SNN-SB: Combining Partial Alignment Using Modified SNN Algorithm with Segment-Based for Multiple Sequence Alignments

Aziz Nasser Boraik Ali¹, Hassan Pyar Ali Hassan² and Hesham Bahamish³

¹Department of Information Technology, Computing Information Technology Faculty, University of Science and Technology, YEMEN.
²Faculty of Environmental Science and Marine Biology, Hadhramout University, YEMEN.
³Computer Science Department, College of Computers & Information Technology, Hadhramout University, YEMEN.

E-mail: aziz_bin_sama@gmail.com

Abstract. Multiple sequence alignment (MSA) is an essential tool in the area of bioinformatics. Many MSA algorithms have been proposed last decade, however there is still opportunity for improvement in accuracy. Including partial alignment into MSA has been proved to be an effective approach to improve the quality of results of final Multiple sequence alignment. This paper presents a novel algorithm known as SNN-SB, which used to detect the related residue of protein sequences to build partial alignments using modified Shared Near Neighbors algorithm and segment-based alignments. The partial alignment is used as guide for DIALIGN-TX algorithm to build the final MSA. In order to evaluate the effectiveness of the SNN-SB, we compared the final result with 10 outstanding MSA tools. The results of SNN-SB got the maximum mean Q score and mean SP score on IRMBASE 2.0 benchmark. Moreover, it got around 8% improvement in BAliBASE 3.0 benchmark regarding to Q score compared to DIALIGN-TX.

1. Introduction

In bioinformatics, sequence alignment is an essential technique of sequence analysis, such as similarity search, phylogenetic tree estimation, protein structure and function prediction [1, 2]. In addition, it could help to find disease related gene variants [3]

Sequence alignment is a technique that used for rearranging biological sequences such as RNA, DNA or proteins. The main purpose of this technique is to detect the similarity between different sequences to discover similar regions. There are two approaches to align the sequences; pairwise alignment and multiple sequence alignments (MSA). For pairwise alignment approach, only two sequences are aligned to each other, while the MSA is used to align more than two sequences. MSA is known to be a very complex problem, either from the algorithmic view point or concerning to the biological science of the output [4, 5].

An extra information about the relation of biological sequences can be embedded into existing MSA tools to increase the accuracy of aligned sequences [5]. The information that helps the MSA tools can be extracted from secondary structure predictions or expert knowledge [6, 7]. In addition, the
information can be extracted from other resources such as the search of database homology to discover local alignments [8] or incorporating some specific partial alignment from an expert to specify particular input sequences to be aligned, which used as a guide for building final output of MSA. Including partial alignment information into MSA tools has been shown to be an efficient technique. This approach was applied in several methods such as DIALIGN [9], DIALIGN-T [10] and DIALIGN-TX [6], and is also have been combined with other MSA methods like MUSCLE [11] T-Coffee [12], and Mafft [13].

In this study, we have proposed an improvement method for detecting partial alignment by incorporating the mutual information for each pair of amino acids with rows information from ungapped segments. This proposed method is developed by extending the work in [14]. In this research, the mutual information among the pairwise alignment is extracted using Shared Near Neighbours method (SNN) [15, 16]. SNN algorithm is a clustering algorithm, where the similarity of the original Shared Nearest Neighbours was based on the number of shared neighbors of any two objects. Therefore, this feature of calculating the similarity has been adaptive to discover the strong similarities of alignments to construct partial alignment in [14]. However, the similarity measurement in SNN which applied in [14] depends only on the number of shared neighbors of two amino acids, where the value of similarity between any two amino acids is solely based on the amount of shared amino acids. It is apparent that the degree of match or mismatch between any two amino acids is taken into consideration. Some studies showed that information contained in the substitution matrix includes biological relationships that can be found among different amino acids, which are used to obtain improved similarity measurement [11, 17]. Thus, the neighbors of amino acids contribute highly to the similarity using their affinities. Therefore, a new similarity based on the degree of matches or mismatches has been proposed in this research to be incorporated into SNN. The generated partial alignment of this research will be used as a guide to construct the final multiple alignment. DIALIGN-TX method will be used to evaluate the partial alignment because it has an explicit option to accept this kind of information.

