A New Method of Gene Information Vectorization and its Application in Similarity Search

Qian Dong¹, Zhaogong Zhang², Dayuan Zheng², Zhiliang Zhang²
Heilongjiang University
*Corresponding author’s e-mail: zhaogong.zhang@qq.com

Abstract. With the development of the Human Genome Project, more and more biological sequence data are generated, and the analysis and processing of these sequence data have promoted the development of bioinformatics. Sequence similarity analysis is the basis of bioinformatics, through which we can use the information of known sequences to study the structure, function and evolutionary relationship of unknown new sequences. This paper performs data compression and retrieval on the genome database based on the dbSNP information of DNA. According to the rule of determining a protein by three bases, the amino acid characters are determined, and the redundant information is removed by using the dbSNP information. It is the first time to propose the construction of a new compressed form of biological sequence structure, which can reflect the strong correlation between the SNP location information and SNP in each sample in the genome. Finally, this paper constructs a complete biological sequence approximate neighbor query system, which can not only greatly reduce the storage and computing overhead, but also improve the query efficiency under the condition of ensuring the retrieval accuracy. The accuracy and scalability of this method are verified by experiments on a large data set of gene database.

1. Introduction
With the launch of the international human genome project, the biological sequencing project to obtain the basic data of DNA sequence has attracted wide attention in recent years. With the development of various sequencing projects, the amount of sequence data generated grows exponentially [1]. The biological sequence database contains a large amount of information and knowledge. How to mine and utilize this information and knowledge is worth studying. The similarity search of biological sequence database is the most basic problem in biological information processing. By comparing the similarity degree of the sequence to be queried and the target sequence of the database, the homology can be inferred to meet the research and application requirements of biological researchers.

This paper proposes for the first time to construct a new compression form for biological sequences, which has a compression rate of 66% and does not require complicated decompression algorithm. The compression is bidirectional mapping. Based on the Spark distributed computing platform, a distributed biological sequence processing method is designed and implemented, to realize rapid and accurate processing of massive data. And this system using dbSNP with the Spark frame to remove redundant information. An embedding method to maintain measurement is used to represent SNP information with vectors, and correspondingly convert SNP sequences into vector domain sequences. By using local sensitive hash algorithm to analyze and retrieve biological sequences, similar sequences are found to find similar individuals.
2. Related Work
DNA sequence data contains certain biological characteristics and similar characteristics. Homology search for these genomic sequence databases is a key step in the inter-protein function and evolutionary relationship reasoning.

The traditional method of analyzing the similarity of DNA sequences is mainly based on the similarity of strings and does not fully utilize the characteristics of DNA data, which is not only computationally intensive but also has certain defects. For example, the Fourier transform method [2] will lose part of the information in the sequence, and it cannot clearly show the similarities and differences between sequences.

With the continuous growth of the number of biological sequences, K-center clustering algorithm represented by Trapnell [3] et al. This algorithm works well on small data sets. However, clustering by pairwise alignment has a large amount of computation and time consuming when clustering large amounts of data. Solovyov [4] et al. proposed a high-flux center clustering algorithm Acluster based on word frequency. The algorithm used K-words to count the frequency of the biological sequence and k-means clustering to determine the center based on word frequency information.

However, in the case of large data volume, with the increase of the number of classes, the efficiency of the algorithm decreases and the convergence of clustering iteration is slow. Jeremy Buhler [5] introduced the LSH-All-PAIRS algorithm on the algorithm of locally sensitive hashing, which uses a random search technique (locally sensitive hash) to find local alignments with no gaps in the genome sequence. Li W, Ren J [6] et al. proposed a new algorithm based on data index structure for biological sequence pattern mining algorithm MPBSMI. Recent ones include trigonometric function method, spectral dynamic method, etc. [7-9].

Generally, "biological sequence similarity search" refers to the problem of retrieving from the database all genes sharing the perceptual characteristics with the query sequence q. However, since the complexity of such methods grows linearly with the size of the database, this problem is difficult to solve when considering the size of the current database. In the face of tremendous growth in database size, we must consider one problem, a genetic structure design that allows us to quickly identify similar genes based on SNP functional information.

