Identification of Short Motifs for Comparing Biological Sequences and Incomplete Genomes

Ramez Mina and Hesham H. Ali
College of Information Science and Technology
University of Nebraska at Omaha
Omaha, NE 68116, USA
hali@unomaha.edu

Abstract — **Sequence comparison remains one of the main computational tools in bioinformatics research. It is an essential starting point for addressing many problems in bioinformatics; including problems associated with recognition and classification of organisms. Although sequence alignment provides a well-studied approach for comparing sequences, it has been well documented and reported that sequence alignment fails to solve several instances of the sequence comparison problem, particularly for those sequences that contains errors or those that represent incomplete genomes. In this work, we propose an approach to identify the relatedness among species based on whether their sequences contain similar short sequences or signals. We cluster species based on biological signals such as restriction enzymes or short sequences that occur in the coding regions, as well as random signals for baseline comparison. We focus on identifying k-mers (motifs) that would produce the best results using this approach. The obtained results showed that specific k-mers with biological significance such as restriction enzymes produce excellent results. They also make it possible to obtain good comparisons while using shorter or incomplete sequences, which is a critical property for comparing genomes obtained from next generation sequencers.**

Keywords–sequence comparison; alignment; biological motifs; alrignment-free; k-mers; restriction enzymes; coding sequences; phylogenetic trees

I. INTRODUCTION

The second generation of sequencing provided the bioinformatics domain with more genomes for comparison and analysis, which in turn motivated more researchers to compare these genomes and identify their similar structures and functionalities. The need for more research in the comparative genomic came from the fact that sequence alignment methods have limitations, such in quality and speed. The focus for this research is to find better results for the comparison process.

Although the default sequence comparison methods in the literature are based on alignment, other methods are alignment-free as discussed by S. Vinga et al. [1]. These alignment-free methods introduced alternative solutions to overcome the limitations of alignment-based methods, which led to the question; “how much have these methods achieved to overcome the addressed limitations?” These limitations could be briefed in two major issues, the speed issue and the quality issue. The speed issue was addressed before in the literature, and several accomplishments were reached based on alignment, such as the work of BLAST by S. F. Altschul et al. [2] for pair-wise comparison. Other work focused on multiple sequence alignment with heuristic speeds like the work done in MUSCLE by R. Edgar [3] and DIALIGN by A. Subramanian [4]. In addition to alignment-based methods, other techniques are alignment-free [1] and had a focus on addressing the speed issue as well the quality issue.

Alignment-free methods are not new subject and they are in the literature for a while as discussed by K. Song et al. [5]. Alignment-free methods are categorized mainly into two categories. The first category is based on compression techniques, which improved the speed problem in comparing the biological sequences; the improvement came from the fact that many of the compression algorithms could be implemented in a linear time complexity. Compression-based techniques also showed very good quality with the results, especially those techniques that are dictionary-based. The two major techniques for compression are Lempel-Ziv complexity and Kolomogrov complexity [1].

The second category of alignment-free methods is based mainly on considering all possible k-mers [1, 6] to identify the relatedness between species, and it is specific for each k value. The core of the k-mer method is accomplished by generating vectors that represent the probability of each k-mer within each sequence. The distance is then measured between these vectors. Several proposed techniques were applied to the second approach, either using different formulas for measuring the distance between the vectors, or integrating several vectors of different k values within the same distance measure’s formula.

Several approaches were introduced to construct the measuring vectors, and several formulas were provided and/or designed to calculate the distance between these vectors. Bonham-Carter et al. [7] surveyed the methods that were conducted in this domain/area in [7], and we are summarizing some of these methods.

Liu et al. [8] explained the development of base-base correlation, which is based on generating frequency vectors for all the possible combination of DNA nucleotides of length two (AA, AC…., GT, GG), and each vector is normalized, then a mathematical distance measure is applied to find how closely are the pair sequences. Another approach was discussed by V. Arnau et al. [9] which is called Feature Frequency Approach, and it is also based on generating vectors of specific k-mer, these vectors could also be normalized, then a mathematical measure is applied; which would result in a numerical value for the distance measure. Also application of block-FFP method was necessary, a method similar to the one described by T. J. Wu et al [10]. In another work Sims et al. [11] applied different distance measures that are based on Jensen-Shannon Divergence and Kullback-Leibler Divergence which were discussed in J. Lin [12]. G. Lu et al [6] discussed the same concept of generating the vectors of the k-mers with more in-depth. In their models; they applied several values for k which led to several groups of vectors, each with a different k value, these vectors are called compositions vectors, then applied basic mathematical distance measure, and then tuned up better distance measure that would produce better results.