2. Background and Related Work

Many of the current multiple sequence alignment methods are built around the concept of progressive method [19]. The most basic idea prevailing these conventional progressive alignments is that they start by aligning the most closely similar sequences, after that a new sequence is incorporated to increasing distance. In many cases, the order for picking a new sequence to be aligned to build the MSA is determined based on a guide tree. ClustalW method [20] is the most common progressive method. Further improvement to guide tree in progressive alignment has been presented in GLProbs-Random and GLProbs-Reference [21]. In GLProbs-Random version, the guide tree is generated randomly. The guide tree in GLProbs-Reference version is generated using maximum-likelihood algorithm. Another enhancement in MSA have been obtained by adapting hidden Markov models which presented in ProbCons [22]. ProbCons is an MSA tool that used Probabilistic and Consistency. It uses progressive technique and iterative refinement with Hidden Markov Model (HMM). Probalign [23] developed a similar approach of ProbCons. However, Probalign obtains posterior probabilities from partition function rather than Pair Hidden Markov Model. In the other hand, ProbPFP [24] combines particle swarm optimization algorithm with partition function. In addition, T-Coffee is a method that includes consistency-based approach with progressive alignment [25]. In the progressive alignment part of T-Coffee, a position-specific scoring is used instead of substitution matrix for aligning the sequences. A hybrid approach has been presented in MSAProbs [26]. This approach used pair hidden Markov model and partition function to calculate posterior probabilities among the aligned sequences.

DIALIGN-TX [6] proposed another approach for solving MSA. DIALIGN-TX build the alignment by composing local segments with high similarity. The most common weakness with this approach, that low homologies sequences could be missed easily, thus they may not seem to be statistically significant in the pairwise alignment. To overcome this weakness, new method called SNN-SB has been developed in this work to combine the information of related amino acids with segment-based alignments to produce partial alignments from sequence similarities.
To evaluate the effectiveness of the proposed method, the results will be compared with other 10 leading MSA tools in terms of Q and TC scores. The proposed methods were comparatively analysed against other methods; Dialign-TX [27], Clustalw [20], Mafft [13], MSAprobs [26], Muscle [11], PicXaa [27], Probalign [28], Probcons [22], POA [29] and GLProbs [21] using the default values of input parameters. In addition, we will compare our results with the method that presented in [14], we call it AP-SNN in this research.

3. Proposed Method

In protein sequences, there are twenty amino acid letters used to represent any sequence. They string over the alphabet letters of amino acids $\Sigma_{\text{Protein}} = \{A, N, R, C, D, H, E, Q, G, I, L, M, K, F, S, P, W, U, T, Y\}$. To represent any group of sequences, $S = (s_1, s_2, \ldots s_i)$, each sequence $s_i$ is a series of finite characters from $\Sigma_{\text{Protein}}$, that is, $s_i \in A$. The multiple sequence alignment of $S$ appears as a matrix containing $i$ rows of symbols of $A \cup \{-\}$, where the number of protein sequences $i > 2$.

The proposed method first entails the construction of all pairwise alignments which used as input. In this study we have used Probcons algorithm [22] to generate the pairwise alignments. The modified SNN method is then adapted to obtain partial alignment columns from pairwise alignments by computing the similarity between any two points.

The original SNN defines the similarity between any pair of points based on their shared neighbours. The two points are considered to be related to the same group if they are nearest neighbours of each other and have more than a specific number of mutual neighbours.

The set of $N$ sequences over the alphabet is denoted by $S = \{S_1, S_2, S_3, \ldots , SN\}$. After deploying the pairwise alignment for all input sequences, each point in each sequence $S_N$ is aligned to at least $N-1$ points, where any point can be aligned to a point or gap $\{-\}$. $S_1$ and $S_2$ are then represented by two input sequences on $S$, specifically $X_i \in S_1$ and $Y_j \in S_2$. $L_N$ denotes the length of the $N$-th sequence, where $1 \leq i \leq L_1$ and $1 \leq j \leq L_2$. If we use the original SNN, then the similarity between any pair of points or characters $X_i$ and $Y_j$ in $S_1$ and $S_2$ is generated based on the following equation:

$$\text{Similarity} (X_i, Y_j) = \text{size}(\text{NN}(X_i)) \cap \text{size}(\text{NN}(Y_j))$$

where $\text{NN}(X_i)$ and $\text{NN}(Y_j)$ are the list of aligned points of $X_i$ and $Y_j$ respectively. The size is equal to the number of points in each list. This is the fundamental idea behind the use of SNN algorithm for similarity measurement between any two points.

The SNN algorithm has following steps:

1) The Nearest Neighbour List ($\text{NNL}$) is first created ($k$). For each point, the $N-1$ points to the current point are kept, where $N$ is the number of sequences.

2) A new shared nearest neighbour similarity is calculated between any two points based on (1).

