ISSN: 2664-9640

A Review of K-mer Frequency-Based Statistics in Biological Sequence Data Processing and Presentation

Zhiyue Su

University College London, London, UK

zhiyue.su.20@ucl.ac.uk

Abstract

K-mer frequency statistics can be used to reveal subsequence distribution patterns in biological sequences. It is a crucial instrument for measuring sequence similarity and thus has important and extensive applications in numerous biological problems, such as haplotyping, motif discovery, species recognition, metagenomic classification, sequence assembly, multiple sequence alignment, variation detection, and sequence error correction. As a biostatistical technique, k-mer frequency statistics are steadily gaining importance in the gathering and analysis of biological sequencing data. In this research, it provides the data compression techniques, algorithm improvement schemes, and data visualization methods based on k-mer frequency statistics to help the academic community comprehend the evolution of k-mer frequency statistics.

Keywords

K-mer Frequency, Biological Sequence .

1. Introduction

Our understanding of biological sequences (primarily RNA, DNA, and protein sequences) has become increasingly sophisticated as science and technology have progressed. There is no question that this view is supported by vast quantities of data. The quantity of biogenetic sequence data being generated is increasing exponentially. Currently, storage is expanding beyond terabytes to petabytes. The academic community must address the difficulty of extracting relevant information from such voluminous data. Therefore, the scientific community must comprehend how to compress, compute, and express biological sequence data. In this regard, k-mer frequency statistics as a biostatistical approach is playing an ever-increasing role in the collecting and analysis of biological sequencing data [1]. A k-mer is a short DNA sequence segment comprised of k consecutive bases. When k is set to an appropriate value, the k-mer frequency distribution of a DNA sequence contains all of the genome's information, thereby representing the sequence in an equivalent representation. Therefore, the examination of DNA sequences by k-mer distribution in DNA sequences can reveal the properties, functions, and structures of the base distribution in biological sequences. We provide the following k-mer definition: If s is an m-length biological sequence, then $s=q_1q_2\cdots q_m$, where $q_1 \in \sum$ ($\sum$ is the symbol space of a biological sequence, for instance, it has $\sum=\{A, T, G, C\}$ for a gene sequence.) A sequence of consecutive symbols starting from any position in a substring of k-length is referred to as a k-mer. We provide the following definition for the issue of frequency statistics of k-mer in a variety of lengths at various offsets for multiple sequences in alignment mode: Let Ω be the sequence set formed by n gene sequences with m-length in alignment mode, then $\Omega=\{S, S_2\ldots, S_n\}$. The starting positions of S_i and S_j must be in alignment for any $1\leq i, j\leq n$. For a given range of k-mer length variations k_1 and k_2 ($k_1\leq k_2$), the frequency of occurrence of different k-mer substrings in a sequence block consisting of k-mer of length k taken from the offset for each

ISSN: 2664-9640

of the n sequences is calculated for each k-mer length k($k_1 \le k \le k_2$) at each offset position I($0 \le I \le$ m-k), respectively.

