Component Vector method and its application in detecting similarities between sequences

Guang-You DUAN
Key Lab. of Bioactive Materials , Ministry of Education and the College of Life Sciences
Nankai University
Tianjin, PR China, 300071
duanguangyou@163.com

Shan GAO
Key Lab. of Bioactive Materials , Ministry of Education and the College of Life Sciences
Nankai University
Tianjin, PR China, 300071
jacky.gao@eyou.com

Ning ZHANG
Key Lab. of Bioactive Materials , Ministry of Education and the College of Life Sciences
Nankai University
Tianjin, PR China, 300071
zhni@mail.nankai.edu.cn

Zhuo YANG
College of Medicine Science
Nankai University
Tianjin, PR China, 300071
zhuoyang@mail.nankai.edu.cn

Tao ZHANG*
Key Lab. of Bioactive Materials , Ministry of Education and the College of Life Sciences
Nankai University
Tianjin, PR China, 300071
zhangtao@nankai.edu.cn

* Corresponding author: Tao Zhang

***Abstract*—With the increase of sequence information, many methods have been used in detecting sequence similarities. In order to avoid the shortcomings of the traditional method based on sequences alignment such as high time and space complexity, component vector method has been successfully used in the reconstruction of phylogenetic tree using genome (proteome) or conserved molecular sequences. With the aim of extension and study the advantage of the method, it has been employed to detect similarities between common short gene and protein sequences and get some good results. Results show that component distance method successfully detects the similarities in our HIV sequences, and have high correlation with the traditional method based on sequence alignment.**

Keywords- component vector method, phylogenetic tree, alignment

I. INTRODUCTION

Comparison of primary sequence information is rapidly becoming the major source of data in the elucidation of the molecular mechanisms of replication and evolution of all organisms. Many multiple sequence alignment programs and various scoring schemes have been developed to analyze potential relationships among sequences [1]. There are two basic software approaches in determining the similarities among proteins and nucleic acids: Global and Local methods [2]. Global alignment programs attempt to align the sequences over their whole length, whereas local programs search only for the most conserved motifs. The most effective alignment algorithm depends on the nature of the sequences to be aligned. Global algorithms produce the most accurate and reliable alignments involving equidistant sequences, divergent families of sequences and the alignment of orphan sequences with a family [1]. One of the objectives of these methods is to detecting similarities between sequences, but there are many problems when the sequences are not in a family and, in some cases, need fine-tunning and adjustment. It also takes much time. From another perspective, component vector method has been brought up. The detailed method will be described in the part of Method. Comparison of G+C content or amino acid composition has long been a standard practice in analyzing biological sequences. By extending single-nucleotide or single-amino acid counting to longer strings, one increases the "resolution power" of the analysis, takes into account short-term correlations in the sequences, and enhances the species specificity of some sequence features[3]. Dinluceotide relative abundance has been used as a genomic signature. The investigation of oligo-nucleotide correlation in a DNA sequence is an important approach to the understanding of

genetic language [4, 5]. Trifonov et al (1986) [6] compiled a dictionary of oligo-nucleotide words from the statistical analysis of nucleotide frequencies in DNA sequences. In the present study, the definition of sequence distance given in literature, which is based on the calculation of n-tupe frequency difference, was used instead of multialignment.

II. METHOD

A. Frequency or Probability of Appearance of K -Strings

Given a DNA or amino acid sequence of length L, the number of appearances of (overlapping) strings of a fixed length K in the sequence is counted. The counting may be performed for a complete genome or for a collection of translated amino acid sequences. There are a total of N possible types of such strings: $N = 4^K$ for DNA and $N = 20^K$ for protein sequences.