3) The density is set for each aligned point. From the Nearest Neighbour List ($\text{NNL}$) the method detected the points that have shared nearest neighbour similarity greater than user specific parameters epsilon $Eps$ (Figure 1).

4) The core points are then found. In this stage, any point has a density greater than specific parameter $\text{MinPts}$ (minimum points) will be marked as a core point.

5) The partial alignment columns are created by aligning the core points to each other if their similarity is great than $Eps$ (Figure 2). In this step, only the points that have strong link are aligned to each other, which will be appear as partial columns.
The initialization of SNN algorithm requires the following three parameters:

- \( k \) - the number of neighbours, which is equal to \( N-1 \), where \( N \) is the number of the input sequences.
- \( Eps \) - the minimum similarity between any two points, which maintains the connections between the similar group of points and break connections to dissimilar points.
- \( MinPts \) - the minimum density of any point used to determine if a point is core or not.

Adapted SNN can be used to detect the related points from pairwise alignment to construct the partial alignment columns using the similarity measurement. However, the similarity measurement in the original SNN algorithm depends on the number of shared amino acids, where the value of similarity between any two amino acids is solely based on the amount of shared amino acids. It is apparent that the degree of match or mismatch between any two amino acids is not taken into consideration. Some studies showed that information contained in the substitution matrix comprises biological relationships that exist among amino acids, which are used for obtaining improved similarity measurement [17]. Thus, the neighbours of amino acids contribute highly to the similarity using their affinities. Therefore, a new similarity based on the degree of matches or mismatches can be incorporated into SNN.

The proposed weighted similarity between any two amino acids \( X \) in sequence \( S_i \) and \( Y \) in sequence \( S_j \) is calculated using the following equation:

\[
Sim(X, Y) = \sum(score(X, L_x) + score(Y, L_y))
\]

(2)

Where, \( L_x \) is the character that is aligned to \( X \) form its \( k \) Nearest Neighbour List (NNL), \( L_y \) symbolizes the character that aligned to \( Y \) form its \( k \) Nearest Neighbour List, and the parameter \( n \) denotes the sequence number. The score between any two amino acids is calculated based on BLOSUM62 substitution matrix [30]. BLOSUM62 was selected for this work because it is most applied and widely accepted approach [31]. Thus, only the scores that are greater or equal to zero are accepted to calculate the proposed weighted similarity, where there is 50% chance of replacement occurring between any two amino acids [32].

It is obvious that the new similarity in (2) is not only dependent on the number of shared amino acids, but also on the affinities of shared aligned amino acids. In SNN method, the value of parameters; \( Eps \) and \( MinPts \) can range between zero and the size of Nearest Neighbour List (NNL). However, by adding the new similarity in (2) to SNN, similarity threshold is applied to accept or reject any point. The threshold ranges from 0 to a maximum value obtainable using this equation. The maximum value can be calculated by replacing all the scores in (2) with the highest score in BLOSUM62 matrix.

The pseudo-code of finding the weighted similarity between the objects where in this case are amino acids and their respective densities after inclusion of the modified similarity to SNN is expresses in Figure 2.
5

Figure 2. The weighted similarities and the densities

The pseudo-code of aligning the core points after including the weighted similarity is shown in Figure 3. This proposed method is known as initial_SNN-SB.

```
01: Begin
02: Read the input pairwise sequences
03: For all points Do
04: | density of the point = 0
05: | get the k-nearest neighbours of the points from the NNL
06: | For all k-nearest neighbours of the points do
07: | Calculate the similarity between the points and its neighbour
08: | If similarity between the points and its neighbour ≥ Eps
09: | density of the object = density of the object + 1
10: | Else Calculate the weighted similarity
11: | If Weighted Similarity ≥ Threshold then
12: | density of the object = density of the object + 1
13: | End if
14: End if
15: End for
16: End for
17: End
```

Figure 2. The weighted similarities and the densities

Figure 3. Aligning the core points

3.1. Including Segment-Based Alignment Approach

In multiple sequences alignment, the basic building blocks are ungapped pairwise local alignments of any pair of protein sequences [33]. Local alignments are named as fragments alignments, where a fragment consists of pair of ungapped segments of two protein sequences [27]. The length of fragments can be determined by a specific value. Therefore, a fragment corresponds to a pair of equal length of ungapped substrings. Multiple sequence alignments can be composed of such fragments [33].

According to Rausch et al. [33], the segment based technique is a useful way for identifying homologies shared between pair of aligned of sequences, which can be used to improve the alignment accuracy. The segment can be defined using global pairwise and local alignments [34].

