Alignment Free Methods in Bioinformatics

In this paper we present a discussion of the methodologies adopted by various types of alignment free methods in bioinformatics.

Kazi Tasnim Zinat(1017052027) and Zarin Tasnim Promi (1017052044)

Department of Computer Science & Engineering, Bangladesh University of Engineering & Technology.
zinat36@gmail.com

Department of Computer Science & Engineering, Bangladesh University of Engineering & Technology.
zarinpromi003@gmail.com

Abstract

Gemone and metagenome sequence analysis has made it's mark in the field of bioinformatics providing better understanding of how genomes evolve and their relationship, genome annotation, molecular phylogeny etc. With the emergence of next generation sequencing techniques the amount of sequence data has increased in manifolds. Though alignment based methods have been used for genomic sequence analysis for years , the gigantic amount of data poses a challenge as alignment based methods are time consuming and computationally expensive. Moreover, when the sequences are divergent, alignment based methods fail to produce reliable results. The Alignment free methods provide a solution to this shortcomings of alignment based methods by being computationally efficient. These approaches are now applied to extract metagenomic information, identification of horizontal gene transfer, infer correlation between host and parasites, detection of gene regulatory regions and many other cases. As there's a wide variety of methods that can be enlisted as alignment free approach, here we provide a brief review of them in broad categories and their applications as well.

Keywords

Sequence comparison,Alignment, Alignment free method, Phylogeny .

Introduction

Sequence alignment, one of the key research phenomenon in the study of computational biology, is the process of organizing sequence data in such a way, that similarity regions within the sequences can be found. There exists a handful of directions in which sequences can be alligned- the alignment can be global(alignment on full length of given sequences), local (alignment on substrings of the given sequences), pairwise (piece-wise best matching of given two sequences), or multiple sequence alignment (aligning three or more sequences together). Many existing alignment-based methods provides the information of match/ mismatch regions within sequences, some methods also permit gaps and indels (insertions or deletions) within the sequences for producing better results. Available Alignment based tools can be categorized as Similarity finders (BLAST, FASTA), Multiple sequence aligners (ClustalW , Muscle , MAFFT) & Whole-genome aligners (BLASTZ, TBA). These tools are significant for the discovery of tasks of varoious genes and proteins.

However, as the knowledge and understanding about evolutionary scenarios and patterns improved, some downsides of alignments based sequencing methods were unveiled.

Limitations of Alignment Based Approaches

Though alignment based methods have been used for years for sequence assembly, annotation and comparison purposes. However, as more advanced information about complex evolutionary phenomenons, patterns and properties of biological sequences were discovered, more about the limitations of alignment based methods uncovered. Here we discuss the shortcomings of methods based on alignment.

Assumption of colinearity

Alignment based methods take up a assumption that homologous sequences contain linearly arranged and conserved sequence sections. They often overlook shuffling and recombination and evolutionary changes. For instance viral genome exhibit great variation in the number and order of genetic elements due to their high mutation rates, frequent genetic recombination events, horizontal gene transfers, gene duplications, and gene gains/losses. Another instance is proteins where the linear and modular organization is not always preserved due to frequent domain swapping, or duplication or deletion of long peptide motifs.

Lack of reliability

Reliability of sequence alignments cannot be determined if sequence identity is lower than a threshold. For instance Protein sequences comprise of 20 different amino acids. So two different sequences can have matching upto 5% . By allowing gaps the percentage can be upto 25%. The area of 20–35% identity is known as the twilight zone, whereas below 20% matching is called the “midnight zone”. So Homologous relationships determined by plain pairwise alignments is not reliable. This limitation is more acute in the annotation of protein superfamilies where the members retain structural kinship even though the average matching between sequences is 8–10%. For RNA/DNA, the accuracy of the alignments is even more disappointing. For instance, two random DNA/RNA sequences can show up to 50% sequence identity when gaps are allowed, and the edge of the twilight zone can encompass nucleotide matches of up to 60–65%.

Time and memory exhaustive

Alignment based methods are memory intense and time consuming, are of restricted use when dealing with large scale data. The number of possible alignments of two sequences grows rapidly with the length of the sequences. There is about $10^{(}$60) possible alignments for two sequences of length 100. Though some methods introduces dynamic programming, they are still Computationally onerous. Despite there being so many alignment based tools the problem of long sequence alignment is not fully resolved. In addition, available sequence comparison models may not directly apply to complete genomes.