In view of the above problems, this paper proposes a novel gene data structure, which uses SNP site information to determine the position of each 10 bases in the upstream and downstream of SNP, and determines 7 amino acid characters according to the rule of determining an amino acid by three bases, so as to compress the data of the whole DNA sequence with the idea of local replacing the whole. An embedding method of retention metric is used to represent the sequences in the vector domain. By using the locally sensitive hash algorithm to analyze and retrieve the biological sequences, the similar sequences are found, and the similar individuals are found.

3. Theoretical Basis
3.1. Sequence Similarity Comparison
Sequence similarity alignment means that two or more sequences are compared in base arrangement to reflect the similarity between the fragments and to clarify the homology of the sequences. Here, the sequence analysis is determined by comparing the sequence of the unknown function with the known sequence.

3.2. Word Embedding
Word embedding is the process of converting words in text into vectors. The word embedding method defines a mapping function that maps words to a vector. After the word embedding mapping, the research on text data can be processed by more mathematical tools.
The word vector is classified according to the vector produced by the word embedding, and the word vector can be divided into a one-hot representation and a distributed representation. Bengio Y, Ducharme R et al. [11] introduced neural networks into the field of word embedding; Mikolov T, Chen K et al. [12] proposed the CBOW and Skip-Gram models, and the performance of the model has been greatly improved.

3.3. Local Sensitive Hash
The Local Sensitive Hash (LSH) [10] theory is used to find approximate similarities in massive high-dimensional data. The basic idea of the LSH algorithm is that two adjacent or similar data sample points in the original space perform the same mapping or projection transformation, the probability that the obtained data points are still adjacent or similar in the new space is very large, and the probability that the non-adjacent data points are mapped to the new space is small.

3.4. Biological Sequences Based On dbSNP
We propose a sequence compression representation based on SNP sites, which performs cutting and conversion at SNP sites. The upstream and downstream of the SNP have 10 base positions respectively, that's 21 base positions. According to the rule of determining an amino acid by three bases, seven amino acid characters (7 characters) are determined. The optimization training is carried out based on the skip model and the continuous word bag model, each sequence consists of several strings of seven characters.

Taking a fragment of the chr22 chromosome of hg19 as an example, a biological sequence containing information of a SNP site is known. As shown in Table 1:

| markname          | SNP      | chrom | position  |
|-------------------|----------|-------|-----------|
| rs7291810XXX      | rs7291810| 22    | 1443520   |
| rs10154759XX      | rs10154759| 22    | 14441342  |
| rs6423472XXX      | rs6423472| 22    | 14467621  |

Table 1 shows the information of each dbSNP on chromosome 22 of hg19, including the name of SNP and the chromosome to which the SNP belongs, as well as the position on this chromosome.

The hg19's chr22 chromosome chr22.fa, and the file fragment is as follows:

```
TACCTGTCTTGCTCCTGCGCTGCAATGCATGATCTTCTCCC
TTCTGAGGCTTTGCTGATGCTTATCTCCACCTGGAACACTCATTTATTCATTATATGCTCA
```

Based on the SNP position information of chr22 in Table 1, we locate the chr22 sequence, find the current position, intercept the position of 10 bases in the upstream and downstream, and convert it into a new biological sequence through standard genetic code table. The sequence conversion process diagram is shown in Figure 1:

![Sequence conversion diagram.](image)

4. Experiment And Result Analysis

4.1. Data Sources
The experiment used two data sets, the self-generated simulation data set and the public GAW20 [13] data set, which were analyzed on the chr22 chromosome of the published human genome hg19. In this experiment, the chromosome chr22 and SNP information marked on this chromosome were selected...
from publicly available data set of GAW20, which contained 822 records in total, and each record contained 10,000 dbSNP information. All data is stored in a file with the extension called ".csv" and each field is separated by ",". The first line of each file contains all the fields in the file. GWASUBJ is used to uniquely identify each sample.

4.2. Experimental Procedure

The experimental steps for performing a similarity search on a sequence are as follows:

1. For the multiple samples in the GAW20 simulation data set, SNP-based sequence compression is performed. Spark distributed computing framework is used for parallel processing of massive data.

2. The processed samples constitute a plurality of new biological sequences. Based on the optimization training of the new sequences based on the skip model and the continuous word bag model, a metric preserving embedding method is used to represent the sequences in the vector domain to generate word vectors.