In initial_SNN-SB, the related amino acids were detected based on the mutual information for each pair of them. It is very clear that a row information of any pair of amino acids was not considered. Accordingly, enthused by the segment based alignment approach, this study incorporates rows information into related points to further improve alignment accuracy. In this approach, the proposed method searches for groups of segments of the input protein sequences that are connected by related points. It is required that each pair of ungapped segments contain at least a pair of shared related point or more to be incorporated into the final partial alignment.

Let \( S_1, \ldots, S_n \in \Sigma_{\text{Protein}} \) be \( n \) sequences over an alphabet \( \Sigma_{\text{Protein}} \). A segment of sequence \( S_i \) is a triple \((S_i, P, L)\), where \( P \) is the position of first amino acid of the segment in sequence \( S_i \) and \( L \) is segment length. A fragment \( f \) between two protein sequences \( S_i \) and \( S_j \) is then a gap free alignment of two
segments \((S_r, P_i, L_i)\) and \((S_t, P_j, L_j)\), where the length of the two segments are equal \((L_i = L_j)\). The two segments are accepted to be part of partial alignment only if they share at least one related point and the score between any two aligned pair of amino acid is more than 0, score \((P_{i+k}, P_{j+k}) > 0\) for all \(0 \leq k < L_i\).

Figure 4 shows an example of combining segment approach with related points method. Given two sequences, \(S_1\) and \(S_2\) with two related points shared between them, where the bold residues represent the related points. In the example, the segment length equal to 5. There are two segments with length equal to 5 can be detected in the two sequences. The first segment is \{“RFLEG”,”RFLEG”\}, and the second segment is \{“HELAS”,”HELAS”\}. However, only the first segment contains related points. Thus, it will be assumed to be part of partial alignment.

![Related Points](image)

**Figure 4.** Example of segment alignment

1.1. Generating Final MSA from Partial Alignment Columns

A set of partial alignment was obtained from the proposal method. The constructed partial alignment columns were incorporated into Dialign-TX method to evaluate its performance and to tune the input parameters. However, DIALIGN-TX combines progressive strategy with segment-based greedy approach. The MSA in DIALIGN-TX is constructed by searching for similar segments. The segments are detected from the pairwise alignments, and they are greedily inserting or rejected according to their consistency with each other. The main weakness of DIALIGN-TX is that weakly conserved homologies can easily be missed and the segments may be outweighed by spurious random similarities in other pairwise alignments. Thus, the related points from the proposed methods can be used to align the missed segments or to correct the wrongly aligned segments.

4. Results and Discussion

The proposed method initial_SNN-SB has been performed using BaliBase benchmarks to test variant settings of new similarity threshold, where the other parameters Eps and MinPts will be set at 30% for both. After that, the performance of initial_SNN-SB will be evaluated by comparing its results to the previous methods using BaliBase and IRMBASE benchmarks.

4.1. Choosing an Appropriate Similarity Threshold

Table 1 and Table 2 show the effects of the different values of scale parameter of similarity threshold on initial_SNN-SB in terms of Q and TC scores based on BaliBase benchmark. The range of scoring value of similarity threshold is fraction of its maximum value, where the maximum value can be calculated based on (2), where the range of similarity threshold is starting from 10% to 60%.

In Table 1 and Table 2, the results from initial_SNN-SB show low sensitivity to change the value of similarity threshold in terms of Q and TC scores. The highest mean values of Q were achieved when the similarity threshold was set to 30%, where Q score is equal to 86.62% and the highest mean values of TC score was achieved when the similarity threshold was set to 30%, where the TC score is equal to 55.50%. Thus, the similarity threshold value at 30% was selected to perform the initial_SNN-SB method in the rest of this research.
4.2. *The effect of combining segment-base*

To study the effect of including segments-base alignment approach in SNN-SB to find related residue of protein sequences (points), SNN-SB was tested using BaliBase benchmark to evaluate different segment lengths. In this testing, the values of other parameters (Eps, MinPts, threshold similarity) were set at 30%.

