# Comparison and Evaluation of Multiple Sequence Alignment Tools In Bininformatics

**Asieh Sedaghatinia, Dr Rodziah Binti Atan, KhairinaTajul Arifin, Masrah Azrifah Binti Azmi Murad**

*Dept of Computer Science and Information Technology, University Putra Malaysia, 43400 UPM-Serdang, Malay*

**Summary**

Comparison and alignment of a series of protein and DNA sequences were among the first and are now established as the most powerful and frequently used bioinformatics methods. A variety of computational algorithms and programs have been created for this purpose. Decision about which tools to use is one of the important problems for bioinformaticians, especially for the majority of biologists who are non-specialist users. Therefore, a comparisons study for the different multiple sequence alignment tools (MSA) is necessary for the biologists and bioinformaticians to use the proper software that interprets correctly their biological data. This study addresses this critical issue in relation to MSA algorithms by systematically comparing and evaluating the functionality, usability and the algorithms of three famous multiple sequence alignment tools. A novel method was proposed for qualifying the MSA tools result by using Scorecons server to compute the conservation scores which was named SCS method (ScoreCons Server method). Furthermore, to assert the accuracy of this method for evaluating the quality of MSA tools, the results were compared with the results of SPS and CS. Finally, based on the achievement some considerations in choosing the proper MSA tools were proposed.

***Key words:***

BALiBASE, Conservation score, MSA, Scorecons server,

## 1. Introduction

MSA (Multiple sequence alignment) is an efficient method to compare and align proteins as well as DNA sequences so that similarities and differences can accurately be detected. This is done through searching for a series of individual characters and patterns which follow the same order in sequential analysis. It is widely employed to identify conserved sequence regions which can be regarded as evolutionary related.

In addition, MSA helps to test, modify and predict the function of specific proteins as well as to identify new members of protein families.

Over the past decades more than fifty MSA packages were developed to present biologically meaningful alignment of multiple sequences. This reflects the importance of MSA tools in day-to-day sequences analysis and the variety of purposes for which they are needed.

Manuscript received July 5, 2009
Manuscript revised July 20, 2009

McClure *et al*. (1994) tested the ability of MSA methods to identify short motifs found in four datasets of homologous. Henikoff and Henikoff (1997) evaluated the ability of multiple alignments in identifying new family members in database search. Thompson et *al. (*1999a) presented a systematic analysis and comparison of several alignment programs using the BALiBASE reference alignments as test cases.

Diamantis and Anna (2005) compared the interfaces, the functionalities and parameterization for the 15 MSA tools and secondly the algorithms and the quality of the results were evaluated by using BALiBASE database.

Nevertheless, the ideal choice in a given setting still eludes non-specialist biologists (Purkinkis, 2006). Misuse compounds the dilemma as it can lead to poor quality or erroneous results. Therefore, it seemed essential to conduct a comparison study in order to provide not only novice users but also experienced bioinformaticians with guidance regarding the top choice of MSA tools.

A more detailed knowledge of all currently available methods helps scientists to opt for the ideal software corresponding to their specific needs.

This study addresses this critical issue in relation to MSA algorithms by systematically comparing and evaluating three famous multiple sequence alignment tools:

1-Clustal (Tompson *et al.,* 1994),
2-MUSCLE (Edgar, 2004),
3-T_Coffee (Notredame *et al.,* 2000)

The rationale behind the choice was that the above-mentioned have been widely used as well-established means of alignment in bioinformatics.

In this study, SCS method computes the conservation score for each column of the alignment with using Scorecons server, in order to assess the quality of the alignment by comparing the obtained values with the human created BALiBASE alignment (Thompson *et al*., 1999b).

Scorecons server shows the relation of physico-chemistry properties among different amino acids residue that are exist in each sequences.