Application of other distance measures to the composition vectors were borrowed from Z. G. Yu et al. [13] as in the work of R. H. Chan et al. [14] and G Lu et al. [6]. A different approach to generate the vectors was based on suffix trees, a data structure that searches for words of length k, and generates the vectors based on its reading; as of the work of Soares et al. [15].

In general Bonham-Carter et al. [7] discussed in depth more statistical (frequency) measures of different k-mers values, with different distance measures, and these methods would address the frequency and also the occurrence of the k-mers, but they never addressed the order of these k-mers.

Our proposed algorithm is primarily based on exploring information embedded in the k-mers of given sequences, it also considers the order of these k-mers as well as the ability of assigning weights for specific signals (k-mers).

The paper is organized as follows: section II is the motivation for this work; section III is the experimental design and the needed algorithms and the utilized methods for this work; section IV is the provided experiments; section V is the results and analysis; and finally section VI is the conclusion followed by references.

II. MOTIVATION

Sequence comparison has been addressed in the literature for several decades, especially with the birth of the very first alignment-based method, Needleman and Wunsch method for sequence alignment was introduced in 1970. This method dominated the domain for a long time, though its limitations showed up with other advances in the bioinformatics sub-domains, especially with the new sequencing machines and the generation of longer genomes, as well as genomes that have sequencing errors or evolutionary history.

Other problems with sequence alignment were addressed by either biologists or computer scientists. Problems like poor quality of results with longer sequences; misinterpretation of results that include biological assumptions (such as the gap filling part of the alignment algorithm; as there is no proof exists that these filled gaps are results of possible evolutionary mutations). Other errors resulted from the genomic translocations; reverse subsequence; mutations; or any other errors that would result from non-biological assumptions. Other errors that are difficult to address with the alignment algorithm; are errors resulting from the sequencing machines; these errors come from mutations and/or assembly errors. Another limitation with sequence alignment is the speed issue, but this work addresses and focuses mainly on the quality issue.

To address the quality issue, integration of biological features and computational theories, and understanding the nature of the DNA sequences are the major motivations for the work, with a hypothesis that considering these major factors would enhance the quality of the comparison results.

DNA sequences are not random in their structures, and it is believed that each fragment/subsequence of the DNA sequence carries a message or a signal. The hypothesis used in this research is that closely related or similar genomes would carry similar signals/fragments.

For example, sequences that carry the same restriction enzymes' cut positions [16] might be related and would have similar functions. It would be the same with sequences that carry transcription factor binding sites; other signals would be motifs of specific nature, unique shortest substrings [17] within the sequences, or just motifs with biological relevance that are not known to the literature.

Another feature that DNA sequences has; is that they carry tandem repeats in their structures. These tandem repeats could also be significant signals, and all of these features need to be addressed when comparing the sequences.

A motivation of the comparison problems is based on the fact that similar genomes have similar structures and functions; although subsequences with similar functions do not necessarily have similar exact structures, they carry similar signals within these structures. By identifying these signals, we would be able to classify these genomes and address better measurement for their relatedness.

Notice that these signals might be hidden and/or overlapping with other signals. They might also be of different lengths. To identify these signals or at least take advantage of using them, we need to consider all of the available features. For the previous reasons, we designed an approach that would consider all or a group of prospective signals of specific length k, which could help in addressing the unknown hidden signals. Our approach is variable and would consider different lengths of k, also would consider the overlapping signals.

The challenge of identifying such hidden and unknown signals is not easy. Trying to identify these signals and their functions, taking advantage of their existence and their relevant order within the sequences, and using them for clustering purposes are collectively the focus of this work. The hypothesis of this work is that we can have an approach that takes advantage of these hidden signals within the sequences. Identifying the relatedness among species would be done by considering all the possible chances for the existence of these signals within the sequences and using them to identify the biological distance between the sequences.

Investigating whether or not addressing such signals would improve the clustering process. and reveal a better measurement for the relatedness among species is the focus and challenge of this work. As we consider different signals of different lengths to compare the sequences, we also consider random groups of these signals in order to measure the quality of the results in each case, and to measure whether randomly selected signals would have better results than those that contained all k-mers or signals with biological nature. In addition, this work also considers the use of signals that have biological relevance like restriction enzymes, as well as signals that occur within specific regions that have biological functionality in the DNA sequence, such as those in CDs regions. Finally and as a conclusion of the strength of this approach, applications to datasets with errors were conducted.

III. Experimental Design

The design of the experiment should meet the needed requirements to test the hypothesis. Recall that comparing DNA sequences results in numerical values that represent biological distances between species. These values are subjective with each dataset and would be meaningless if they are not used to address the relationship for the entire group of species.