3. Based on the local sensitive hash, the generated word vector is hashed to find the similar candidate set, and the distance calculation is performed to find the similar top k (k can be arbitrarily specified) samples.

4.3. Experimental Results And Analysis

In this experiment, the chr22 chromosome of hg19 was selected, which is about 0.6 billion bp in length and contains about 10,000 SNP site information. The 2000, 4000, 6000, 8000 and 10000 SNP site information were tested to calculate the execution time and accuracy of the results. The accuracy and execution time of the calculated results were different with 800,1600,2400,3200,4000 samples. The experimental results are shown in Figure 2 and Figure 3.

![Figure 2](image1.png)

**Figure 2.** Effect of different SNP site on accuracy and execution time.

![Figure 3](image2.png)

**Figure 3.** Effect of different number of samples on accuracy and execution time.

Figure 2 show that with the increase of the number of SNP bits, the accuracy of obtaining similar samples increases gradually, and the execution time also increases gradually.

Figure 3 show that as the number of samples increases, the accuracy of our similar samples is gradually reduced, and the execution time is gradually increased.
Table 2 shows the experimental results we obtained on the GAW20 data set, with blue marking as test sample id and yellow marking as the correct id we obtained similar to the test sample.

5. Conclusion
The biological sequence mode based on dbSNP constructed in this paper compresses the biological sequence, based on the effective use of biological genetic information to ensure the integrity of sequence information, this mode compresses biological sequences. The size of biological sequences constructed by this structure is much smaller than the original sequence data set and can be stored in memory for quick query. Experiments show that the biological sequence data constructed based on dbSNP can improve the performance of sequence query, especially when the sequence database is larger, the advantages of the algorithm are more obvious.

Further work includes improving the speed of sequence conversion vectorization and optimizing on local sensitive hashing algorithms and the future research will focus on the paralleliization algorithm based on this single gene vectorization method, as well as the individual genome-oriented information vectorization method based on this single gene vectorization method, and apply it to the discovery of pathogenic genes.

References
[1] Cochrane G, Akhtar R, Bonfield J, et al. (2009) Petabyte-scale innovations at the European Nucleotide Archive. [J]. Nucleic Acids Research, 37(1): D19-25.
[2] ZHANG S, DAI X, et al. (2005) Alignment-based biological sequences similarity analysis [J]. China Journal of Bioinformatics, 3(2):81-84.
[3] Trapnell C, Cacchiarelli D, Grimsby J, et al. (2014) The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells [J]. Nature biotechnology, 32(4): 381.
[4] Solovyov A, Lipkin W I. (2013) Centroid based clustering of high throughput sequencing reads based on n-mer counts [J]. Bmc Bioinformatics, 14(1):268.
[5] Buhler, J. (2001) Efficient large-scale sequence comparison by locality-sensitive hashing [J]. Bioinformatics, 17(5):419-428.
[6] Li W, Ren J, Li X. (2018) MpBsmi: A new algorithm for the recognition of continuous biological sequence pattern based on index structure [J]. PLoS ONE, 13(4).
[7] Zhong S, Liu Y, Li R, et al. (2010) A novel non-degenerate 2D graphical representation and numerical characterization of DNA sequences[C]// IEEE Fifth International Conference on Bio-inspired Computing: Theories & Applications. IEEE.

[8] Guo-Sen X, Xiao-Bo J, Chunlei Y, et al. (2018) Correction to: Graphical Representation and Similarity Analysis of DNA Sequences Based on Trigonometric Functions[J]. Acta Biotheoretica.

[9] Ho P H, Nguyen N A T, Vo T H . (2018) DNA Sequences Representation Derived from Discrete Wavelet Transformation for Text Similarity Recognition[J].

[10] Leskovec J, Rajaraman A, Ullman J D. (2014) Mining of massive datasets[M]. Cambridge university press.

[11] Bengio Y, Ducharme R, Vincent P, et al. (2003) A neural probabilistic language model[J]. Journal of machine learning research, 3(Feb): 1137-1155.

[12] Mikolov T, Chen K, Corrado G, et al. (2013) Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv:1301.3781.

[13] GAW20 Homepage, https://www.gaworkshop.org/.