To assert the accuracy of this method, the result are then to be compared with the data obtained from two other

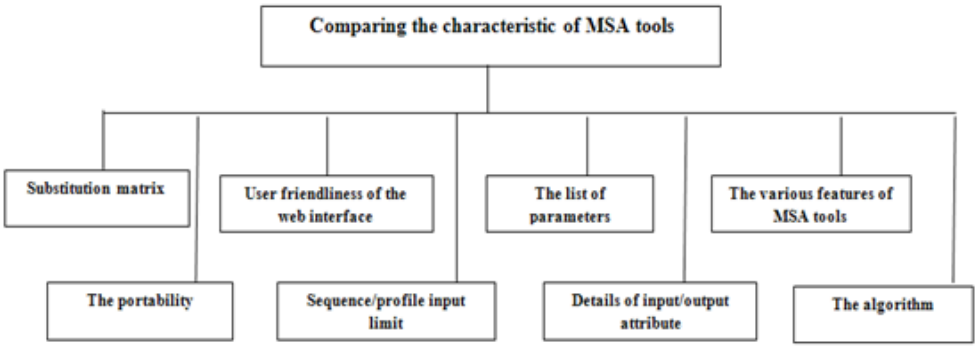



Fig.1 The Characteristics that are evaluated in MSA Tools

common methods; Sum-of-Pair Score (SCS) and Column Score (CS).

This new perspective offers both advantages and disadvantages in regard to the choice of particular MSA tools by users according to specific biological problems.

## 2. Methodology

The latest version of MSA tools which are available as web interfaces were compared and evaluated. Two main aspects were given special importance: functionality and features as well as accuracy and precision.

### 2.1 Functionality and Features

Main features and specifications were selected in view of functionality, as were listed in Fig 1. These features affect the usability and therefore the popularity of the program. Comparison and contrast yielded detailed criteria as can be seen in the summary in Table I.

### 2.2 Algorithms and Accuracy

The latter process is comparing the "heart" of the programs, i.e., the algorithms that define the quality and biological meaning of their results.

BALiBASE version 3 (Thompson *et al.*, 2005) was used as the globally accepted benchmark. Multiple sequence alignment tools were run through the web interface separately with the protein groups of BALiBASE reference datasets.

Defaults parameters were used according to the defined setting. The quality of alignments was initially acquired through a score system implemented in BALiBASE.

Sum-of-Pair score and Column score were obtained for every alignment from Clustal, Muscle and T_Coffee respectively. We used Scorecons server to achieve residue conservation score for every column in all sequence alignments.

Results were then plotted graphically to make visual comparison possible. Needles to say that the most accurate measure was the closest to BALiBASE.

Fig.2 illustrates Scorecons results for a certain series of alignments, compared against BALiBASE scores.

The minimum distance with BALiBASE conservation score was computed, followed by the credit given to the specified tool.

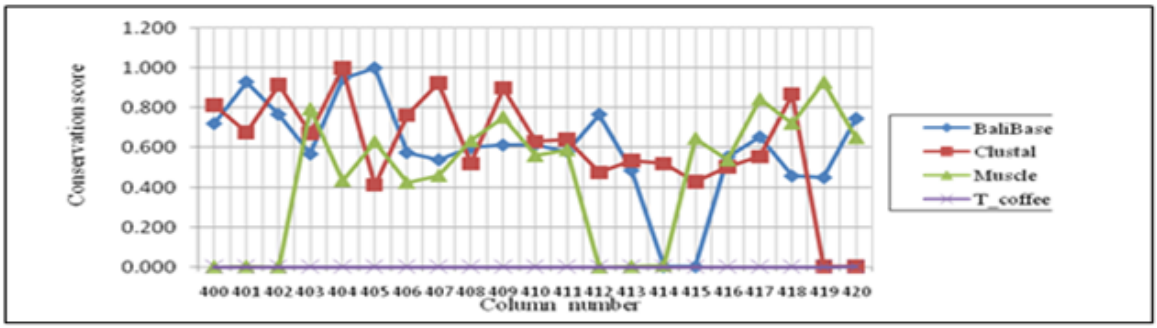
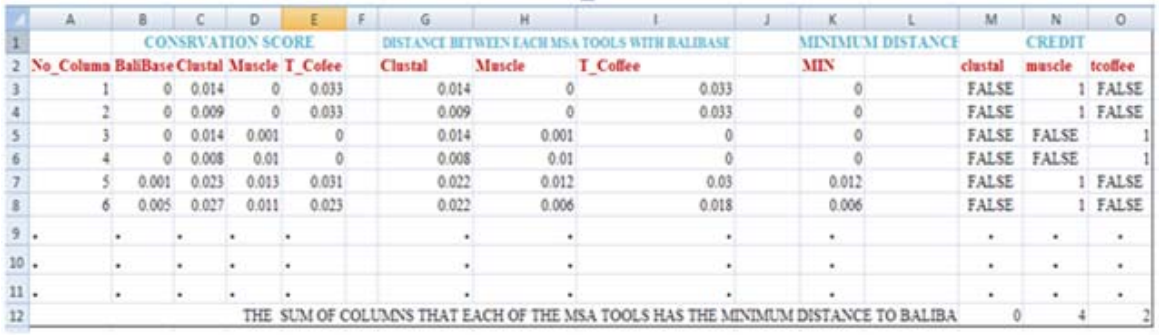



