Different Algorithms for Alignment Free DNA Sequence Comparison in Current Scenario

Kshatrapal Singh$^{1*[0000-0002-5965-0783]}$, Raja Sarath Kumar Boddu^{2}

Primary affiliation
[1, 2,]School of Computer Engineering, Lincoln University College, Petaling Jaya, 47810, Malaysia
pdf.kshatrapal@lincoln.edu.my
Secondary affiliation
[1]Department of CSE, KCC Institute of Technology and Management, Greater Noida, 201308, India,
mekpsingh1@gmail.com
[2]Department of AI&ML, Raghu Engineering College, Vishakhapatnam, 531162, India
iamrajaboddu@gmail.com

Abstract: The method of determining the precise arrangement of the four chemical bases (A, T, C, and G) in a DNA molecule—which contains the genetic instructions for every living thing—allows scientists to interpret the code of life, comprehend gene function, identify illnesses, and monitor evolution. Alignment free sequence studies used to solve challenges varying from complete-genome phylogenetics to protein family categorization, horizontally transmitted gene identification, and rejoined sequence discovery. These approaches are very effective for processing and analyzing data from next-generation sequencing because of their strength. Many researchers, however, are unsure of how these approaches function, how they collate to alignment-based approaches, and how they may be used in their research. We answer these issues and present an overview of the alignment-free sequence analysis algorithms that are currently available. Next-Generation Sequencing (NGS) is a novel method that much outperforms Sanger sequencing by enabling fast, high-speed reading of large volumes of DNA.

Keywords: Biological sequences; Information Theory; Euclidean Distance; Phylogeny; Data mining

1. Introduction

The introduction of sequence comparison methods changed the computation as well as molecular biology sciences in the 1990s, making bioinformatics a thriving field. Many computational biologists rose to prominence at the period by inventing programmes for sequence alignment, which is a tool for identifying regions of similarity in biological sequences that could have implications for functional, structural, or evolutionary relation [1].

Any technique of assessing sequence similarity which does not utilize or create alignment at any phase of the algorithmic application is known as alignment-free techniques to sequence analysis. Due to the lack of dynamic programming, alignment-free methods are computationally less expensive and thus suited for complete genome comparisons. Alignment free approaches are robust to shuffle and recombine events, as well as are useful when alignment could not consistently handle low sequence conservation [2, 3]. Finally, unlike alignment based approaches, they are not predicated on assumption about sequence change evolutionary paths.

Approaches on the basis of frequency of sub-sequences of a specific count (word centric approaches) and approaches that find the informational content among full length sequences (information theory centric approaches) are the two types of alignment free approaches. There are also techniques that cannot be classified into either group, such as those based on the count of matching words (common, longest common, or minimal absent words among sequences), chaos game representation, iterated maps, and graphical representation of DNA sequences.

All of the alignment free methods calculate pair wise measure of dissimilarity or distance among sequences and are mathematically effectively established in domains of linear algebra and information theory [4, 5, 6, 7]. Most of these metrics may be easily entered into common tree-building tools like Phylip or MEGA, which is quite convenient.

2. Algorithm for word frequency based approach

The logic behind such approaches is straightforward: similar sequences have similar words/k-mers, so mathematical functions on the frequencies of the words provide a decent comparative scale of sequence dissimilarity [8, 9, 10, 11]. The approach is closely linked to the concept of genetic patterns, which were first developed for dinucleotide compositions and have now been expanded to include prolonged words. This procedure can be divided into three distinct parts (Table 1).

The sequences to be analyzed first be broken into groupings of distinct words of a predetermined size. Two DNA sequences, a = ATGCATC and b = CATATG, and a word size of 3 nucleotides (3-mers), for example, create two sets of unique words: W_a^3 = ATG, TGC, GCA, CAT, ATC and W_b^3 = CAT, ATA, TAT, ATG. We build a complete set of words that belongs to at least W_a^3 or W_b^3 more to streamline the computations, leading in the combine set W_k^3 = ATG, TGC, GCA, CAT, ATC, ATA, TAT, because certain words are frequently included in one sequence but not in the another.

The next pace is to change every sequence into a vector (an array of integers) (for example, by counting how many times each word (from W_3) appears inside the sequences). We find two real-valued vectors for sequences a and b: O_a^3 = (1, 1, 1, 1, 1, 1, 0, 0) and O_b^3 = (1, 0, 0, 1, 0, 1, 1). The final step involves applying a distance function to the sequence representation vectors O_a^3 as well as O_b^3 in order to quantify the dissimilarity among sequences. The Euclidean distance is a popular method for calculating this difference, however some metric can be used. The larger dissimilarity number, the further apart the sequences are; consequently, two identical sequence will have a distance of zero between both.