The segment length of SNN-SB was selected in the range of 2 to 10, i.e. a segment can consist of 2 to 10 amino acid residues. The average Q score for each experiment were calculated. Figure 5 depicts the plot of adapting the segment size parameters within the chosen range in BaliBase benchmark in terms of Q score. The best average result in Q score was obtained at the segment length of 5. Thus, the segment length at 5 was chosen to perform the SNN-SB method in this study.

| Dataset | Threshold |
|---------|-----------|
|         | 10%   | 20%   | 30%   | 40%   | 50%   | 60%   |
| RV11    | 67.76 | 67.92 | 67.99 | 67.96 | 68.01 | 67.99 |
| RV12    | **94.00** | 93.96 | 93.97 | 93.97 | 93.98 | 93.98 |
| RV20    | 93.40 | 93.42 | **93.43** | **93.43** | 93.42 | 93.41 |
| RV30    | 84.52 | 84.64 | **84.67** | 83.67 | 83.69 | 83.68 |
| RV40    | 90.22 | 90.26 | 90.33 | **90.36** | 90.31 | 90.30 |
| RV50    | 89.29 | 89.24 | 89.31 | 89.30 | **89.33** | 89.32 |
| Mean    | 86.53 | 86.57 | **86.62** | 86.45 | 86.46 | 86.45 |

Table 2. Mean of scores of different similarity threshold in initial_SNN-SB on BaliBase 3.0

| Dataset | Threshold |
|---------|-----------|
|         | 10%   | 20%   | 30%   | 40%   | 50%   | 60%   |
| RV11    | 45.28 | 45.66 | 45.59 | 45.59 | 45.62 | 45.62 |
| RV12    | **85.33** | 85.18 | 85.20 | 85.19 | 85.19 | 85.19 |
| RV20    | 46.92 | 46.92 | **47.02** | 47.00 | 46.82 | 46.82 |
| RV30    | 46.08 | 46.07 | **46.26** | 45.93 | 45.87 | 45.84 |
| RV40    | 52.70 | 53.06 | 53.14 | **53.17** | 53.05 | 53.04 |
| RV50    | **56.03** | 55.77 | 55.77 | 55.78 | 55.76 | 55.76 |
| Mean    | 55.39 | 55.44 | **55.50** | 55.44 | 55.38 | 55.38 |
4.3. Comparing the Results of Proposed Methods with Commonly MSA Methods

This section compares the proposed methods with 10 commonly used methods in terms of Q and TC score based on global alignment benchmark BaliBase 3.0 and local alignment IRMBASE 2.0. The comparison between the computation durations of the proposed method and other methods are subsequently presented.

4.3.1. Global alignment benchmark: BaliBase 3.0

Table 3 and Table 4 summarize the comparative analysis between the accuracy of the proposed methods (initial_SNN-SB, and SNN-SB), AP-SNN and 10 commonly used methods based on Q and TC scores. The comparisons were made based on global reference alignment (BaliBase 3.0). The comparison was performed on the six subset references and the overall dataset.

The proposed method SNN-SB produced more effective alignment and better accuracy in terms of Q and TC scores compared to Dialign-TX, Clustalw, Muscle and POA with statistically significant superiority of SNN-SB (see Table 3). However, the results in Table 3 and Table 4 show that, the MSAprops method has obtained the highest overall Q and TC score (89.25 and 63.76, respectively) among all other tested methods. The Dialign-TX method exhibited a relatively lower alignment accuracy comparing with other MSA tools in terms of Q and TC score values in global alignment benchmarks (79.51 and 45.53, respectively). From the results, it can be inferred that the SNN-SB method considerably improved the Dialign-TX method by 9.17% and 22.84% in Q and TC scores, respectively, with statistically significant differences. The improvements by SNN-SB method accrued in all six subsets compared with Dialign-TX. In RV11 reference, the SNN-SB method obtained the highest improvement compared to Dialign-TX by 22.36% and 43.63% in terms of Q and TC scores respectively.

Further analysis of the results indicated that the proposed method SNN-SB yielded good performance in RV20 subset, where the SNN-SB method outperformed eight out of ten other commonly MSA methods in terms of Q score, although the RV20 reference contains aligned families with a highly divergent “orphan” protein sequence.
The 1st International Conference on Engineering and Technology (ICoEngTech) 2021
Journal of Physics: Conference Series 1962 (2021) 012048
doi:10.1088/1742-6596/1962/1/012048