Verification of the correctness of these distances is not an easy task and simply looking at these numerical values will not reveal the correctness of the results. Hence, we propose another way to measure the correctness.

Clustering the species based on the resulting distances would provide a way to evaluate the correctness of these results. The clustering would be done using bi-clustering algorithms for phylogeny. Using the resulting trees of the phylogeny would be a good way to evaluate the quality of the results. Evaluating the correctness of these trees can then be done by comparing them to known gold standard trees; those are trees that have been verified biologically, hence would be a good proof of the quality of the approach and sequence alignment is used as a baseline for comparison.

The steps necessary to accomplish the proposed experiment in this work are listed as follows:

1. Generate the list of the k-mers. For example, k = 3 for all the possible 3-mers, would result in 64 words (4^3). Alternatively, the list could be a random selection of about 20 percent of all possible words, which would be 13 random 4-mers. It could also be a list of biological signals of different lengths.
2. Convert the DNA sequences according to the compiled list of k-mers (refer to Figure 1).
3. Generate the scoring matrix based on pair-wise comparison, using longest common subsequence (LCS) and Lempel-Ziv complexity of distance measure 2 (LZC).
4. Build the phylogenetic trees using UPGMA and Neighbor-Joining (NJ) phylogenetic algorithms.
5. Repeat Step 4 using the scoring matrix generated by multiple sequence alignment (MSA) [3].
6. Measure the distance between the generated trees and the gold standard tree; the method used to measure this distance is the path-length-difference.

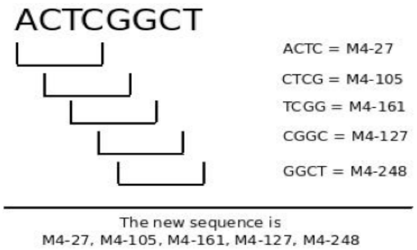

Figure 1. The list on the right side includes the preferred signals to be used for the approach. The sequence on the left is parsed to subsequences each of the same length as the signal lengths on the right, If there is a match, it will be reported with the proper code and in the correct order.

A. Conversion of DNA sequence to sequence of signals

Consider all possible signals of specific length k (all possible k-mers); a production of all the possible combination of the length k is generated. This would result in a word's list of size 4^k, where 4 is the number of the used nucleotides in a DNA sequence (A, C, T, and G).

The content of the generated list is used as the main seeds for the signals needed to be identified within the sequences. We substitute any existence of a signal in the DNA sequence with a unique code, which would save conservation of the order for the signals within the sequences. This design would also save computational time when the list is small and the sequences are longer.

Figure 1 shows how to identify the existence of these signals in the sequences, as well as how to convert the DNA sequence to a sequence of signals/words in the proper order of these signals. The used motifs/signals list in Figure 1 is on the right side of the figure. In this list, each motif/signal has a name (code). The left side of the figure has the original sequence, parsed as words of length k (k = 4 in this case).

The used motifs/signals list in Figure 1 is on the right of the figure. Each motif/signal has a name (code). On the left side is the original DNA sequence. We identify the signals from this list that exist in the sequence. If they occur, their codes would be assigned with the proper order to the new sequence of signals. Thus we convert a DNA sequence to a sequence of signals. Also notice that this approach considers all the overlapped signals.

We also need to mention that occasionally some of these signals do not exist in the sequence, or they occur more frequent. Either way would impact the results of the relatedness between the sequences; this would be a major difference between the converted sequences and would address similarity or dissimilarity among species.

B. The experimental design steps and discussion on the remaining steps

The conversion step is the heart of this work. In this step, we address the signals in preference, but the work remains incomplete as long as there is no way to compare the converted sequences.

The nature of the converted sequences is that they carry two main features. The first feature is a new alphabet of preferred signals; this motivates us to use similar comparison algorithms/approaches as in regular DNA sequences. Simple and efficient algorithms such as the longest common subsequence (LCS) [18] would address the distances between the converted sequences.

The second feature of converted sequences is the conservation of the signals' order within the sequences. This was missed by other research that was also based on k-mers [6, 7]. The order of the signals would have a great impact on the results, with some possibility of having few or more mobile subsequences. We would still be able to use an algorithm that would address the order feature, like Lempel-Ziv Complexity (LZC) [19]. LZC is based on compression complexity and has a great success in identifying the relatedness of different strings. Please notice that LCS addresses the order factor as well.