Table 1: Alignment free comparison of word based distance among 2 DNA samples.

DNA Sequences	a = ATGCATC	b = CATATG
K = 3	W_a^3 ATG TGC GCA CAT ATC	W_b^3 CAT ATA TAT ATG
Union of 2 sets $W_k^{3} = W_a^{3}$ U W_b^{3}	ATG TGC GCA CAT ATC ATA TAT	
Word Counts	O_a^3 = (1, 1, 1, 1, 1, 1, 0, 0)	O_b^3 = (1, 0, 0, 1, 0, 1, 1)
ED	$\|\| O_a^3 - O_b^3 \|\| = sqrt(0+1+1+0+1+1+1) = \sqrt{5}$ = 2.23	

You can test various word length approximations, but it's vital to pick words that aren't likely to occur in a sequence frequently (the smaller the word, the further possibly it is to look randomly in a series).

3. Algorithm for information theory based approach

The quantity of information shared among two studied biological sequences is recognized as well as computed using information theory-based approaches. Nucleotide as well as amino acid sequences are essentially strings of symbol, and their digital arrangement can be easily deciphered using information theory ideas [12, 13, 14]. It is extremely simple to calculate a distance among sequences employing complexity (compression) (Table 2).

This process concatenates the two sequences that are being evaluated (a = ATGTGTG and b = CATGTG) to get one larger sequence (ab = ATGTGTGCATGTG). If a and b are identical, ab's complexity (compression length) will be quite near to the complexity of the individual a or b. If a and b are not similar, however, ab complexity will tend to be greater than the sum of a and b complexities. Naturally, there are more alternative information based distances as there are complexity assessment techniques. LZ complexity [15, 16, 17], such as, is a prominent metric that estimates the count of unlike subsequences encountered when examining a sequence from start to finish. One time sequences' complexity have been determined, a measure of their differences can be simply calculated. Currently, various DNA-specific compression techniques are being used to solve latest kinds of challenges.

Table 2: Alignment free estimation of CD using LZ algorithm

DNA Sequences		
a = ATGTGTG	b = CATGTG	ab = ATGTGTGCATGTG
LZ Complexity		
C(a) = 4 A T G TG	C(b) = 5 C A T G TG	C(ab) = 7 A T G TG C AT GT
Compression Distance (CD)		
	$\frac{(C(ab)\ -\ \min\{C(a), C(b)\})}{\max\{C(a), C(b)\}}$	(7-4)/5 = 0.6

4. Machine Learning Algorithms for DNA Sequencing

We have seen the tremendous expansion of biomedical data and the innovative advancement of biotechnology and biomedical research during the last few decades. The growing amount of biological data is no longer the issue; instead, it is how to extract valuable insights from the data. On one hand, a difficult new area called bioinformatics has emerged as a result of rapid growth of biotechnology and biological methodologies for analyzing data. However, the ongoing advancement of biological data mining methods has led to the invention of numerous efficient and scalable algorithms. It is worthwhile to focus on and do research on how to effectively process biomedical data by bridging the two domains of machine learning and bioinformatics [25-26]. We should specifically examine how data mining may be used to evaluate biomedical data effectively (Figure 1) and formulate certain research objectives that could encourage the development of more potent biological machine learning algorithms.