Table 3. The Q score of proposed methods and commonly MSA tools and on BaliBase 3.0

| Algorithm       | RV11  | RV12  | RV20  | RV30  | RV40  | RV50  | Overall |
|-----------------|-------|-------|-------|-------|-------|-------|---------|
| SNN-SB          | 67.90 | 94.02 | 93.55 | 85.00 | 90.42 | 89.93 | 86.80   |
| initial_SNN-SB  | 67.99 | 93.97 | 93.43 | 84.67 | 90.33 | 89.31 | 86.62   |
| AP-SNN          | 60.43 | 91.60 | 92.40 | 81.85 | 89.42 | 87.10 | 83.80   |
| Dialign-TX      | 55.49 | 89.43 | 89.09 | 76.90 | 83.80 | 82.33 | 79.51   |
| MSAprobs        | 74.63 | **94.86** | **94.36** | **88.19** | 92.54 | **90.91** | **89.25** |
| PicXaa          | 70.60 | 94.65 | 93.30 | 86.81 | **93.23** | 89.38 | 88.00   |
| Clustalw        | 58.15 | 88.41 | 87.77 | 77.14 | 78.93 | 76.91 | 78.05   |
| POA             | 39.35 | 83.11 | 87.07 | 73.61 | 78.23 | 71.19 | 72.09   |
| Mafft           | 69.18 | 93.62 | 93.50 | 87.97 | 92.14 | 90.14 | 87.76   |
| Probalign       | 71.05 | 94.64 | 93.54 | 86.44 | 92.20 | 89.09 | 87.83   |
| Probcons        | 73.65 | 94.64 | 93.73 | 87.62 | 90.32 | 90.06 | 88.34   |
| Muscle          | 65.75 | 92.32 | 91.51 | 84.24 | 86.48 | 85.29 | 84.26   |
| GLProbs         | **74.89** | 94.83 | 93.39 | 87.72 | 92.50 | 90.53 | 88.98   |

Table 4. The TC score of proposed methods and commonly MSA tools and on BaliBase 3.0

| Algorithm       | RV11  | RV12  | RV20  | RV30  | RV40  | RV50  | Overall |
|-----------------|-------|-------|-------|-------|-------|-------|---------|
| SNN-SB          | 45.56 | 85.45 | 47.38 | 46.55 | 54.04 | 56.62 | 55.93   |
| initial_SNN-SB  | 45.59 | 85.26 | 47.02 | 46.26 | 53.14 | 55.77 | 55.50   |
| AP-SNN          | 37.41 | 80.15 | 43.43 | 45.57 | 53.83 | 53.87 | 52.38   |
| Dialign-TX      | 31.72 | 76.00 | 34.69 | 39.91 | 45.17 | 45.66 | 45.53   |
| MSAprobs        | **53.70** | **87.43** | **54.07** | **63.43** | 62.54 | **61.38** | **63.76** |
| PicXaa          | 47.86 | 86.81 | 43.64 | 58.79 | **63.58** | 53.93 | 59.10   |
| Clustalw        | 32.53 | 75.59 | 33.86 | 38.18 | 39.82 | 36.49 | 42.75   |
| POA             | 16.52 | 64.38 | 27.76 | 32.96 | 34.00 | 27.37 | 33.83   |
| Mafft           | 48.35 | 84.35 | 48.87 | 60.60 | 60.65 | 58.28 | 60.18   |
| Probalign       | 48.15 | 86.77 | 46.69 | 59.72 | 60.73 | 54.35 | 59.40   |
| Probcons        | 51.45 | 86.91 | 51.00 | 60.11 | 54.03 | 58.92 | 60.40   |
| Muscle          | 43.32 | 82.00 | 42.22 | 47.68 | 45.32 | 47.50 | 51.34   |
| GLProbs         | 53.01 | 87.35 | 47.07 | 61.04 | 59.73 | 58.32 | 61.09   |

4.3.2. Local Alignment Benchmark: IRMBASE 2.0

For locally related alignment benchmark, Table 5 and Table 6 provide a brief comparative summary of alignment accuracy of the proposed methods and other 10 frequently used aligners based on Q and TC scores on IRMBASE 2.0 dataset.

The SNN-SB method achieved the highest Q and TC scores on the overall IRMBASE dataset with relatively higher statistically significant differences compared to other methods. In the subsets, the SNN-SB method achieved the highest Q and TC scores in Ref3 and Ref4. Although, the method AP-SNN achieved the highest Q score in Ref1 subset. In the terms of TC score in subset Ref1, the method
Mafft achieved the highest score over SNN-SB. In the Ref2 subset, Dialign-TX outperformed SNN-SB in terms of Q and TC scores, although the superiority is not statistically significant.