Computation of accurate MSA is NP-hard

The computation of an accurate multiple sequence alignment is NP-hard. Many speed optimized faster method apply heurictics , often causing inaccuracy which affects the quality of many downstream analyses such as phylogenetic.

Alignment Free Sequencing Methods

Any process of computing sequence similarity that does not apply alignment at any step of the algorithm can be considered as alignment free method Unlike methods entirely relying on alignment of sequences alignment-free methods does not depend on Dynamic Programming, as a result they are computationally less demanding and therefore fit for whole genome comparisons. Besides, they are resilient to shuffling and recombination events and applicable when low sequence conservation cannot be handled. Alignment free methods do not depend on evolutionary change assumptions.

In the following section, we discuss the categories of alignment free methods.

Categories of Alignment Free Methods

Alignment free methods can be broadly classified in 5 different groups

- Methods based on K-mer frequency
- Methods based in Sub-string
- Methods based on information theory
- Methods based on graphical representation

- Methods based on Sequence representation designed by chaos theory

The working procedures adopted by different categories are explained in the subsequent sections.

Methods based on K-mer frequency

The principle of k-mer frequency based methods is that similar sequences share k-mers (subsequences of length k) among themselves. Computation of word frequency among given sequences provides measure of their similarity.
The key steps of these methods are noted below.

- Detection of unique words for given length K
- Transformation of each sequence into a vector
- Quantification of sequence similarity/ dissimilarity, where the distance between two sequences is zero if they are identical.

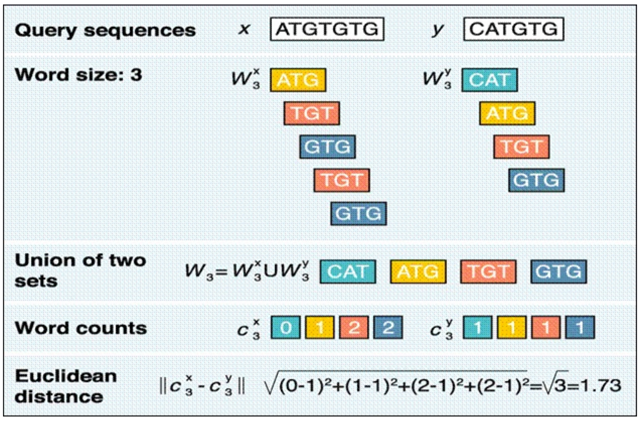


Figure 1. Alignment-free calculation of the word-based distance between two sample DNA sequences using the Euclidean distance

Word-based alignment-free algorithms vary at the three basic steps, resulting in different methods. Now we take a look at some of the major word-based alignment free methods.

FFP (Feature Frequency Profile)

In this method, all possible k-mers are extracted first, then a Feature Frequency Profile is constructed based on K-mer count in each sequence normalized by total k-mer count for that specific sequence. The distance between the feature frequency vector of two sequences is measured by Jensen-Shannon divergence.

Composition Vector

Similar to the previous method, the frequencies of all possible k-mers are calculated. Then a vector for each given sequence is generated by subtraction of random background of frequencies using Markov model, ensuring reduction of random neutral mutation and emphasizing on selective evolution. A composition Vector is formed keeping normalized frequency in fixed order. The distance between the vectors is measured by Cosine Distance.

Return Time Distribution

Instead of keeping track of k-mer frequencies the time required for the reappearance of a K-mer is recorded. Here, time implies number of residues between successive appearance of a k-mer. The vector is formed with Standard deviation & mean from time sequence of k-mers. The distance between the vectors is measured by Euclidean Distance.

Choosing the value of K for word based methods

Choosing the value of the length of k –mers plays a crucial role in methods based on word frequency. The basis for choosing k is that the words are unlikely to commonly appear in a sequence. Rule of thumb in this case is given below.

- Smaller k-mers for unrelated sequences.
- Longer k-mers for similar sequences.

In practice the word size (k) of 2–6 residues produces optimal protein sequence comparisons. For genes or RNA, value of k can safely be set to 8–10. K value of 9–14 bases is optimal for general phylogenetic analyses. And optimal K value is up to 25 bases in case of comparison of isolates of the same bacterial species

Methods based on sub-String

These methods were adopted from String processing techniques that are applied on Strings in the field of Computer Science.

Average common substring (ACS)

This method works by calculating longest substring lengths starting at different positions in first sequence and having exact matches at some positions in the second sequence.

k-mismatch average common substring approach (kmacs)

This method is the generalized form of ACS, tolerating up to K mismatches between the chosen longest substrings.