Fig. 2 Part of the Scorecons results for RV20:BB20019



| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | CONSRVATION SCORE | | | | | DISTANCE BETWEEN EACH MSA TOOLS WITH BALIBASE | | | | MINIMUM DISTANCE | | CREDIT | | |
| 2 | No_Column | BaliBase | Clustal | Muscle | T_Cofee | | Clustal | Muscle | T_Coffee | | MIN | | clustal | muscle | tcoffee |
| 3 | 1 | 0 | 0.014 | 0 | 0.033 | | 0.014 | 0 | 0.033 | | 0 | | FALSE | 1 | FALSE |
| 4 | 2 | 0 | 0.009 | 0 | 0.033 | | 0.009 | 0 | 0.033 | | 0 | | FALSE | 1 | FALSE |
| 5 | 3 | 0 | 0.014 | 0.001 | 0 | | 0.014 | 0.001 | 0 | | 0 | | FALSE | FALSE | 1 |
| 6 | 4 | 0 | 0.008 | 0.01 | 0 | | 0.008 | 0.01 | 0 | | 0 | | FALSE | FALSE | 1 |
| 7 | 5 | 0.001 | 0.023 | 0.013 | 0.031 | | 0.022 | 0.012 | 0.03 | | 0.012 | | FALSE | 1 | FALSE |
| 8 | 6 | 0.005 | 0.027 | 0.011 | 0.023 | | 0.022 | 0.006 | 0.018 | | 0.006 | | FALSE | 1 | FALSE |
| 9 | . | . | . | . | . | | . | . | . | | . | | . | . | . |
| 10 | . | . | . | . | . | | . | . | . | | . | | . | . | . |
| 11 | . | . | . | . | . | | . | . | . | | . | | . | . | . |
| 12 | THE SUM OF COLUMNS THAT EACH OF THE MSA TOOLS HAS THE MINIMUM DISTANCE TO BALIBA | | | | | | | | | | | | 0 | 4 | 2 |

Fig. 3 The view of the proposed method for computing SCS

Fig. 3 shows few excel files which includes conservation score results and the calculation of minimum distance for each. Fig. 4 shows the algorithm used to find the overall distance between each of these tools with BALiBASE benchmark.

```
For I:=1 to number of column  BALIBASE
{
Distance1:=Conservation score (Clustal)
Distancne2:=Conservation score (MUSCLE)
Distrance3:=Conservation score (T-Coffee)
Minimum score: =min (distance1, distance2, distance3)
If minimum score: =distance1
{  Clustal.count:=Clustal count+1
   If minimum score: =distance2
   MUSCLE.count:=MUSCLE.count+1
Else
 T-Coffee.count :=T-Coffee.count+1
}}
}
Minimum of distance: =min (T-Coffee.count,
MUSCLE.count, Clustal.count)
```

Fig. 4 The proposed algorithm for computing the SCS.

## 3. Analysis

Although the three mentioned programs have similar functionality, this study only concerns multiple sequence alignment and thus functionality and the characteristics of MSA tools were observed in this particular setting:

Given the input-output format available in tools, Muscle turned out to be the one more limitations as the input can only be in Fasta, which might not be the desired format in certain performances. Other tools favor wider possibilities of format as input sequence.

Another determining factor is the maximum number and length of sequences used to create the alignment. While Muscle can process infinite numbers as well as a remarkable length of 50000 characters, T_Coffee is limited to a mere number of 2000 sequences and thus unsuitable for such calculation.

Portability among different operating systems is of paramount significance as users may intend to run the program on their PC rather than web interfaces. T_Coffee seems deficient since it can not be run in windows and requires Cygwin to provide a Linux-like environment.

## 4. Experimental Results

Fig. 5, Fig. 6 and Fig .7 summarize the results of Friedman test pertaining to the data obtained from Scorecons Score, Sum-of-Pair Score, and Column Score for each of the reference datasets in BALiBASE 3.0 respectively. Noticeably, there is a statistically significant difference in the comparison.