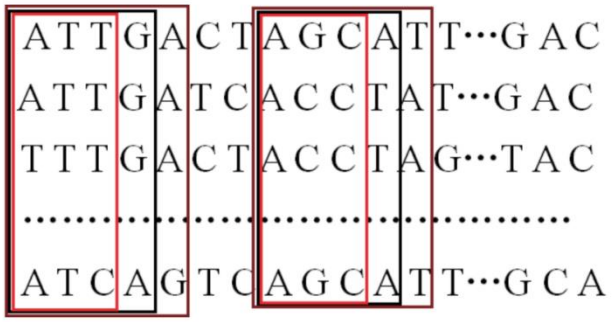

Figure 1 k-mer Sample Diagram

In Figure 1, a clear illustration is provided. All sequences' initial positions are aligned, and the k-mer lengths range from 3 to 5. First, when the offset is 0, the number of instances of each k-mer in the red sequence block of length three is counted. Then, up until the maximum value of 5, we count the occurrences of each k-mer in the 4-length black block. The offset position is then moved one bit to the right, to 2, and we then count the frequency of each substring with each iteration of the k-mer length. The process described above is then repeated until the sequence's conclusion [2]. k-mer frequency statistics can be used to reveal subsequence distribution patterns in biological sequences. It is a crucial instrument for measuring sequence similarity and thus has important and extensive applications in numerous biological problems, such as haplotyping, motif discovery, species recognition, metagenomic classification, sequence assembly, multiple sequence alignment, variation detection, and sequence error correction. Whether it be in image visualization or a new algorithm to deal with k-mer frequency statistics, many researchers have been working on improving the k-mer frequency statistics for decades. In this paper, it will examine recent developments in the processing and presentation of biological sequence data using k-mer frequency statistics. The contributions of k-mer to biostatistics are also outlined to aid the academic community in better understanding how k-mer frequency statistics were developed.

Table1 Review Structure

Processing and presentation of biological sequence data based on k-mer frequency statistics	Groups
DNA data compression	DNA data compression algorithm based on segmented coding
	SBGSO, a hybrid GA-PSO optimization algorithm based on SVM touch-type guided grouping
Algorithm improvement	FTKC algorithm
	BTKC algorithm
Visualization of biological data	Biological diagrams and grayscale
	Models and morphology of genomic k-mer frequency spectrum (mammalian and viral)

ISSN: 2664-9640

2. k-mer Frequency Statistics-Based DNA Data Compression

DNA data compression algorithm based on segmented coding

The academic community has proposed a DNA data compression algorithm based on segmentation coding in response to the feature that the k-mer distribution of various fragment regions in DNA sequences has great variability. The DNA sequence is divided into short sequence fragments of 64 bases during the preprocessing stage, and each fragment is assessed separately and independently. The DNA sequence is compressed by alternative coding based on information such as the total number of occurrences of this k-mer subsequence and its displacement in the fragment after the k-mer subsequence with the highest repetition rate in the fragment is counted.

GA-PSO hybrid optimization algorithm SBGSO based on SVM shape guidance grouping

GA and PSO are both global parallel optimization affine algorithms and are two common artificial intelligence techniques. Using stochastic optimization methods, they get optimal solutions in the global space by updating the population and historical search for optimal locations. GA employs the Darwinian concept of survival of the fittest to eliminate maladaptive elements in the solution, while PSO is an evolutionary computation technique based on swarm intelligence. Although the mutation operator has a certain local search ability, crossover and mutation operation have randomness. When particles are close to the optimal solution, convergence speed is sluggish and local searchability is not optimal. The PSO method, in which individuals are mostly updated by their internal velocity, is user-friendly and has fewer parameters. There are disadvantages of premature convergence and poor convergence [3], especially in the optimization of high-dimensional functions. Combining GA and PSO can effectively leverage the respective benefits of both methods to boost optimization performance.

The academic community has created the hybrid GA-PSO optimization technique SBGSO (SVM Based GSO) based on SVM (Support Vector Machine) model-guided grouping and has utilized it to determine the optimal k-mer combination in DNA sequences for DNA data encoding compression. SBGSO optimization method preserves the autonomy of the GA algorithm while maintaining the independence of the GA and PSO algorithms. The SBGSO optimization algorithm utilizes a support vector machine model to divide the DNA base particle swarm into two groups before each round of the optimization search iteration and optimizes the two groups independently with the GA and PSO algorithms. Then, the two groups were mixed to form a new population, then grouped again for optimization, and then traversed successively until the termination condition was satisfied, so as to achieve the purpose of hybrid optimization [4].

