.

D. Comparison Process

In order to evaluate the alignment methods, three scores, namely, the dNNI(M), the dRF(M) and the Id_tree(M) are determined. These scores are based on metrics, namely, T_dNNI, T_dRF and T_Identity designed to compare two trees. The higher are the dNNI(M), dRF(M) and Id_tree(M) scores, the more performing is the alignment method.

- ***dNNI(M)*:**

TABLE I
BENCHMARK DATA SETS USED IN THIS ANALYSIS AND THEIR SUBSETS

Database (363)	Category Name	Number of tests	Average Percent Identity	Average Sequence Number	Average Sequence Length
BaliBASE (137)					
	BB_10	86	2.50	11	241
	BB_20	29	13.69	4	293
	BB_30	19	23.37	5	269
	BB_50	3	32.39	4	321
OxBench (226)					
	OXB_10	109	4.20	9	138
	OXB_20	43	14.77	8	137
	OXB_30	31	24.7	7	144
	OXB_50	43	39.25	7	115

BaliBase reference sets are prefixed with "BB", and those of OXBench with "OXB". For each reference set, the suffix denotes the percent identity of the sequences within the reference alignment. Numbers in brackets show the total number of data sets.

TABLE II
URLS OF THE MULTIPLE ALIGNMENT PROGRAMS

Method	URLs
ClustalX	ftp://ftp-igbmc.u-strasbg.fr/pub/ClustalX/
SAGA	http://www.tcoffee.org/Projects_home_page/saga_home_page.html
ProbCons	http://probcons.stanford.edu
DIALIGN	http://bibiserv.techfak.uni-bielefeld.de/dialign/
T-Coffee	http://www.tcoffee.org/Project_home_page/t_coffee_home_page.html
MAFFT	http://www.biophys.kyoto-uac.jp/~katoh/programs
MUSCLE	http://www.drive5.com/muscle
Align-m	http://bioinformatics.vub.ac.be/

The T_dNNI distance is given by the following equation:

$$T\_dNNI\left(TT_{ij}, RT_{ij}\right) = 0, \text{ if the NNIs} = 0 \quad (1)$$
$$\neq 0, \text{ otherwise},$$

where TT_{ij} and RT_{ij} are, respectively, the test tree j and the reference tree j inside the category i of each benchmark. This distance gives the minimum number of NNIs required to change one tree to another. The COMPONENT 2.0 software [18] is used to compare two trees. The test tree is better if it requires minimum number of NNIs. Thus, the distance between two given trees, T_dNNI, is optimal if it is equal to zero. This means that no interchanges are necessary to transform a given test tree into a reference one.

The dNNI(M, i) is given by the following equation:

$$dNNI(M,i) = \sum_{j=1}^{k} \left(T\_dNNI\left(TT_{ij}, RT_{ij}\right) = 0\right), \quad (2)$$

TABLE III
TREE SCORES PRODUCED BY THE ALIGNERS ON EACH CATEGORY OF THE BALIBASE BENCHMARK ALIGNMENT DATABASE

Method	BB_10 (86)			BB _20 (29)			BB _30 (19)			BB _50 (3)			Overall (137)		
	dRF (M.1)	dNNI (M.1)	Id_tree (M.1)	dRF (M.2)	dNNI (M.2)	Id_tree (M.2)	dRF (M.3)	dNNI (M.3)	Id_tree (M.3)	dRF (M.4)	dNNI (M.4)	Id_tree (M.4)	dRF (M)	dNNI (M)	Id_tree (M)
CLUSTALX	33	33	38.4	23	23	79.3	16	16	84.2	**3**	**3**	**100**	75	75	54.7
ALIGNM	26	26	30.2	22	22	75.9	17	17	89.5	**3**	**3**	**100**	68	68	49.6
T-COFFEE	**34**	**34**	**39.5**	24	24	82.8	15	15	78.9	**3**	**3**	**100**	76	76	55.5
SAGA	26	26	30.2	21	21	72.4	16	16	84.2	2	2	66.67	65	65	47.4
PROBCONS	33	33	38.4	24	24	82.8	16	16	84.2	**3**	**3**	**100**	76	76	55.5
MAFFT	25	25	29.1	24	24	82.8	16	16	84.2	2	2	66.67	67	67	48.9
MUSCLE	33	33	38.4	**26**	**26**	**89.7**	**18**	**18**	**94.7**	**3**	**3**	**100**	**80**	**80**	**58.4**
DIALIGN	25	25	29.1	23	23	79.3	14	14	73.7	**3**	**3**	**100**	65	65	47.4