The comparison method would result in numerical values that represent the distances between species. As previously discussed, it is necessary to cluster the species to measure the correctness of the resulting distances using hierarchical clustering algorithms such as UPGMA and NJ. Therefore the results of these algorithms are in the form of trees. Although most researchers use visual inspection to evaluate phylogenetic trees, we don't recommend it for the following reasons:

- Visual inspection uses personal judgment, and personal judgment is not usually accurate. It can mislead the evaluation process, especially if it is not compared against some reference.
- Visual inspection cannot identify the correctness of trees with large numbers of species. In fact with some trees that have 1,000 species or more, it would be impossible to find out the relationships between each species.
- Visual inspection does not provide numerical value for the comparison. Therefore, no clear decision could be achieved based on its results. Using a computational method to measure the distance of the resulting tree to a reference tree would yield a decision for the entire experiment.

For these reasons, a computational approach to measuring the distance between resulting trees to a gold standard tree was used. This approach is called path-length-difference, and it was modified to give normalized values. Finally, it is important to compare the trees from our approach to the resulting values from MSA and evaluate whether our approach would have better results.

C. *Different algorithms of the experiments*

This subsection describes some of the methods used to verify the hypothesis of this work, specifically methods that are new to the reader or those that have been modified to fit the work.

- *Normalizing Longest Common Subsequences*

LCS is based on dynamic programming and has a well-established reputation and implementations. However, the generated scores are not normalized, and these scores cannot be used to build a phylogenetic tree. To understand this problem, consider these sequences:

S1 : GTTAATGCCACCAAAAAAAAA (length 21)
S2 : GTTAATGCCACCGA (length 14)
S3 : TCCCTAGCT (length 9)

The LCS for all the pair-wise comparisons is as follows:
S1 : GTTAATGCCACCAAAAAAAAA
S2 : GTTAATGCCACCGA
LCS is GTTAATGCCACCA and the score is 13

S1 : GTTAATGCCACCAAAAAAAAA
S3 : TCCCTAGCT
LCS is TTAGT and the score is 5
S2 : GTTAATGCCACCGA
S3 : TCCCTAGCT
LCS is TTAGC and the score is 5

TABLE I. THE SCORES OF USING LCS ON THE EXAMPLE SEQUENCES

	S3	S2	S1
S3	9	5	5
S2	5	14	13
S1	5	13	21

The resulting scores of using LCS for these sequences are shown in Table 1. Note that the scores in the table are not normalized. To address this issue, we divide the resulting score by the length of the shortest sequences of the measured pair. In addition, to simplify the clustering step we reverse the meaning (average) of the score, by subtracting the normalized score from one, so smaller values imply closer species or higher degree of similarity. The resulting values are shown in Table 2 below.

TABLE II. THE RESULTS AFTER USING THE NORMALIZING FUNCTION AS SUGGESTED

	S3	S2	S1
S3	0	1-5/9	1-5/9
S2	1-5/9	0	1-13/14
S1	1-5/9	1-13/14	0

- *Lempel-Ziv complexity:*

Lempel-Ziv complexity of distance measure 2 was used. Please refer to [19, 20] for more details.

- *Path-Length-Difference:*

The comparison between trees was done by estimating the path-length-difference metric [20]. The main concept here is to give penalties for changing relative positions of species in the generated tree, relative to the reference tree (gold standard tree). Each change would make a species closer to a group of species and further from the rest of species; that should cause a penalty value.

The method begins by generating two matrices; one for the generated tree (resulting tree of our approach), and the other matrix is for the gold standard tree. The dimensions of the matrices are *m* x *m*, where *m* represents the number of species for the dataset (or the tree). Each cell has a value

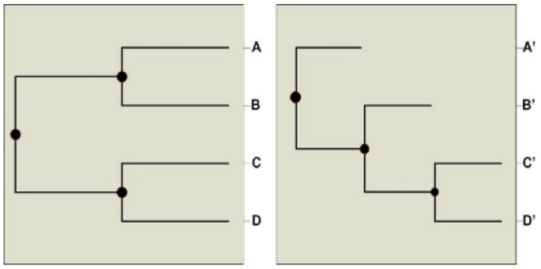

Figure 2. Two trees for the comparison method. The one on the left is the gold standard tree, while the one on the right is the algorithmic tree.

IV. Experiments

The *experiments* are designed and carried out to answer the following questions:

1. Would some motifs/words/signals provide good results for sequence comparison? Would these signals have better comparison results over traditional sequence comparison methods such as those that are alignment based?
2. If the answer for question 1 is positive, is it possible to change the selection of the k-mers for the experiment? Would that enhance the results? In other words, are there certain words/k-mer(s) that would improve the clustering pattern?
3. If the answer for question 2 is positive, would we be able to use signals with biological relevance to improve the results? Like restriction enzymes..etc.
4. If the answer for question 3 is positive, would it be possible to find hidden signals within the sequence with biological relevance and then use them to have valid results?
5. Finally, if the first four questions have been answered positively, is it possible to use the approach on datasets that have errors and still get valid results?