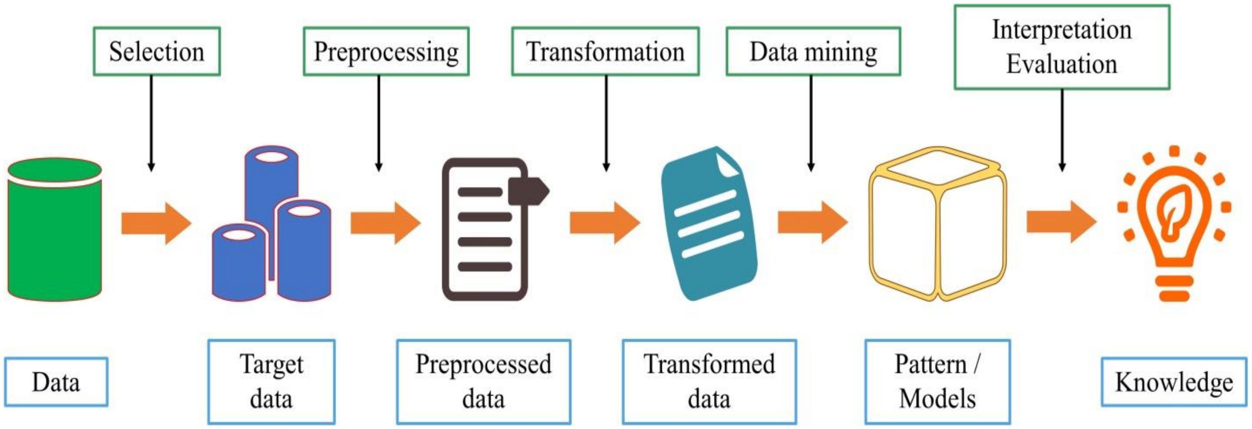

Figure 1: Sequence data mining of DNA.

4.1 Association Rule Mining Algorithm

Association rule mining, one of the most significant subfields of data mining, can find frequently occurring trends and associations among a group of items in a certain database. There are two smaller issues with it: (1) Use the minimal support and set the minimum support criterion. Locate often occurring itemsets in the database; (2) Utilize minimum confidence to identify association rules that meet predetermined requirements for frequently occurring item-sets. In addition to being crucial for corporate data analysis, association rule mining has shown effective in a variety of other domains, including shopping basket analysis and healthcare data analysis. The prediction of protein structures and sequence pattern mining are two uses for the Apriori algorithm, a common association rule-based mining technique. Apriori is the foundation for many machine learning techniques used in data mining [32]. The database's strong relationships are analyzed using a few metrics. Minimum confidence and minimum support are the most widely used measurement techniques. The Apriori algorithm mines association rules between data elements in the database using a guided way.

4.2 Classification Algorithm

One of the most widely researched machine learning domains is classification. In order to forecast the class of the desired attribute that the user has provided, the classification principle relies on the anticipated attribute. The two main problems in genetics are sequence annotation and genome categorization. Fuzzy sets, neural networks and genetic algorithms are popular techniques in biological sequence mining. Numerous broad categorization methods are also available, including decision trees, neural networks, naïve Bayesian networks, and rule learning with evolutionary algorithms.

4.3 Clustering Algorithm

In machine learning, the technique of clustering can group related sequences that share certain traits and investigate the useful information of unfamiliar sequences from known structures and functions. As a result, biological sequence grouping is crucial to bioinformatics studies. Clustering differs from classification in that it does not apply a predetermined category. Every cluster has unique traits in common. Cluster analysis's goal is to group data that shares similar features into a single category, after which the data can be analyzed using further techniques.

The clustering method has gained popularity as a machine learning study topic in the past few years due to the advancement of artificial intelligence. Scholars both domestically and internationally have studied clustering algorithms in greater detail in order to increase the processing capability of large-scale data. Numerous outstanding clustering techniques have been developed, primarily based on granularity, uncertainty, entropy, clustering integration, and other factors. There are numerous algorithms in addition to the ones listed above [22]. Every algorithm has unique qualities, and no algorithm can be used in every circumstance. Knowing each algorithm's benefits and drawbacks could improve our application and analysis.

5. Applications of alignment free approaches in next generation

The amount of data generated by the sequenced specimens has already beyond the storage and computing capabilities of advanced systems. Because of the computationally costly multiple alignment procedure, the volume of data created by next-generation genomics is rapidly surpassing analytics capabilities [18, 19]. Alignment-free techniques often give a significant speed boost over traditional next-generation sequencing applications, but they offer way to extract biological relevant information directly from raw next generation sequence information.

Alignment-free techniques for transcripts measurement, for example, reveal utmost of the data offered by aligners is not required for higher quality transcript level assessment. Such techniques use a standard set of transcripts to create an indexing of k-mers, which are then used to estimate expressions by directly matching them to each sequencing read [20, 21]. The association among read and a collection of matching transcripts is described as "pseudoalignment." When pseudoalignments from the same set of transcripts are grouped together, the output of each transcript models can be inferred directly.

Screening of genome variability's, such as single nucleotide/variant polymorphisms, seems to be another essential feature of next-generation sequencing methods. Genotype calling on mapped reads is commonly used to detect these genomic changes. Yet, alignment free methods based on k-mer evaluation allow genotyping of familiar variants straightly from next-gen sequencing information [23, 24]. These approaches appear to be particularly applicable for clinical uses, where sequencing information from a huge count of individuals must be analyzed in a timely way, because they are 1–2 substantially faster than standard mapping-based identification.