Values show the dRF(M, i), the dNNI(M, i) and the Id_tree(M, i) produced by each aligner for a given category i (i = 1 to 4). The three last columns show the values of dRF(M), dNNI(M) and Id_tree(M) achieved by each aligner for the entire BaliBase categories. Values of Id_tree(M, i) and Id_tree(M) are in percentage. The number of sequences in each reference dataset is given in parentheses. The best results in each column are shown in bold.

TABLE IV
TREE SCORES PRODUCED BY THE ALIGNERS ON EACH CATEGORY OF THE OXBENCH BENCHMARK ALIGNMENT DATABASE

Method	OXB_10 (109)			OXB_20 (43)			OXB_30 (31)			OXB_50 (43)			Overall (226)		
	dRF (M.1)	dNNI (M.1)	Id_tree (M.1)	dRF (M.2)	dNNI (M.2)	Id_tree (M.2)	dRF (M.3)	dNNI (M.3)	Id_tree (M.3)	dRF (M.4)	dNNI (M.4)	Id_tree (M.4)	dRF (M)	dNNI (M)	Id_tree (M)
CLUSTALX	**73**	**73**	**67**	35	35	81.4	27	27	87.1	**36**	**36**	**83.7**	**171**	**171**	**75.7**
ALIGNM	67	67	61.5	**36**	**36**	**83.7**	**30**	**30**	**96.8**	35	35	81.4	168	168	74.3
T-COFFEE	**73**	**73**	**67**	35	35	81.4	27	27	87.1	34	34	79.1	169	169	74.8
SAGA	65	65	59.6	31	31	72.1	22	22	71	33	33	76.7	151	151	66.8
PROBCONS	69	69	63.3	33	33	76.7	26	26	83.9	**36**	**36**	**83.7**	164	164	72.6
MAFFT	69	69	63.3	35	35	81.4	25	25	80.6	35	35	81.4	164	164	72.6
MUSCLE	72	72	66.1	**36**	**36**	**83.7**	28	28	90.3	35	35	81.4	**171**	**171**	**75.7**
DIALIGN	66	66	60.6	**36**	**36**	**83.7**	27	27	87.1	35	35	81.4	164	164	72.6

Details are as in Table III.

where M is an alignment method and k is the number of reference trees inside the category i of each benchmark. This distance is defined as the total number of T_dNNI equal to zero between each test and reference tree for a given category. Thus, dNNI(M), given by the following equation

$$dNNI(M) = \sum_{i=1}^{4} dNNI(M, i), \tag{3}$$

estimates the total number of dNNI equal to zero produced by each alignment method on the four categories of each benchmark.

- ***dRF(M)*:**

The T_dRF distance is given by the following equation:

$$T\_dRF\left(TT_{ij}, RT_{ij}\right) = 0, \text{ if the two trees are identical} \tag{4}$$
$$\neq 0, \text{ otherwise.}$$

It is a topological measure based on splits. The Vanilla package fronted to PAL 1.2 [19] is used to compare trees based on this measure. It 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. If trees are identical, the T_dRF is equal to zero.