This lends to the need for improvement as there is a considerable gap between MSA tools findings and the already established BALiBASE benchmark values.

What is also noteworthy is that SCS results were similar to the ones achieved by SPS and CS for each category of reference datasets of BALiBASE.

For references RV11, RV12, RV20, RV40 and RV50, T_Coffee achieved the highest SPS and SCS while RV30 was best aligned by Muscle in CS and SCS.

Table1: Summary of the Comparison of the functionality and usability of the MSA tools.

| Characters | Clustal | MUSCLE | T_Coffee |
|---|---|---|---|
| **Input format** | NBRF/PIR<br>EMBL/UNIPRO<br>TKB/<br>SWISS_PROT<br>Pearson(Fasta)<br>GDE<br>ALN/CLUSTAL<br>W<br>GCG/MSF<br>RSF | Fasta | NBRF/PIR<br>EMBL/<br>UniprotKB/<br>Swiss-Prot<br>Pearson(FA<br>STA)<br>GDE<br>ALN/CLU<br>STALW2<br>GCG/MSF/<br>RSF |
| **Output format** | ALN<br>GCG<br>PHYLIP<br>PIR<br>GDE | Fasta<br>Clustalw2<br>MSF<br>Html | Clustalw2<br>MSF<br>HTML<br>PHYLIP |
| **Portability** | UNIX<br>LINUX<br>MAC<br>MS-WINDOWS | Linux<br>Unix<br>Windows<br>XP<br>Mac OS X. | UNIX<br>Windows/<br>Cygwin<br>LINUX<br>Mac OS X. |
| **Substitution matrix** | Blosum<br>Pam<br>Gonnet<br>Id | Blosum<br>Pam<br>Gonnet<br>Id | Blosum<br>Pam<br>Gonnet |
| **Parameters** | Pair wise alignment method, Word method, MSA method and Guide tree parameters. | Output tree parameters. | Matrix parameters. |
| **WEB** | yes | yes | yes |
| **Stand-alone** | yes | yes | yes |
| **Result I/E*** | I/E | I/E | I/E |
| **Algorithms** | Progressive method | Iterative method | Progressive method with extended library |
| **Max sequences** | Maximum of 500 sequences | No limitation | Maximum of 50 sequences |
| **Max length of sequences** | NO limitation | Maximum of 50,000 characters. | Maximum of 2000 characters. |

*The Result 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).

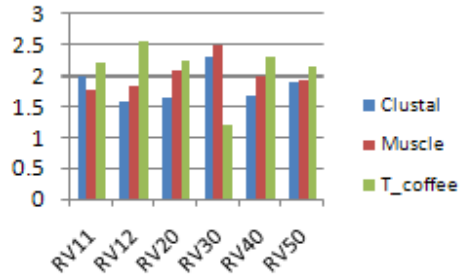



Fig.5 Bar chart of the Mean rank computed by the Friedman test on the SCS for each reference test.

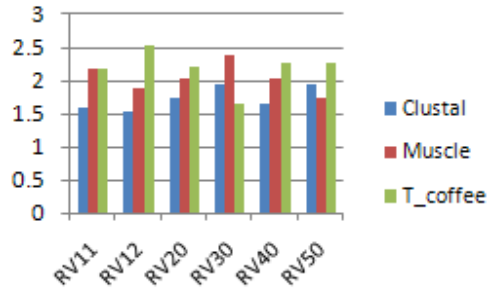



Fig.6 Bar chart of the Mean rank computed by the Friedman test on the SPS for each reference test.

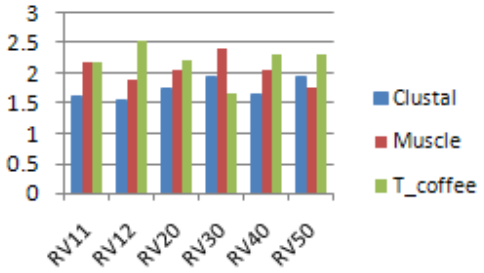



Fig.7 Bar chart of the Mean Rank Computed by the Friedman test on the CS for each reference test.

## 5. Discussion

Based on all the results that achieved from qualifying the quality of the alignment and with considering the information based on the quality of the characteristics of MSA tools, Table 2 shows some consideration about choosing the proper tools between three online multiple sequence alignment tools.