Mutation distances (Kr)

Computes the number of changes in each position between two DNA sequences using the shortest absent substring(SHUstring).

Methods based on information theory

The procedures based on this concept analyzes the quantity of information shared between two biological sequences. The information can be represented by entropy or complexity of the sequences.

Methods measuring information via complexity

This procedures are established on the idea that length of a compressed sequence gives an approximation of its complexity. The sequences being compared to are concatenated to create one longer sequence. If the sequences are similar, then the complexity (compressed length) will be very close to the complexity of the individual sequences. If the sequences are dissimilar, then the complexity will tend to the cumulative complexities of the individual sequences.

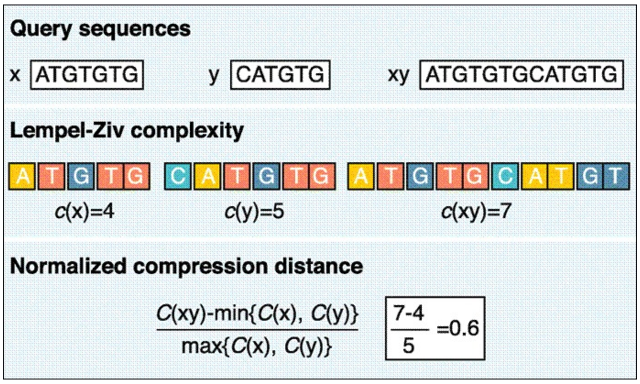


Figure 2. **Lempel–Ziv complexity**

Methods measuring information via entropy

Kullback and Leibler introduced a relative entropy measuring similarity of sequences. The procedure involves the calculation of the frequencies of symbols or words in a sequence and the summation of their entropies in the compared sequences.

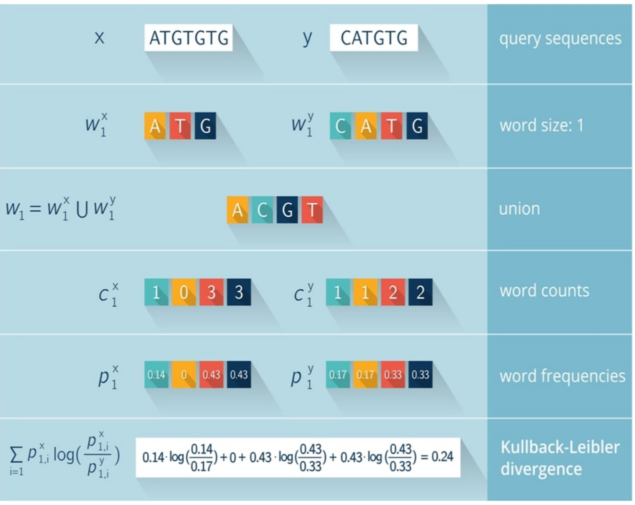

x ATGTGTG
y CATGTG
query sequences
w_1^x A T G
w_1^y C A T G
word size: 1
$w_1 = w_1^x \cup w_1^y$
A C G T
union
c_1^x 1 0 3 3
c_1^y 1 1 2 2
word counts
p_1^x 0.14 0 0.43 0.43
p_1^y 0.17 0.17 0.33 0.33
word frequencies
$\sum_{i=1} p_{1,i}^x \log(\frac{p_{1,i}^x}{p_{1,i}^y})$
$0.14 \cdot \log(\frac{0.14}{0.17}) + 0 + 0.43 \cdot \log(\frac{0.43}{0.33}) + 0.43 \cdot \log(\frac{0.43}{0.33}) = 0.24$
Kullback-Leibler divergence

Figure 3. **Kullback-Leibler Divergence**

Complexity and entropy have a relationship even with their methodical differences. For instance, a low-complexity sequence will have smaller entropy than a high complexity sequence.

Methods based on Graphical Representation

Iterated Maps

Iterated maps provide a bijective map between the symbolic space and numeric space

Methods based on Sequence representation designed by Chaos Theory

Chaos game representation (CGR) technique offers scale independent representation for genomic sequences. The CGRs can be divided by grid lines where each grid square denotes the occurrence of oligonucleotides of a specific length in the sequence.