The dRF(M, i), given by the following equation

$$dRF(M, i) = \sum_{j=1}^{k} \left(T\_dRF\left(TT_{ij}, RT_{ij}\right) = 0\right), \tag{5}$$

defines the total number of T_dRF equal to zero between each test and reference tree for a given category. Thus, dRF(M), given by the following equation

$$dRF(M) = \sum_{i=1}^{4} dRF(M, i), \tag{6}$$

TABLE V
TREE SCORES PRODUCED BY THE ALIGNERS ON THE WHOLE DATABASES

Method	0_10 (195)			10 _20 (72)			20_30 (50)			30_50 (46)			Overall (363)		
	dRF (M.1)	dNNI (M.1)	Id_tree (M.1)	dRF (M.2)	dNNI (M.2)	Id_tree (M.2)	dRF (M.3)	dNNI (M.3)	Id_tree (M.3)	dRF (M.4)	dNNI (M.4)	Id_tree (M.4)	dRF (M)	dNNI (M)	Id_tree (M)
CLUSTALX	106	106	54.4	58	58	80.6	43	43	86.0	**39**	**39**	**84.8**	246	246	67.8
ALIGNM	93	93	47.7	58	58	80.6	**47**	**47**	**94.0**	38	38	82.6	236	236	65.0
T-COFFEE	**107**	**107**	**54.9**	59	59	81.9	42	42	84.0	37	37	80.4	245	245	67.5
SAGA	91	91	46.7	52	52	72.2	38	38	76.0	35	35	76.1	216	216	59.5
PROBCONS	102	102	52.3	57	57	79.2	42	42	84.0	**39**	**39**	**84.8**	240	240	66.1
MAFFT	94	94	48.2	59	59	81.9	41	41	82.0	37	37	80.4	231	231	63.6
MUSCLE	105	105	53.8	**62**	**62**	**86.1**	46	46	92.0	38	38	82.6	**251**	**251**	**69.1**
DIALIGN	91	91	46.7	59	59	81.9	41	41	82.0	38	38	82.6	229	229	63.1

Columns show the values of the three comparison tree criteria described in this analysis for the entire databases. The parameter *k* in Equations 2, 5, and 8 takes the number of sequences in each category for the entire databases (the numbers given in parentheses). The parameter *N* in Equation 9 takes the total number of sequences (363). The best results in each column are shown in bold. Values of Id_tree(M, i) and Id_tree(M) are in percentage.

computes the total number of identical trees produced by each alignment method on each of the BaliBase and the OXBench categories.

- ***Id_tree(M)*:**

A score called the T_Identity for each test tree is given by the following equation:

$$T\_Identity\left(TT_{ij}, RT_{ij}\right) = 1, if\ T\_dRF\left(TT_{ij}, RT_{ij}\right) = 0 \qquad (7)$$
$$= 0, otherwise.$$

It evaluates the similarity of the structure and positions of leaves between test and reference trees. This metric is equal to 1 if the test tree is identical to the reference tree and 0, otherwise. Two trees are identical if the Robinson-Foulds distance (T_dRF) is equal to 0 (Equation 4). This means that the test tree has the same structure and positions of leaves as in the reference tree.

A score Id_tree(M, i) is given by the following equation:

$$Id\_tree(M,i) = \sum_{j=1}^{k} T\_Identity\left(TT_{ij}, RT_{ij}\right) \Big/ k, \qquad (8)$$

where *k* is the number of reference trees in the category *i* for each benchmark. It is defined as the total sum of T_Identity generated by each alignment method divided by the total number of reference trees for a given category. This distance estimates the percent of identical test trees on each category. Thus, Id_Tree(M) is given by the following equation:

$$Id\_tree(M) = \sum_{i=1}^{4}\sum_{j=1}^{k} T\_Identity\left(TT_{ij}, RT_{ij}\right) \Big/ N, \qquad (9)$$

where *N* is the total number of reference trees for each benchmark. It is defined as the total sum of T_Identity divided by the total number of reference trees for each benchmark. This score estimates the proportion of identical trees given by each alignment method on the BaliBase and OXBench databases.

III. Results

Test trees of the eight aligners are compared to the reference trees of BaliBase and OXBench data sets. Values of the dRF, dNNI and Id_tree for the eight alignment methods, on each reference subset (BB_10 to BB_50 for BaliBase and OXB_10 to OXB_50 for OXBench) are given in Tables III and IV. The values of dRF(M), dNNI(M) and Id_tree(M) for each method are shown in the last columns. The best results are highlighted in bold. T-Coffee and ClustalX are the best on OXBench data sets when sequences are distantly related (0-10%), followed by MUSCLE with a slight difference of 0.9% lower Id_tree. T-Coffee is the most accurate alignment method on BaliBase data sets for sequence identity lower than 10%, followed closely by ClustalX, ProbCons and MUSCLE.