V. Results and Analysis

To answer the above questions, an experiment for each question was conducted. All the experiments follow the same process, as discussed in the Methodology section, with each experiment using a different list of used k-mers.

Datasets:

- The first dataset used was the mycobacterium dataset. We used it for the first three experiments.
- The second dataset was a mitochondrial genomic dataset. We used it for experiments 3, 4, and 5.

A. The First Experiment: Viability of the method

The goal of this experiment is to evaluate whether using generic signals within the sequences would provide good results, as well as whether those results would be better than the results of traditional alignment-based methods (MSA). This experiment deals with all the possible k-mers, as some of them might be hidden signals with strength and within the sequences. The used list of k-mers includes all possible k-mers.

Figure 3 shows the results of using all possible k-mers, and it shows very good results. All distances of any value for k were less than 1.25 percent, while with MSA the results were above 1.8 percent, be aware that smaller values express better distance measures, and are interpreted as closer distance to the gold standard tree.

Figure 3 shows significance in the results using our approach compared to those of MSA. That proves our hypothesis that emphasizing such signals would improve the results and would answer the first question. Please notice that MSA refers to the result of applying Multiple Sequence Alignment, and 7LCS means applying longest common subsequence on k-mer of length 7.

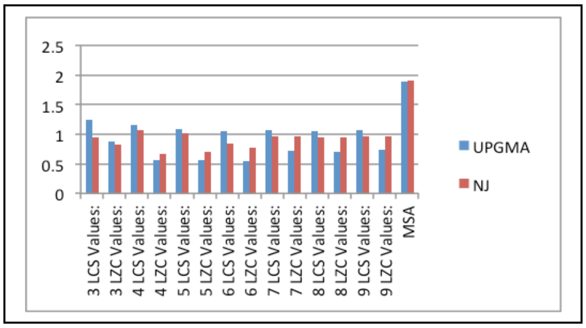

Figure 3. This shows the results of using our algorithm with different parameters. Here, k ranges from 3 to 9. The used methods of comparison are LCS and LZC, and the clustering methods are UPGMA and NJ. The chart shows that in all cases our approach outperformed MSA (multiple sequence comparison) with significant results.

B. The Second Experiment: Using random k-mers

For this experiment, we used lists of random signals selected from all possible k-mers with percentages of 10–90 percent. These selections were applied to k values of 3 to 9, using comparison methods LCS and LZC and clustering NJ and UPGMA. We tested the impact of charging the parameter k on the quality of the results by measuring how similar the obtained phylogenetic tree produced by our approach and MSA from the gold standard tree.

Expectations for the randomly generated lists were that the list carries strong signals, carries weak signals, or carries both. The purpose of carrying on this experiment was to test whether results would be better, worse, or close to those obtained from the first experiment.

Overall, the majority of the experiments we conducted using random k-mers showed better results obtained from our proposed approach as compared to MSA. Some of them were even better than the results obtained in the first experiment. Few cases produced results slightly below the level produced by the first experiment, and some were very close to the results of the first experiment. We tested the approach using the two comparison methods LCS and LZC while using two clustering algorithms NJ and UPGMA.

For example, when using using LCS as a comparison method and NJ as a clustering method. The results of this experiment showed that with just a random selection of k-mers, the approach would still provide better performance than using alignment-based methods for all the values of k that we used. Even with a small list for k-mers (up to 10 percent of all the possible k-mers of specific k), the results would still outperform MSA.

In addition, some runs showed better results than those obtained in the case when for all possible k-mers are used, as in the case with 60 percent random selection of k-mers of length 6. The resulting tree has a distance to the gold standard tree of 0.489 percent, which outperformed any result of all possible motifs, which you can compare by referring to Figure 3. When the random selection of 10 percent for k-mers of length 5, the distance to the standard tree has a value of 1.877005 percent, which is slight worse than any value in Figure 3. That shows that while some signals would do better when they are used alone, others would do slightly worse.

Similar results were produced when the algorithm used LZC as a comparison method and NJ or UPGMA were as the clustering method.

C. *The Third Experiment: Using restriction enzymes' cut positions as the words list*

The second experiment showed that results would be impacted with the selection of the words (k-mers) list, and that some signals would have a higher impact over others. This motivated us to proceed with the third experiment that deals with words that have biological relevance and to see how these words would impact the results. The used signals were obtained from a database of restriction enzymes' cut positions.

Restriction enzymes are special nucleotide signals that cut the DNA double- or single-stranded sequence at specific recognition positions. We believe that DNA sequences that share similar restriction enzymes' cut positions would also have similarities in their functions and structures.