Synthesis of recently genomic sequences is among the most difficult jobs in modern biology. It involves an error correction phase and the development of a genomic scaffold depending on read similarity in standard applications (sequence overlaps) [27, 28, 29]. Numerous alignment-free techniques for correcting sequencing reads have been developed, with the goal of being both quick and memory effective as well as extremely accurate.

When the first alignment free metric was introduced absolutely 30 years ago, the efficiency of alignment free approaches has greatly enhanced. The problem nowadays isn't a shortage of alignment-free techniques, but rather the count of benchmark techniques to alignment free sequence analysis every time an advance technique is reported, an advance assessment method and specified dataset are released as well [30, 31]. The bulk of methods, for example, have been tested on simulated DNA sequences, primate/mammalian mitochondrial genomes, complete prokaryotic genomes/proteomes, chosen plant genomes, limited subsets of homologous genes, and possible combinations of these.

6. Conclusion

The computing problems of sequence analysis will be even more significant as sequencing approach gets less costly and more widely available. This problem forces developers to shift their attention to speedier, alignment-independent alternatives. Will conventional alignments be doomed as a result of these new methods? Probably not throughout the authors' lifetimes. Many aspects of modern biology, including the author's note of conserved protein domains, rebuilding of ancestral Nucleotide sequence, evaluating the cost of study sample, as well as homology-based modelling of 3-dimensional protein molecules, still rely on alignment.

Alignment-free methods are quickly expanding their applicability including addressing formerly unanswered problems in phylogenomics as well as horizontal gene transfer, regulating sequence evolution, as well as genome-epigenome linkages. The alignment-free solutions appear to be particularly effective in addressing the limitations of next generation sequencing data processing. The currently prevalent k mer techniques are tied to new measurements for biomedical application.

7. Future scope

The most popular technique to deal with data is machine learning, which is the foundation of data mining. The ability of machine learning technologies to filter vast amounts of data in order to investigate patterns that could otherwise go unnoticed is one of their main advantages. Machine learning is essential for identifying recurring trends in biological processes in the era of massive data of biomedical studies. Machine learning relies heavily on large data sets. Currently, the majority of biological data sets are still too tiny to satisfy machine learning algorithms. Even if the overall amount of biological data is enormous and growing daily, the data is gathered from a variety of platforms. Integrating various data sources is extremely challenging because of the gaps between biology and technology. Machine learning algorithms based on one data set may not be successfully adapted to different data sets since the variations in biological data directly. The machine learning algorithm's analytical outputs tend to be inaccurate if the new data differs greatly from the training data. Biological applications have additional difficulties due to the black-box nature of machine learning models. The applicability of a model is often limited by the difficulty of interpreting its results from a biological perspective.

References

[1]. Berlin K, Koren S, Chin C-S, Drake JP, Landolin JM, Phillippy AM. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. Nat Biotechnol. 2015; 33:623–30.

[2]. Li H. Minimap and miniasm: Fast mapping and de novo assembly for noisy long sequences. Bioinformatics. 2016; 32:2103–10.

[3]. Ren J, Song K, Deng M, Reinert G, Cannon CH, Sun F. Inference of Markovian properties of molecular sequences from NGS data and applications to comparative genomics. Bioinformatics. 2016; 32:993–1000.

[4]. Gardner SN, Slezak T, Hall BG. kSNP3.0: SNP detection and phylogenetic analysis of genomes without genome alignment or reference genome. Bioinformatics. 2015; 31:2877–8.

[5]. Ounit R, Lonardi S. Higher classification sensitivity of short metagenomic reads with CLARK-S. Bioinformatics. 2016; 32:3823–5.

[6]. Roosaare M, Vaher M, Kaplinski L, Möls M, Andreson R, Lepamets M, et al. Strain Seeker: fast identification of bacterial strains from raw sequencing reads using user-provided guide trees. PeerJ. 2017; 5:e3353.

[7]. Everaert C, Luypaert M, Maag JLV, Cheng QX, Dinger ME, Hellemans J, et al. Benchmarking of RNA-sequencing analysis workflows using whole transcriptome RT-qPCR expression data. Sci Rep. 2017; 7:1559.