At 10-20% sequence identity, on the OXBench data sets, Align-m, MUSCLE and DIALIGN outperform the other aligners with 2.3 to 11.6% better Id_tree. However, MUSCLE generates better test trees on the BaliBase data sets.

When sequence identities are within 20 and 30%, on the OXBench data sets, Align-m achieves the highest results, followed by MUSCLE with 6.5% better Id_tree. On BaliBase (BB_30), MUSCLE outperforms Align-m with 5.2% higher Id_tree.

At 30-50% sequence identities (OXB_50 and BB_50), most alignment methods achieve similar results. However, on the OXBench data set, ClustalX and ProbCons slightly outperform Align-m, MAFFT, MUSCLE and DIALIGN with 2.3% better Id_tree, followed by T-Coffee with 4.6% and SAGA with 7%. On BaliBase, approximately all methods agree and give similar results.

The last columns in Table IV show that ClustalX and MUSCLE are the best on all categories of OXBench database, with 0.9-8.9% higher Id_tree. However, as shown in Table III, MUSCLE outperforms different alignment methods on the overall BaliBase data sets with 2.9-11% Id_tree range.

Table V shows the values of the dRF, the dNNI and the Id_tree given by each alignment method on the four categories for all databases. MUSCLE is the optimal aligner with a

TABLE VI
AVERAGE CS AND SP SCORES GIVEN BY THE ALIGNERS ON THE FOUR CATEGORIES OF BALIBASE, OXBENCH AND FOR THE WHOLE DATABASES

	BaliBase									
	BB_10 (86)		BB_20 (29)		BB_30 (19)		BB_50 (3)		Overall (137)	
	CS	SP	CS	SP	CS	SP	CS	SP	CS	SP
CLUSTALX	42.1	71.3	77.7	87.1	86.2	92.4	88.3	92.6	56.8	78.0
ALIGNM	37.4	67.6	73.4	81.3	79.9	86.8	81.0	87.8	51.9	73.6
T-COFFEE	43.8	72.7	78.2	87.2	83.8	89.9	82.3	89.9	57.5	78.5
SAGA	31.6	62.4	67.2	78.6	77.4	87.6	72.7	84.8	46.4	69.8
PROBCONS	**51.4**	**77.8**	**81.5**	**89.4**	**87.8**	**93.2**	**89.0**	**93.7**	**63.6**	**82.8**
MAFFT	42.2	71.7	72.8	84.4	85.1	91.1	81.7	90.0	55.5	77.4
MUSCLE	45.7	73.8	78.1	86.9	85.0	91.5	83.7	91.1	58.8	79.4
DIALIGN	35.0	67.0	69.3	81.2	78.6	86.4	74.7	83.4	49.2	73.1
	OXBench									
	OXB_10 (109)		OXB_20 (43)		OXB_30 (31)		OXB_50 (43)		Overall (226)	
	CS	SP	CS	SP	CS	SP	CS	SP	CS	SP
CLUSTALX	46.5	69.7	79.8	89.7	89.5	95.2	**93.6**	**97.2**	67.7	82.2
ALIGNM	39.5	61.5	78.8	89.6	88.9	94.7	93.4	**97.2**	64.0	78.2
T-COFFEE	46.0	68.4	**80.5**	90.1	88.7	94.6	93.2	97.0	67.4	81.6
SAGA	41.5	65.6	76.9	88.0	88.0	94.3	**93.6**	**97.2**	64.5	79.8
PROBCONS	48.7	70.3	80.4	**90.2**	**90.8**	**95.7**	93.3	96.9	**69.0**	82.6
MAFFT	44.4	67.7	78.0	88.9	88.2	94.2	92.7	96.7	66.0	80.9
MUSCLE	**49.4**	**71.1**	79.8	89.9	88.8	94.5	93.5	97.1	**69.0**	**82.8**
DIALIGN	37.2	59.4	75.0	86.7	83.7	91.6	92.1	96.4	61.2	76.1
	All databases									
	0-10 (195)		10-20 (72)		20-30 (50)		30-50 (46)		Overall (363)	
	CS	SP	CS	SP	CS	SP	CS	SP	CS	SP
CLUSTALX	44.5	70.4	78.9	88.6	88.3	94.1	**93.3**	**96.9**	63.6	80.6
ALIGNM	38.6	64.2	76.7	86.2	85.5	91.7	92.6	96.6	59.4	76.5
T-COFFEE	45.0	70.3	79.6	89.0	86.9	92.8	92.5	96.5	63.7	80.4
SAGA	37.1	64.2	73.0	84.2	84.0	91.7	92.2	96.4	57.7	76.0
PROBCONS	**49.9**	**73.6**	**80.8**	**89.9**	**89.6**	**94.7**	93.0	96.7	**67.0**	**82.7**
MAFFT	43.4	69.4	75.9	87.1	87.0	93.0	92.0	96.3	62.0	79.6
MUSCLE	47.7	72.3	79.1	88.7	87.4	93.3	92.8	96.7	65.1	81.5
DIALIGN	36.2	62.8	72.7	84.5	81.8	89.6	91.0	95.5	56.7	74.9