It was observed in last subsection that the MSAprobs method obtained the highest alignment score in global dataset BaliBase, where it yielded higher accuracy scores compared to SNN-SB, with percentage differences of 2.82% and 14% in terms of Q and TC scores, respectively (as shown in Table 3 and Table 4). However, it is discernible that the proposed method SNN-SB outperforms the MSAprobs method in local alignment dataset IRMBASE, with percentage differences of 11.40% and 29.98% in terms of Q and TC scores, respectively. The reason behind that, the MSAprobs method designed for global alignment. ClustalW method obtained the lowest Q and TC scores in the local alignment dataset. Similar results of ClustalW have been reported in [35]. ClustalW was run with default settings. In ClustalW, there is no refinement process to recorrect any error that can be made during any step of building of MSA, and then the errors can be propagated to the final alignment. Moreover, the quality of MSA in ClustalW is affected by the way the guide tree is constructed.

The overall results of SNN-SB method showed some improvement in alignment accuracy by incorporating the segment based approach to find more related points. The segments were valuable in determining the homology among the sequences and assisting in improving the alignment quality by finding many undetected similarities.

Table 5. The Q score of proposed methods and commonly MSA tools and on IRMBASE 2.0

| Algorithm     | Ref1 | Ref2 | Ref3 | Ref4 | Mean  |
|---------------|------|------|------|------|-------|
| SNN-SB        | 88.87| 94.65| 95.27| 95.47| 93.57 |
| initial_SNN-SB| 88.72| 93.84| 94.87| 95.01| 93.11 |
| AP-SNN        | **90.03**| 94.33| 94.51| 94.72| 93.40 |
| Dialign-TX    | 89.41| **94.90**| 93.75| 93.73| 92.94 |
| MSAprobs      | 78.20| 83.02| 87.59| 85.10| 83.48 |
| PicXaa        | 89.89| 89.34| 89.21| 86.97| 88.85 |
| Clustalw      | 07.14| 10.61| 19.86| 26.35| 15.99 |
| Muscle        | 30.36| 34.54| 54.00| 57.39| 44.07 |
| Mafft         | 87.63| 91.96| 89.86| 88.50| 89.49 |
| Probalign     | 81.74| 82.79| 81.69| 82.57| 82.20 |
| Probcons      | 79.04| 87.23| 86.84| 88.12| 85.31 |
| POA           | 87.56| 49.57| 41.89| 38.24| 54.32 |
| GLProbs       | 79.88| 87.16| 85.29| 87.92| 85.06 |
5. Conclusion

In this paper, we have developed initial_SNN-SB and SNN-SB, new approaches for building partial alignments of MSA. In initial_SNN-SB, Shared Near Neighbours clustering algorithm (SNN) was adapted and modified to obtain partial alignment columns by detecting of the mutual information for every pair of amino acids from pairwise alignments. A new similarity measurement was included to the proposed method to benefit from the degree of match or mismatch between any two amino acids. SNN-SB is an extended version of initial_SNN-SB, a segment based approach was combined with the partial alignment of initial_SNN-SB. Segment based approach was used in SNN-SB to identify homogeneity shared among the portion sequences in order to improve the alignment accuracy. The comparative analysis of the proposed methods is presented based on two benchmarks; BaliBase and IRMBASE. The proposed method SNN-SB shows good accuracy compared to other proposed methods in BaliBase and IRMBASE. On the benchmark BAliBASE, the Q and TC scores of SNN-SB algorithm are close to the highest one in many cases. In IRMBASE dataset, the SNN-SB outperformed the 10 leading methods. These findings show that the accuracy of the final MSA can be greatly improved by combining the relevant points with segment alignments.

References

[1] Edgar RC, Batzoglou S. 2006 Multiple sequence alignment. Current opinion in structural biology. 16(3):368-73.A
[2] Notredame C. 2002 Recent progress in multiple sequence alignment: a survey. Pharmacogenomics. 3(1):131-44.
[3] Cheng L, Yang H, Zhao H, Pei X, Shi H, Sun J, Zhang Y, Wang Z, Zhou M. 2019 MetSigDis: a manually curated resource for the metabolic signatures of diseases. Briefings in bioinformatics. 20(1):203-9.
[4] Hussein AM, Abdullah R, AbdulRashid N, Ali AN. 2017 Protein multiple sequence alignment by basic flower pollination algorithm. In2017 8th International Conf. on Information Technology (ICIT) (pp. 833–838). IEEE.
[5] Pitschi F, Devauchelle C, Corel E. 2010 Automatic detection of anchor points for multiple sequence alignment. BMC bioinformatics. 11(1):1-1.
[6] Subramanian AR, Kaufmann M, Morgenstern B. 2008 DIALIGN-TX: greedy and progressive approaches for segment-based multiple sequence alignment. Algorithms for Molecular
Biology. 3(1):1-1.