3. Algorithm Improvement based on k-mer Frequency Statistics: from FTKC Algorithm to BTKC Algorithm

The conventional algorithm in use today is the forward traversal counting algorithm or FTKC for short. In the FTKC algorithm, for each offset position l, traverse in the range of k-mer length variation from the minimum value of 1 k to the maximum value of 2k. Each time a traversal is performed, all n sequences are scanned. The substrings of the current k-mer length at the current offset of each sequence are retrieved, and the frequency of each substring is recorded using hash tables. The operation of the FTKC algorithm demonstrates that in a sequence space containing n sequences of length m, for each k-mer of length k ($k_1 \le k \le k_2$), m-k+1 blocks sliding from left to right are formed. To calculate k-mer frequency statistics for each block, it must iteratively visit n sequences and perform n hash operations, making the algorithm's time complexity $O((k_2-k_1)\ mn)$. The BTKC algorithm first traverses all n sequences for each offset. The substring whose length at the current offset is K_2 (maximum k-mer length) is used as the

ISSN: 2664-9640

key, while the substring's occurrence frequency is used as the value. The result of using a hash table for statistics and storage is the letter H. Next, we create an empty H_{k2-1} hash table. Then, I iterate over H_{k2}, which is obtained in the preceding step. As the new key for each record is traversed, the prefix substring with the key length K_{2-1} is selected. When the k-mer length at the offset is k-1, the frequency statistical result is obtained by adding the value of the record to the value corresponding to the key in H_{K2-1}. Repeat the process. By traversing the results obtained when the k-mer length is k+1, the statistical results when the k-mer length is K are obtained until the result H_{k1} is obtained when the k-mer length at the offset is the smallest k_1. The process is then repeated with the subsequent offset. Slide to the extreme right offset and obtain the final statistics [5].

The related study demonstrates that the execution time of both algorithms grows linearly with the number of sequence entries. This observation is consistent with the algorithm complexity analysis results, namely that the time complexity of both algorithms is proportional to the number of sequence entries. The researchers also discovered that the execution time of both algorithms increases linearly as the sequence length increases. Nevertheless, the running time of the FTKC algorithm increases linearly with the k-mer length variation range, whereas the running time of the BTKC algorithm increases very slowly with the k-mer length variation range without a significant trend. This result indicates that, compared to the FTKC algorithm, the time complexity of the researcher's proposed BTKC algorithm is not significantly correlated with the range of k-mer length variation. The time performance of the BTKC algorithm is nearly k_2-k_1 times that of the FTKC algorithm, especially as the maximum length of k-mer increases and the BTKC algorithm's performance advantage becomes more apparent [6].

4. Biological data visualization based on k-mer frequency statistics

Visualization can facilitate intuitive comprehension of complex phenomena and large-scale data, and it plays a crucial role in the study of biological sequences. Using visualization techniques to convert biological sequences into a variety of two-dimensional or multi-dimensional image images, followed by signal and image processing techniques, provides us with new ways to investigate the properties and functions of biological sequences. The academic community initially proposed a visualization scheme for constructing biological sequence barcodes using k-mer frequencies of biological sequence data. The scheme is predicated on the premise that the combined frequency distribution of k-mers and their backward complementary k-mers for each gene sequence is stable. This scheme permits the efficient resolution of two difficult problems: macro-genome classification and identification of horizontal transfer genes. Each gene sequence is divided into multiple segments of length M, which do not overlap. Then, every k-mer within each fragment is extracted. For each k-mer, the frequency of that k-mer in combination with its inverse complementary k-mer in that fragment is separately tallied. Finally, each gene sequence's barcode is a matrix with L/M rows and N(k) columns. Each element in the matrix represents the frequency of the corresponding k-mer in the corresponding sequence fragment, where N(k) is the number of unique combinations of k-mer species. In the subsequent step of data representation, each barcode is mapped to a corresponding grayscale map.