Available Tools

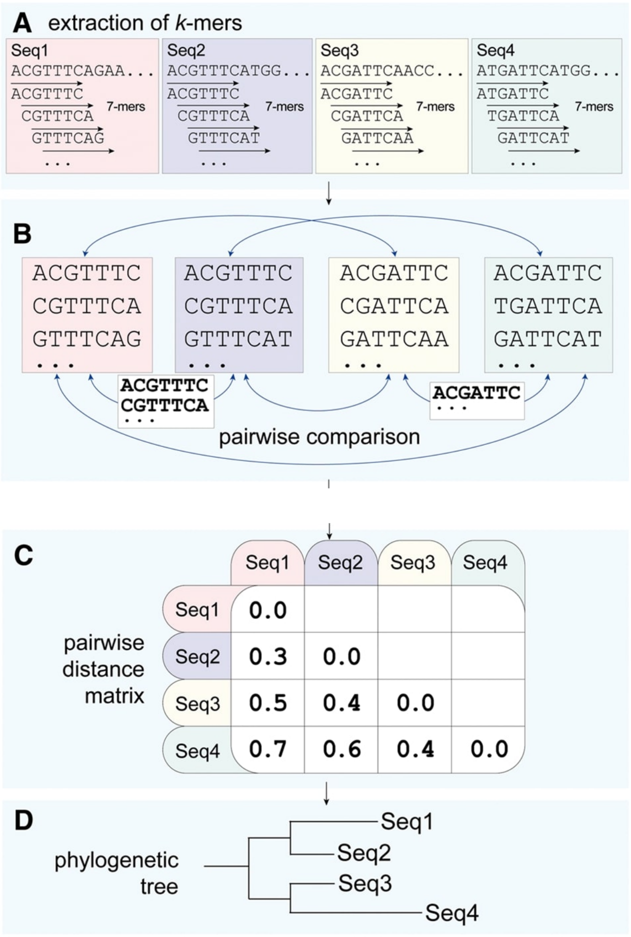
- Sequence Mapping
  - Transcript Abundance Quantification
    Kallisto, RNA-skim
  - Variant Calling
    ChimeRScope, LAVA, FastGT
  - Overall Mapping
    Minimap
- Assembly
  - De-novo Assembly
    MHAP, Miniasm, Links
  - Reads Error Correction
    Trowel, Lighter

Applications of alignment free methods

Alignment free methods have gone from being a curisity to an important sequence analysis tool in bioinformatics over the last 30 years. There are now almost 100 methods with numerous applications in various fields.

Phylogeny

Inference of genealogical relationship must be based on homologous elements, hence the identity in sequences. Frequency methods can be of help to find homologous elements. As sequences diverge over time from a common ancestor, the tend to share shorter K-mers. By disallowing mismatch, degeneracy and indels, k-mer statistics become simpler and the computation more efficient.
Phylogenetic inference based on k-mers consists of 4 major steps. First is the extraction of K-mers from the sequences. We define the substrings of length k from the sequences.Then pairwise comparison of the sequences is done. From the comparison a dstance matrix is computed. In the example we see an elastration of the steps. There are 4 sequences, which are sorted into k-mers of length seven. Then pairwise comaprison is done for each of the sequences. From the comparison a distance matrix is constructed. At the final step a phylogenetic tree is constructed using distance algorithms such as neighbour joining etc.



	Seq1	Seq2	Seq3	Seq4
Seq1	0.0			
Seq2	0.3	0.0		
Seq3	0.5	0.4	0.0	
Seq4	0.7	0.6	0.4	0.0

Figure 4. **Phylogeny**

Sequence Classification

Remote sequences that evolve beyond recognization are one of the most classic applications of alignment-free methods. For instance, alignment-free approaches successfully recognized unknown G-protein-coupled receptor sequences that could not be assigned to any previously known receptor family. Another rising trend for the use of word-based alignment-free methods is the detection of functional and evolutionary similarities among regulatory sequences such ass promoters, enhancers, and silencers to estimate their in vivo activities in different organisms eg flies and mammals, including humans etc.

Sequence Assembly

The assembly of the newly sequenced genomes is one of the most demanding tasks today. It requires an error correction and construction of the genome scaffold based on sequence overlaps. Several alignment free tools give correct sequencing reads which are fast and memory efficient. Example: Quorum , Lighter , Trowel. The reconstruction of whole-genome of E.coli O104:H4, which caused the 2011 outbreak in Germany revealed a direct line of ancestry leading from a putative typical enteroaggregative E.coli ancestor through the 2001 strain to the 2011 outbreak strain.

HGT

Horizontal gene transfer is referred as the phenomenon where genetic material is shared between unrelated organisms. This suggests a parallel evolutionary history. The discovery of similar regions of DNA between two enormous genomes is not a trivial task, and is determined from word frequency data.