We used restriction enzymes' cut positions that have lengths of 4 to 8 nucleotides. As the number of words for each length was small, we had to use all of them as the words list. Therefore, we used a modified implementation for the conversion algorithm, which would integrate different lengths of the words. The following subsection shows how we modified our conversion approach to take advantage of all restriction enzymes' cut positions.

Since there are a limited number of restriction enzymes, we had to integrate all of them in the converted sequence. To do so, we looked at restriction enzymes of length 4 and identified their locations in the sequences. Then we moved on to restriction enzymes of length 5, 6, 7, and 8. This would give priority to words with shorter lengths first, then move up with longer words.

Again, these words have names/codes in their list, so the generated sequences would have a new alphabet that represents words of different lengths and biological relevance. The rest of the experiment would be the same as in the previous two experiments. The following example shows the new modification for the conversion approach. For example, assume this sequence: ACCGTGC, the restriction enzymes list we have with their codes is:

ACCG	= RE1
CGTG	= RE2
ACCGT	= RE3

Applying the restriction enzymes of length 4 would generate: RE1 (at position 1), RE2 (at position 3), while applying the restriction enzymes of length 5 would generate RE3 (at position 1). The final sequence of restriction enzymes after integrating both lengths would be: RE1, RE3, RE2.

Figure 4 shows the results of using a list of restriction enzymes' cut positions on the mitochondria dataset. The results showed better quality for the application of the restriction enzymes' list than those results from using MSA. Similar results are shown in Figure 9, using mitochondrial genomes. Again the results of the proposed approach outperformed those obtained by MSA.

Figures 4 and 5 show that the results obtained using restriction enzymes are generally better than those obtained using multiple sequence alignment. However in some cases, the random selection (refer to the second experiment) might produce better results, as shown in Figure 4. Using k = 6 and the random selection of 60 percent, we got a 0.489 percent of tree distance difference to the gold standard tree.

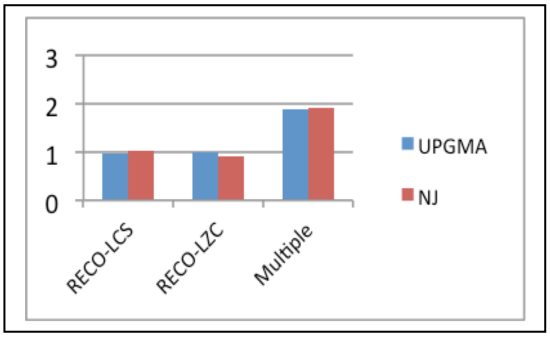

Figure 4. These are the results of using the algorithm with a list of restriction enzymes' cut positions on the mitochondrial dataset and using LZC; MSA results are included for comparison.

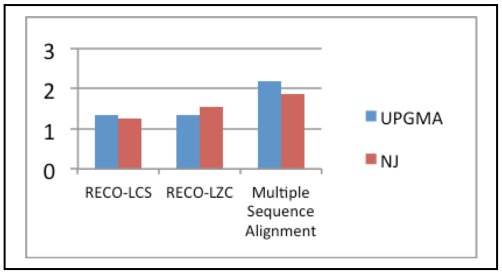

Figure 5. These are the results of using our algorithm with a list of restriction enzymes on the mitochondrial dataset and using LCS. MSA results are included for comparison.

That proves that there are some strong signals known to the literature, and those signals would improve the results of the comparison method. This yields a positive answer for the third question.

D. *The Fourth Experiment: Using k-mers that occur only in CDs regions of the genomes*

As our hypothesis of using words with biological relevance provided promising results, we continued searching for more signals that would also give high quality. One way to find such signals is to use signals/words from the CDs regions of the genomes. As these regions are rich with biological information, we proposed that they would improve the results. In fact, CDs are the main DNA source for functional genes, and a lot of species that are closely related would have similar functions, and in turn genes with similar structures that exist in these CDs regions. Therefore, we generated a list of k-mers that occur in the CDs regions.

We eliminated word lists of lengths 3, 4, and 5, as those lists were all possible k-mers of these lengths and would have same exact results as in the first experiment. The used dataset here was entire genomes. These mitochondrial genomes are rich with CDs regions and were a good fit for this experiment, as they also have a gold standard tree. Figure 6 shows better results when signals from the coding sequences are used in our approach. Figure 6 shows that these signals are rich with information that would improve the quality of the method. Therefore, these signals would be a major source as input lists of the approach. This yields a positive answer for the fourth question.