Columns show the average column scores (CS) and sum-of-pairs (SP) achieved by each aligner for each of the four BaliBase, OXBench and the entire databases reference.subsets, respectively. All scores have been multiplied by 100. The number of sequences in each category is given in parentheses. The best results in each column are shown in bold.

difference range of 1.3-9.6% on Id_tree. It is followed by ClustalX, T-Coffee and ProbCons with a difference of 1.3%, 1.6% and 3% lower Id_tree, respectively.

The phylogenetic trees estimated depend heavily on the quality of alignments produced. Therefore, the accuracies of the multiple alignments of each aligner are estimated using two metrics, CS (Column Score) and SP (Sum of Pairs score) implemented in BaliBase scoring scheme. CS is the number of columns of residues that are identical in both test and reference alignment, divided by the length of the reference. The SP score is defined as the number of correctly aligned residue pairs found in the test alignment divided by the total number of aligned residue pairs in the reference alignment.

Table VI shows the average CS and SP scores achieved by the different aligners, on each category of BaliBase, OXBench and on the whole databases. ProbCons and MUSCLE achieves best CS score on the whole OXBench database. MUSCLE gives 0.2% higher SP score than ProbCons. In addition, ProbCons is more performing on the overall categories of

BaliBase for CS and SP scores. However, ProbCons is followed by MUSCLE with 4.8% lower CS scores and 3.4% SP scores. On the overall databases, it is observed that ProbCons is the best on CS and SP scores, followed by MUSCLE with 1.2% lower SP score and 1.9% lower CS score than ProbCons.

It is noticed that the method which produces accurate alignment generates consequently accurate test trees. MUSCLE achieves accurate alignment as ProbCons, which is known as the most performing method.

IV. Discussion

In this paper, eight multiple alignment methods, namely, ClustalX, T-Coffee, SAGA, MUSCLE, MAFFT, DIALIGN, ProbCons and Align-m are evaluated by comparing their generated phylogenetic test trees to the reference trees. The multiple alignments are produced on the BaliBase and OXBench data sets categorized according to percent sequence identity within the reference alignment (0-10%, 10-20%, 20-30%, 30-50%). The test and reference alignments are given as input to the Neighbor-Joining method to estimate the different phylogenetic trees.

The goal of this comparison is to determine, for each given data set, which alignment method produces the closer test trees to the reference ones. Effectively, there is no single alignment method currently available that consistently outperforms the rest. On a given data set, a ranking of each method is clearly observed. When highly divergent sequences are used, low alignments accuracies are returned and consequently lower tree quality. As sequences become closely related, the difference between methods become marginal and the alignments and tree scores converge. This is clearly observed in Tables V and VI.