[8]. Lindgreen S, Adair KL, Gardner PP. An evaluation of the accuracy and speed of metagenome analysis tools. Sci Rep. 2016; 6:19233.

[9]. Pizzi C. MissMax: alignment-free sequence comparison with mismatches through filtering and heuristics. Algorithms Mol Biol. 2016; 11:6.

[10]. alfpy. https://github.com/aziele/alfpy. Accessed 23 Aug 2022.

[11]. Alfree: Benchmark. http://www.combio.pl/alfree/benchmark. Accessed 23 Aug 2021.

[12]. Bernard G, Ragan MA, Chan CX. Recapitulating phylogenies using k-mers: from trees to networks. F1000Research. 2016; 5:2789.

[13]. Pinello L, Lo Bosco G, Yuan G-C. Applications of alignment-free methods in epigenomics. Brief Bioinform. 2019; 15:419–30.

[14]. Comin M, Leoni A, Schimd M. Clustering of reads with alignment-free measures and quality values. Algorithms Mol Biol. 2015; 10:4.

[15]. Yin C, Yau SS-T. A coevolution analysis for identifying protein-protein interactions by Fourier transform. PLoS One. 2017; 12, e0174862.

[16]. Lu YY, Tang K, Ren J, Fuhrman JA, Waterman MS, Sun F. CAFE: accelerated Alignment-Free sequence analysis. Nucleic Acids Res. 2017; 45:2015–7.

[17]. Kuksa P, Pavlovic V. Efficient alignment-free DNA barcode analytics. BMC Bioinformatics. 2020; 10:S9.

[18]. Wei D, Jiang Q, Wei Y, Wang S. A novel hierarchical clustering algorithm for gene sequences. BMC Bioinformatics. 2018; 13:174.

[19]. Hatje K, Kollmar M. A phylogenetic analysis of the brassicales clade based on an alignment-free sequence comparison method. Front Plant Sci. 2019; 3:192.

[20]. Joseph J, Sasikumar R. Chaos game representation for comparison of whole genomes. BMC Bioinformatics. 2021; 7:243.

[21]. Ames SK, Hysom DA, Gardner SN, Lloyd GS, Gokhale MB, Allen JE. Scalable metagenomic taxonomy classification using a reference genome database. Bioinformatics. 2013; 29:2253–60.

[22]. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 2014; 15: R46.

[23]. Roosaare M, Vaher M, Kaplinski L, Möls M, Andreson R, Lepamets M, et al. StrainSeeker: fast identification of bacterial strains from raw sequencing reads using user-provided guide trees. PeerJ. 2017; 5: e3353

[24]. Everaert C, Luypaert M, Maag JLV, Cheng QX, Dinger ME, Hellemans J, et al. Benchmarking of RNA-sequencing analysis workflows using wholetranscriptome RT-qPCR expression data. Sci Rep. 2017; 7: 1559.

[25]. Lim EC, Müller J, Hagmann J, Henz SR, Kim ST, Weigel D. Trowel: A fast and accurate error correction module for Illumina sequencing reads. Bioinformatics. 2014; 30: 3264–5.

[26]. Suwa M. Bioinformatics tools for predicting GPCR gene functions. In: Filizola M, editor. G protein-coupled receptors – modeling and simulation. Springer: Netherlands; 2014. p. 205–24.

[27]. Giancarlo R, Rombo SE, Utro F. Compressive biological sequence analysis and archival in the era of high-throughput sequencing technologies. Brief Bioinform. 2014; 15: 390–406.

[28]. Wang Y, Liu L, Chen L, Chen T, Sun F. Comparison of metatranscriptomic samples based on k-tuple frequencies. PLoS One. 2014; 9, e84348.

[29]. La Rosa M, Fiannaca A, Rizzo R, Urso A. Alignment-free analysis of barcode sequences by means of compression-based methods. BMC Bioinformatics. 2013; 14: S4.

[30]. Comin M, Verzotto D. Alignment-free phylogeny of whole genomes using underlying subwords. Algorithms Mol Biol. 2012; 7: 34.

[31]. Fox NK, Brenner SE, Chandonia J-M. SCOPe: Structural Classification of Proteins–extended, integrating SCOP and ASTRAL data and classification of new structures. Nucleic Acids Res. 2014; 42: D304–9.

[32]. Delibas, E., and Arslan, A. (2020). DNA sequence similarity analysis using image texture analysis based on first-order statistics. *J. Mol. Graph. Model.* 99:107603.