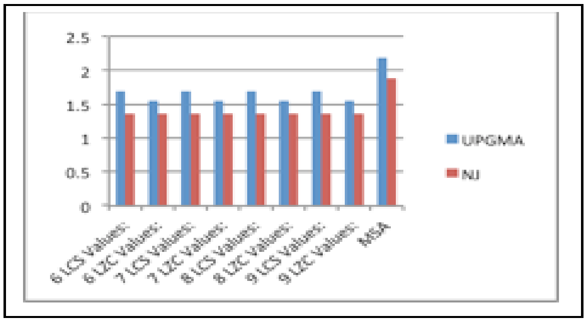

Figure 6. This figure shows the results of using our algorithm with lists that were generated from CDs regions, k ranges from 6 to 9. The used methods of comparison are LCS and LZC, and the clustering methods are UPGMA and NJ.

E. *The fifth experiment: Application of the approach to datasets with different level of gaps errors*

We finally applied the approach to special datasets. These datasets were generated and manufactured from the mitochondrial genome dataset. They are incomplete genomes and/or with errors. The reason for applying the approach to such datasets is to measure if it would be possible to identify the relatedness among species with errors.

These datasets are divided into three categories. The first category is for a dataset where each sequence is a fragment from the original genome, each fragment's content is a percentage of the original genome's content, and was chosen randomly from the genome's content. For this category, we generated two datasets: one with 50 percent content, and the second for 70 percent content.

The second group was for datasets composed of several fragments from the original genomes, and these fragments are in order. So each sequence would be the merging of several fragments from the original sequence, and these fragments would have a content represented as a percentage of the original genome. These fragments were chosen randomly from the genome's content and did not overlap. For this category, we generated two datasets with percentages 30 percent and 90 percent.

The third category is similar to the second one, but the fragments were switched randomly. This means that now a sequence has fragments that are not in order, yet would still have a content that is represented as a percentage amount of the original genome. These datasets were generated with percentages of 40 percent and 80 percent.

We compare the results of using our approach on these datasets to those resulting from MSA on the same datasets. We are evaluating whether our approach would identify the relatedness of the species in these datasets, even if they have errors, and whether these results would be better than those of MSA.

Figure 7 shows the results of applying the approach to these datasets. Each group of columns (blue, red, and green) represents one dataset and the use of one clustering algorithm (NJ or UPGMA). Each column shows the result of using LCS, LZC, or MSA. The results show that in most cases our approach outperforms MSA, except in two cases. As with the dataset of using several fragments, with 30 percent contents of the original sequences, and using all possible 4-mers, and LCS comparison method with UPGMA clustering algorithm, the quality of result was lower than MSA, same with the dataset of (80% contents, several fragments not in order, 6-mers selected from CDs and using LCS and UPGMA), the result again was lower than MSA.

Using the motif-based approach for comparing sequences in datasets that contain errors would be more effective than using MSA, as most of the results of our approach outperformed MSA results. Thirty-four results were better than MSA out of 36 runs (94.44 percent).

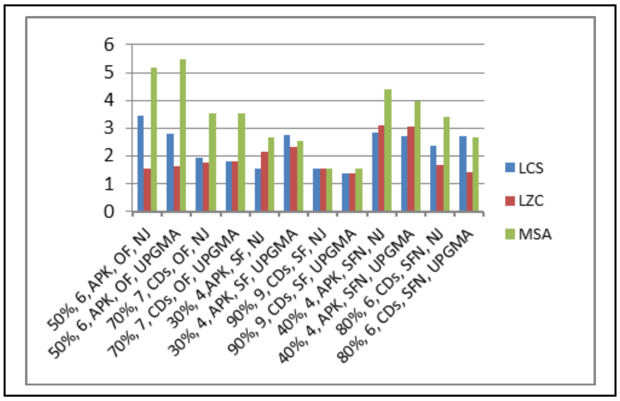

Figure 7. The results of applying proposed approach to datasets with high degree of errors. Abbreviations: APK(All Possible K-mers), CDs(Coding Regions), OF(One Fragment), SF(Several Fragments), SFN(Several Fragments Not in order).

VI. CONCLUSIONS

In this paper, we introduced an alternative approach to compare biological sequences, and that method in several cases outperforms the traditional alignment-based approaches. The proposed method is developed to compare sequences based on their inclusion of short biological signals or motifs. The conducted experiments showed that the effectiveness of the approach depends on which motifs used. In particular, the results showed that with a number of biologically significant signals/words, we should be able to produce results far superior than those obtained using alignment.

The proposed approach produced comparable results, and even better in certain cases, when random or generic motifs are used to compare sequences. Better results were obtained when certain motifs were used. Future work should focus on identifying different types of biological signals that can be utilized for better classification.