Specifically, the target value is minimized by counting the frequency of each k-mer in all gene sequences, sorting and listing them in ascending order of frequency, and then minimizing the count. Lastly, to assign the corresponding grayscale value, if the k-mer frequency falls within any sublist, the value is mapped to the corresponding grayscale value in that sublist, and the corresponding grayscale map is obtained by performing the aforementioned mapping for each element in the QR code [7]. This paper must also discuss the academic studies conducted on the model and morphology of the genomic k-mer frequency spectrum. The results of the study

ISSN: 2664-9640

indicate that the k-mer frequency spectrum of the majority of species is single-peaked. However, a few species, including all mammals, have a k-mer frequency spectrum with multiple peaks. Other researchers have employed genetic barcode visualization techniques for the identification and classification of enteroviruses. The two-dimensional codes of each genome were constructed by counting the respective k-mer frequencies of enterovirus and other viral genomes. Next, a phylogenetic tree was constructed using the two-dimensional code distance and neighbor-joining method to identify and classify enteroviruses. The experimental results demonstrated that the identification and classification method based on genetic barcodes is more accurate than the comparison-based method [8].

5. Conclusion and Future Direction

From the standpoint of biological sequence data compression, both methods reviewed in this article begin with precise and highly repetitive k-mer short sequence segments in DNA. The algorithms are capable of achieving relatively good overall compression with good robustness, but they all suffer from poor time efficiency. Since the segments are of fixed length and short, the DNA data compression algorithm based on segment encoding has a strong relationship between the time required for encoding and the sequence length. The algorithm time increases proportionally with the length of the sequence. The following work will address the issue of balancing the length of segmentation and the number of binary bits used to encode each item in order to improve time efficiency without sacrificing compression rate. Due to the algorithm's complexity, the hybrid GA-PSO optimization-based DNA data compression algorithm has a relatively large time overhead. Future research will focus on optimizing the algorithm to reduce the time overhead while increasing the flexibility of the number of k-mer subsequence selection and k-mer length selection in the particles [9]. From the standpoint of algorithmic improvement of biological data, it proposes and implements a backward traversal k-mer counting algorithm, BTKC, as opposed to the conventional forward traversal-based FTKC. The BTKC algorithm first traverses all sequences at each offset position to determine the frequency statistics of the longest k-mer. The k-mer statistics of other lengths are used to calculate the k-mer statistics of previously obtained lengths. This makes it unnecessary for the FTKC algorithm to traverse all sequences for each length of k-mer information in order to obtain the length k-mer statistics. Analysis of the algorithm's time complexity and experimental results demonstrate that the performance of the BTKC algorithm is significantly superior to that of the FTKC algorithm. In addition, its time complexity is independent of the variation range of k-mer length, making it highly applicable in the case of a large variation range. From the perspective of biological data visualization, whether for the full sequence or inter-sequence k-mer frequency counting problem, the use of the hash-based algorithm to determine whether the corresponding k-mer exists and the frequency it has appeared will result in a large number of duplicate characters appearing, causing a large waste of space (e.g., for k-mer ATATTA and ATATTC, these two k-mers have a common prefix ATATT, but the hash-based method needs to store this common prefix twice). For multiple k-mer strings that share a common prefix, we only need to store one copy of the common prefix if we choose to store the occurring k-mer strings in a suffix tree. This strategy, when compared to the hash method, can significantly reduce the amount of storage space needed to store all the k-mer, but it also adds complexity to the implementation and lengthens the time needed to access the k-mer. As a result, it must strike a fair balance between the demands on time and space for a given problem's analysis [10].

ISSN: 2664-9640

References

[1] Yao Y T , Wu Y W , Lin P T . A Two-Stage Multi-Fidelity Design Optimization for K-mer-Based Pattern Recognition (KPR) in Image Processing[C]// ASME 2020 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference. 2020.

[2] Lee H J , Shuaibi A , Bell J M , et al. Unique k -mer sequences for validating cancer-related substitution, insertion and deletion mutations[J]. NAR Cancer, 2020, 2(4).

[3] Astakhov S , Astakhov O , Fadeeva N , et al. A ring generator of two- and three-frequency quasiperiodic self-oscillations based on the van der Pol oscillator.[J]. Chaos (Woodbury, N.Y.), 2021, 31(8):083108.