Performance Analysis of Alignment free methods

The introduction of first alignment free method was done 30 years ago. Today there are almost 100 published methods. But the lack of benchmarking approaches to alignment free comparison exists. With every new method there is a new evaluation procedure and/or selected dataset introduced. The majority of algorithms have been evaluated using various sets of simulated DNA sequences, selected plant genomes, small subsets of homologous genes etc. The first benchmark, by Vinga and Almeida in 2004, evaluated the accuracy of six word-based methods in recognition of structural and evolutionary relationships among proteins. Höhl and Ragan compared the accuracy of nine alignment-free methods in the construction of phylogenetic trees using homologous proteins representing a wide range of phylogenetic distances. Both research groups showed that, in general, tested alignment-free methods can be as good as alignment algorithms. May perform even better for protein sequences that underwent domain shuffling events. Dai and colleagues (2008) tested nine alignment-free distance measures and two alignment-based approaches (Needleman–Wunsch and Smith–Waterman alignment methods) in annotation of functionally related regulatory sequences in human and fly. Virtually all tested alignment-free methods perceived statistically relevant similarities in sequence compositions whereas alignment-based methods showed only limited correspondence recognizable by alignments. In recent becnhmark, Bernard and colleagues (2016) used simulated and empirical microbial genomes to test the sensitivity of nine alignment-free methods under different evolutionary schemes. All approaches generated biologically meaningful phylogenies. Alignment-free methods were most sensitive to the extent of sequence divergence, less sensitive to low and moderate frequencies of horizontal gene transfer, and most robust against genome rearrangements.

Zielezinski, Vinga (2017) compared 33 alignment-free methods, 25 word-based and 8 information theory-based and Smith–Waterman algorithm in the classification of structural and evolutionary relationships between protein sequences from the SCOPe/ASTRAL database. This resource provides a high-quality structural classification of proteins at four levels: class, folds, superfamilies, and families. The alignment-based algorithm (Smith–Waterman algorithm) was outperformed at all levels by two word-based measures: normalized Google distance and Bray–Curtis distance. These results support the assumption that alignment-free methods can produce more accurate results than alignment-based solutions when applied to homologous sequences of low similarity. Also word-based methods achieved higher accuracy than information-theory based solutions. The run time for the calculation of approximately 22 million pairwise protein comparisons by the Smith–Waterman algorithm took exactly 3 days, which was more than 1000-fold slower than the alignment-free methods. On average, these methods need 4 minutes to complete the task, and the fastest approach (Hamming distance) ran the analysis in only 19 seconds.

Conclusions

As research on alignment free method is still at developing stage, the alignment based methods has broad applications in the reconstruction of ancestral DNA sequences, determining the rate of sequence evolution, and homology-based modeling of three-dimensional protein structures and many other fields. Lack of well-defined benchmarking methods inhibits the process of finding perfect tool for a specific job Using lengthy k-mers in word-based methods might inflict a significant memory overhead (the total number of possible DNA words of length 14 is 414, which is about 4 GB). Although information-theory methods that are based on the compression algorithms are more memory efficient and computationally inexpensive, they may fail to decipher complex organization levels in the sequences. Alignment-free algorithms are applicable in resolving problems in phylogenomics and horizontal gene transfer, population genetics, evolution of regulatory sequences, and links between the genome and epigenome. Disadvantages of next-generation sequencing data processing and analysis seem to be particularly well addressed by the alignment-free methods . The currently dominant k-mer approaches are bound to novel measures for biological applications (e.g., Google distance) and application of advanced information theory-based methods should improve the available alignment-free and alignment-based tool box.

References

- Vinga, S; Almeida, J (2003). "Alignment-free sequence comparison-a review".
- Song, K; Ren, J; Reinert, G; Deng, M; Waterman, MS; Sun, F (2014). "New developments of alignment-free sequence comparison: measures, statistics and next-generation sequencing".
- Haubold, B (2014). "Alignment-free phylogenetics and population genetics".
- Bonham-Carter, O; Steele, J; Bastola, D (2013). "Alignment-free genetic sequence comparisons: a review of recent approaches by word analysis".
- Zielezinski, A; Vinga, S; Almeida, J; Karlowski, WM (2017). "Alignment-free sequence comparison: benefits, applications, and tool".
- Ren, J; Bai, X; Lu, YY; Tang, K; Reinert, G; Sun, F (2018). "Alignment-free sequence analysis and applications".