When we used short motifs associated with biological significance, such as restriction enzymes, the motif-based comparison produced even better results. Similar results were obtained when motifs are selected from rich regions in the genome such as coding regions. These motifs made it possible for the approach to outperform traditional methods like MSA in identifying the relatedness between genomes.

We also compared the proposed method with MSA using datasets that contain sequencing errors, and again for the majority of the cases, the signals-based comparison produced better results and better identified relationships among species. So for genomic datasets that have errors, it would be better to use this approach instead of using traditional alignment-based methods. The overlapped signals would identity such relationships among species. Overlapped signals are powerful marks for comparing biological sequences and would identify more accurate relationships between species as compared to alignment-based methods.

REFERENCES

[1] S. Vinga, J. Almeida, “Alignment-free sequence comparison – a review,” Bioinformatics, 19(4), pp.513–23, 2003.

[2] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, “Basic local alignment search tool,” J Mol Biol 215 (3), pp. 403–410, 1990.

[3] R. Edgar, “MUSCLE: Multiple sequence alignment with high accuracy and high throughput,” Nucleic Acids Research 32(5), pp. 1792-9, 2004.

[4] A. Subramanian**,** M. Kaufmann, and B. Morgenstern, “DIALIGN-TX: Greedy and progressive approaches for segment-based multiple sequence alignment,” Algorithms for Molecular Biology, 3:6, 27 May 2008

[5] K. Song, J. Ren, G. Reinert, M. Deng, M. S. Waterman, F. Sun: "New developments of alignment-free sequence comparison: measures, statistics and next-generation sequencing", Brief Bioinform (2013) doi: 10.1093/bib/bbt067

[6] G. Lu, Sh. Zhang, and X. Fang: “An improved string composition method for sequence comparison,” Symposium of Computations in Bioinformatics and Bioscience (SCBB07) Iowa City, 28 May 2008.

[7] O. Bonham-Carter, J. Steele and D. Bastola**, “**Alignment-free genetic sequence comparisons: a review of recent approaches by word analysis”, Briefing in Bioinformatics, August 23, 2013

[8] Z. Liu, J. Meng, X. Sun, "A novel feature-based method for whole genome phylogenetic analysis without alignment: application to HEV genotyping and subtyping." Biochem Biophys Res Commun 2008; 223:223–30.

[9] V Arnau, M Gallach, I Marín. “Fast comparison of DNA sequences by oligonucleotide profiling.” BMC Res Notes 2008;1:5.

[10] T. J. Wu, Y. H. Huang, L. A. Li . "Optimal word sizes for dissimilarity measures and estimation of the degree of dissimilarity between DNA sequences." Bioinformatics 2005; 21:4125–32.

[11] G. E. Sims, S. R. Jun, G. A. Wu, SH. Kim. “Alignment-free genome comparison with feature frequency profiles (ffp) and optimal resolutions.” Proc Natl Acad Sci USA 2008;106:2677–82.

[12] J. Lin, "Divergence measures based on the shannon entropy." IEEETrans InfTheory 1991;37:145–51.

[13] Z. G. Yu, L. Q. Zhou, VV Anh, "Phylogeny of prokaryotes and chloroplasts revealed by a simple composition approach on all protein sequences from complete genomes without sequence alignment." J Mol Evol 2005;60:538–45.

[14] R. H. Chan, T. H. Chan, H. M. Yeung and R. W. Wang, “Composition vector method based on maximum entropy principle for sequence comparison.” IEEE/ACM Trans in Comput Biol Bioinf, Mar 3, 2011.

[15] I. Soares, A. Goios, A. Amorim. “Sequence comparison alignment-free approach based on suffix tree and l-words frequency.” SciWorldJ 2012;2012:450124.

[16] R. Sengupta, D. Bastola, H. Ali, “Classification and identifMSAcation of fungal sequences using characteristic restriction endonuclease cut order,” J. Bioinformatics and Comp Biology, pp. 181–198, 2010.

[17] B. Haubold, N. Pierstorff, F. Möller, and T. Wiehe, “Genome comparison without alignment using shortest unique substrings,” BMC Bioinf, 6:123, 23 May 2005.

[18] S. Aluru, Handbook of Computational Molecular Biology, pp. 15–10, 2005.

[19] H. Otu, K. Sayood, “A new sequence measure for phylogenetic tree construction,” Bioinformatics Vol. 19, no. 16, pp. 2122–2130, 2003.

[20] R Mina, D Bastola and H Ali, “Compression- based Algorithms for Comparing Fragmented Genomic Sequences”, BIOTECHNO 2013, The Fifth Int Conf on Bioinformatics, Biocomputational Systems and Biotechnologies, Lisbon, April 2013.