Table 2: Some consideration in choosing the proper MSA tools

| **Program** | **Some major Advantages** |
|---|---|
| Clustal | >This tool is known as one of the old and famous and creditable MSA tools.<br>>There are a series of remarkable parameters that are accessible for user to select.<br>>There is no limitation on the length of the sequences. |
| MUSCLE | >There is no limitation on the number of input sequences to be aligned.<br>>Faster and more accurate than Clustal.<br>>So useful for huge amount of data. |
| T-Coffee | >It is useful when high accuracy and high quality of the alignment is needed.<br>>There are so much useful features that T-Coffee is able to do, compare with the other MSA tools. |
| **Program** | **Some major Disadvantages** |
| Clustal | >Less accurate or scalable than modern programs. |
| MUSCLE | >The acceptable format as input sequences is limited to only FASTA format. |
| T-Coffee | >The number of sequences that can be aligned is limited to 50 sequences.<br>>This program does not install on Windows alone and needs to have a Linux-like environment. |

Past studies stated that T_Coffee achieved the highest score in all reference sets (Notredame *et al.*, 2000) while Muscle had the highest CS score in the entire reference categories (Edgar, 2004).

In contrast, in this study the difference could be justified in regards to the BALiBASE version used in each comparison. Former research ran version 1, 1999 whereas while we used version 3, 2006. The discrepancy can also be explained in view of different version of programs in which different series of parameters were employed.

## 6. Conclusion

The aim of this study was to evaluate well-known MSA tools used by biologists and bioinformaticians in order to select the proper software which corresponds best to their specific needs. Alignment results were compared to the BALiBASE benchmark output while scorecons server was employed to achieve scorecons score (SCS) as a new method to assess MSA tools. SCS results were close to SPS and CS finding as T_Coffee had the highest quality among the five reference datasets. The downside was the limitation as to the number of input sequences in addition to Linux-like environment the tools require to run rather than conventional windows.

Muscle receive the second score for the accuracy of the produced alignment while the only possible input format is Fasta; however as opposed to T_Coffee there is no limit to the number of input sequences, which makes it the ideal choice in case of vast data input.

Clustal turned out to be the least accurate as well as scalable program. Nevertheless, there is no denying that it favors remarkable parameters with no limitation as to the length of sequences.

Evidently, the quality of alignment depends on several parameters since highly divergent sequences make results less accurate whereas sequence conservation improves the discrepancy. Nevertheless, there is still need to improve these tools so that higher quality can be achieved.

**Acknowledgments**

A. SedaghatiniAa thanks Dr Waqas Awan because of all his useful guidance, critical advice, encouragement and suggestions during this study.

## References

[1] McClure, M.A. *et al., Mol. Biol. Evol.*, 11: 4 (1994) .PMID 8078398.

[2] Henikoff, S. & Henikoff, J.G. (1997). *Protein. Sci.*, 6: 3 .PMID 9070452.

[3] Diamantis, S. & Anna, C. (2005).Comparison of multiple sequence alignment programs, National and Kapodistrian university of Athens.

[4] Purkinkis, E. (2006). Opening a window in the blackbox:improving bioinformatics tools by exposing their innards to biologists, Indiana university. Retrieved Dec 2, 2005, from

http://www.snowedin.net/windowinthebox/Paper.pdf.

[5] Notredame, C., Higgins, D.G. Heringa, J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol 302(1), 205-17.

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

[7] Edgar, R.C. (2004).MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 32(5), 1792-97.

[8] Valdar, W.S.J. (2002). Scoring residue conservation. Proteins, 48(2), 227-241.

[9] Thompson, J.D., Koehl, P., Ripp, R. & Poch, O. (2005). BALiBASE 3.0: latest developments of the multiple sequence alignment benchmark. Proteins, 61, 127-136.

[10] Thompson, J.D., Plewniak F. & Poch O. (1999a). A comprehensive comparison of multiple sequence alignment programs. Nucleic Acids Res 27, 2682-90. doi:10.1093/nar/27.13.2682. PMID 10373585.

[11] Thompson, J.D., Plewniak, F. & Poch, O. (1999b). BALiBASE: a benchmark alignment database for the evaluation of multiple alignment programs: Nucleic Acids Res 29, 3110-20