[7] Deng X, Cheng J. 2011 MSACompro: protein multiple sequence alignment using predicted secondary structure, solvent accessibility, and residue-residue contacts. BMC bioinformatics. 12(1):1-6.

[8] Thompson JD, Plewniak F, Thierry JC, Poch O. 2000 DbClustal: rapid and reliable global multiple alignments of protein sequences detected by database searches. Nucleic Acids Research. 28(15):2919-26.

[9] Morgenstern B, Dress A, Werner T. 1996 Multiple DNA and protein sequence alignment based on segment-to-segment comparison. Proceedings of the National Academy of Sciences. 93(22):12098-103.

[10] Subramanian AR, Weyer-Menkhoff J, Kaufmann M, Morgenstern B. 2005 DIALIGN-T: an improved algorithm for segment-based multiple sequence alignment. BMC bioinformatics. 6(1):1-3.

[11] Edgar RC. 2004 MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic acids research. 32(5):1792-7.

[12] Notredame C, Higgins DG, Heringa J. 2000 T-Coffee: A novel method for fast and accurate multiple sequence alignment. Journal of molecular biology. 302(1):205-17.

[13] Katoh K, Misawa K, Kuma KI, Miyata T. 2002 MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic acids research. 30(14):3059-66.

[14] Boraik AN, Abdullah R, Venkat I. 2013 Detect Anchor Points by Using Shared Near Neighbors for Multiple Sequence Alignment. In Int. Multi-Conf. on Artificial Intelligence Technology 2013 Aug 28 (pp. 171-182). Springer, Berlin, Heidelberg.

[15] Jarvis RA, Patrick EA. 1973 Clustering using a similarity measure based on shared near neighbors. IEEE Transactions on computers. 100(11):1025-34.

[16] Ertöz L, Steinbach M, Kumar V. 2003 Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data. In Proc. of the 2003 SIAM int. conf. data mining (pp. 47-58). Society for Industrial and Applied Mathematics.

[17] Kelil A, Wang S, Brzezinski R, Fleury A. 2007 CLUSS: clustering of protein sequences based on a new similarity measure. BMC bioinformatics. 8(1):1-9.

[18] Notredame C. 2002 Recent progress in multiple sequence alignment: a survey. Pharmacogenomics. 3(1):131-44.

[19] Chatzou M, Magis C, Chang JM, Kemenia C, Bussotti G, Erb I, Notredame C. 2016 Multiple sequence alignment modeling: methods and applications. Briefings in bioinformatics. 17(6):1009-23.

[20] Thompson JD, Higgins DG, Gibson TJ. 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):4673-80.

[21] Zhan Q, Ye Y, Lam TW, Yiu SM, Wang Y, Ting HF. 2015 Improving multiple sequence alignment by using better guide trees. In BMC bioinformatics (Vol. 16, No. 5, pp. 1-6). BioMed Central.
[27] Sahraeian SM, Yoon BJ. 2010 PicXAA: greedy probabilistic construction of maximum expected accuracy alignment of multiple sequences. Nucleic acids research. 38(15):4917-28.

[28] Roshan U, Livesay DR. 2006 Probalign: multiple sequence alignment using partition function posterior probabilities. Bioinformatics. 22(22):2715-21.

[29] Lee C, Grasso C, Sharlow MF. 2002 Multiple sequence alignment using partial order graphs. Bioinformatics. 18(3):452-64.

[30] Henikoff S, Henikoff JG. 1992 Amino acid substitution matrices from protein blocks. Proceedings of the National Academy of Sciences. 89(22):10915-9.

[31] Sun J, Wu X, Fang W, Ding Y, Long H, Xu W. 2012 Multiple sequence alignment using the Hidden Markov Model trained by an improved quantum-behaved particle swarm optimization. Information Sciences. 182(1):93-114.

[32] AbdulRashid NA. 2008 Enhancement of Hirschberg Algorithm Using N-Gram and Parallel Methods for Fast Protein Homologous Search. PhD, School of Computer Sciences, Universiti Sains Malaysia.

[33] Rausch T, Emde AK, Weese D, Döring A, Notredame C, Reinert K. 2008 Segment-based multiple sequence alignment. Bioinformatics. 24(16):i187-92.

[34] Sammeth M, Morgenstern B, Stoye J. 2003 Divide-and-conquer multiple alignment with segment-based constraints. Bioinformatics.19(suppl_2):ii189-95.

[35] Papadopoulos JS, Agarwala R. 2007 COBALT: constraint-based alignment tool for multiple protein sequences. Bioinformatics. 23(9):1073-9.