At lower sequence identity (<10%), T-Coffee produces the highest percentage of identical trees, followed by ClustalX and MUSCLE with 0.5 and 1.1% lower, respectively. MUSCLE shows high performance within 10-20% sequence identity. However, MUSCLE is ranked second when sequence identities are within 20-30%. It gives 2% lower Id_Tree than Align-m, which produces better results on this given category. The trees scores of the different aligners converge when sequence identity is between 30-50%. However, ClustalX and ProbCons are the best on this data set, followed by Align-m, MUSCLE and DIALIGN with 2.2% lower Id_Tree.

MUSCLE outperforms the different alignment methods in producing more identical test trees to the reference ones on all datasets used in this analysis. MUSCLE is also shown to produce as accurate alignments as ProbCons. These results show that the more accurate is the alignment method the closer are its test trees to reference trees.

References

[1] J.D. Thompson, et al. "The ClustalX:windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tools," Nucleic Acids Res., vol. 25, 1997, pp. 4876–4882.

[2] C. NotreDame, et al. "T-Coffee: A novel method for multiple sequence alignments", J. Mol. Biol., vol. 302, 2000, pp. 205–217.

[3] R.C. Edgar, "MUSCLE: multiple sequence alignment with high accuracy and high throughput," Nucleic Acids Res., vol. 32, 2004, pp. 1792–1797.

[4] K. Katoh, et al. "MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier Transform," Nucleic Acids Res., vol.30, 2002, pp. 3059-3066.

[5] C. B. Do, "ProbCons: Probabilistic consistency-based multiple sequence alignmentg," Genome Res., vol. 15, 2005, pp. 330-340.

[6] I. V. Walle, et al. "Align-m—A new algorithm for multiple alignment of high divergent sequences," Bioinformatics., vol. 20, 2004, pp. 1428-1435.

[7] B. Morgenstern, "DIALIGN2: improvement of the segment-to-segment approach to multiple sequence alignment," Bioinformatics., vol. 15, 1999, pp. 211-218.

[8] C. NotreDame and D. G. Higgins, "SAGA: sequence alignment by genetic algorithm," Nucleic Acids Res., vol. 24, 1996, pp. 1515–1524.

[9] M. A. McClure, et al., "Comparative analysis of multiple protein-sequence alignment methods," Mol. Biol. Evol., vol. 11, 1994, pp. 571–592.

[10] S. Henikoff and J. G. Henikoff, "Embedding strategies for effective use of information from multiple sequence alignments," Protein Sci., vol. 6, 1997, pp. 698-705.

[11] P. Briffeuil, et al., "Comparative analysis of multiple protein sequence alignment servers: clues to enhance reliability of prdictions," Bioinformatics, vol. 14, 1998, pp. 357-366.

[12] J.D. Thompson, et al. "BAliBASE: A benchmark alignment database for the evaluation of multiple alignment programs," Bioinformatics., vol. 15, 1999, pp. 87–88.

[13] M. S. Waterman and T. F. Smith "On the similarity of dendrograms," J. Theor. Biol., vol. 73, 1978, pp. 789-800.

[14] D. F. Robinson and L. R. Foulds "Comparison of phylogenetic trees," Math. Bios., vol. 53, 1981, pp. 131 -147.

[15] G. P. S. Raghava et al., "OXBench: A benchmark for evaluation of protein multiple sequence alignment accuracy," BMC Bioinformatics, vol. 4, 2003.

[16] J.D. Thompson, et al. "ClustalW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice," Nucleic Acids Res., vol. 22, 1994, pp. 4673–4680.

[17] N. Saitou and M. Nei "The Neighbor-Joining method: a new method for reconstructing phylogenetic trees," Mol. Biol. Evol., vol. 4, 1987, pp. 406-425.

[18] R. D. M. Page "COMPONENT: Tree comparison software for Microsoft Windows, version 2.0," The Natural History Museum, London, 1993.

[19] A. Drummond and K. Strimmer "PAL: An object-oriented programming library for molecular evolution and phylogenetics," Bioinformatics, vol. 17, 2001, pp. 662-663.