For concreteness consider the case of one DNA sequence of length L. Denote the frequency of appearance of the K-string $\alpha_1\alpha_2\ldots\alpha_K$ by $f(\alpha_1\alpha_2\ldots\alpha_K)$, where each α_i is 1 of the 4 nucleotide single-letter symbols. The frequency, divided by the total number (L - K + 1) of K-strings in the given DNA sequence, may be taken as the probability $p(\alpha_1\alpha_2\ldots\alpha_K)$ of appearance of the string $\alpha_1\alpha_2\ldots\alpha_K$ in the protein:

$$p(\alpha_1\alpha_2\ldots\alpha_k)=\frac{f(\alpha_1\alpha_2\ldots\alpha_k)}{L-K+1} \quad (1)$$

The collection of such frequencies or probabilities reflects both the result of random mutations and selective evolution in terms of K-strings as "building blocks".

B. Substraction of Random Background

Mutations happen in a more or less random manner at the molecular level, while selections shape the direction of evolution. Neutral mutations lead to some randomness in the K-string composition. In order to highlight the selective diversification of sequence composition, one must subtract a random background from the simple counting results.

Suppose we have done direct counting for all strings of length (K - 1) and (K - 2). The probability of appearance of K-strings is predicted by using a Markov model:

$$p^0(\alpha 1\alpha 2\ldots\alpha k)=\frac{p(\alpha_1\alpha_2\ldots\alpha_{k-1})p(\alpha_2\alpha_3\ldots\alpha_k)}{p(\alpha_2\alpha_3\ldots\alpha_{k-1})} \quad (2)$$

The superscript 0 on p^0 indicates the fact that it is a predicted quantity. We note that the denominator comes from the frequency of (K-2)-strings. This kind of Markov model prediction has been used in biological sequence analysis for a long time [7]. It can be justified by virtue of a maximal entropy principle with appropriate constraints [8].

C. Composition Vectors and Distance Matrix

It is the difference between the actual counting result p and the predicted value p^0 that really reflects the shaping role of selective evolution. Therefore, we collect

$$a(\alpha_1\alpha_2\ldots\alpha_K)=\begin{cases}\dfrac{p(\alpha_1\alpha_2\ldots\alpha_K)-p^0(\alpha 1\alpha 2\ldots\alpha k)}{p^0(\alpha_1\alpha_2\ldots\alpha_K)} & p^0\neq 0\\ 0 & p^0=0\end{cases} \quad (3)$$

For all possible strings $\alpha_1\alpha_2\ldots\alpha_K$ as components to form a composition vector for a species. To simplify the notations further, we define a_i to be the *ith* component corresponding to string type *i*, where *i* runs from 1 to $N = 20^K$. Putting these components in a fixed order, a composition vector for sequence A was obtained:

$A=(a_1, a_2, \ldots, a_N)$

Likewise, there was a composition vector for sequence B

$B=(b_1, b_2, \ldots, b_N)$

The correlation C(A, B) between any two species (A and B) is calculated as the cosine function of the angle between the two representative vectors in the N-dimensional space of composition vectors:

$$C(A,B)=\frac{\sum_{i=1}^{N}a_ib_i}{\left(\sum_{i=1}^{N}a_i^2\times\sum_{i=1}^{N}b_i^2\right)^{\frac{1}{2}}} \quad (4)$$

The distance D(A,B) between the two species is defined as

$$D(A,B)=\frac{1-C(A,B)}{2} \quad (5)$$

Since C(A,B) may vary between -1 and 1, the distance is normalized to the interval (0, 1). The detailed method can also be seen in [3].

III. APPLICATION AND RESULT

There are 305 gene sequences, which encode envelope glycoprotein (env) protein in HIV. They are obtained from the Chinese center for disease control and prevention. In order to detect some features in these sequences, the correlation between these sequences was examined. Using the component vector method, we calculated the correlation matrix by setting the k=15 and compared it with the traditional multi-alignment method using clustalX [9]. Since the correlation of the eigenvalue of the two matrixes is 0.9985, these two methods are similar to detect the similarities between the sequences. And then the phylogeny tree using kitsch method in the package of PHYLIP [10] was constructed. Through the rooted evolution tree (Fig 1) it was found that the HIV strain (1544.1174.) played an important in these HIV sequences.

*This work was supported by grants from the NBRPC (2007CB914803) and the NSFC (30870827)

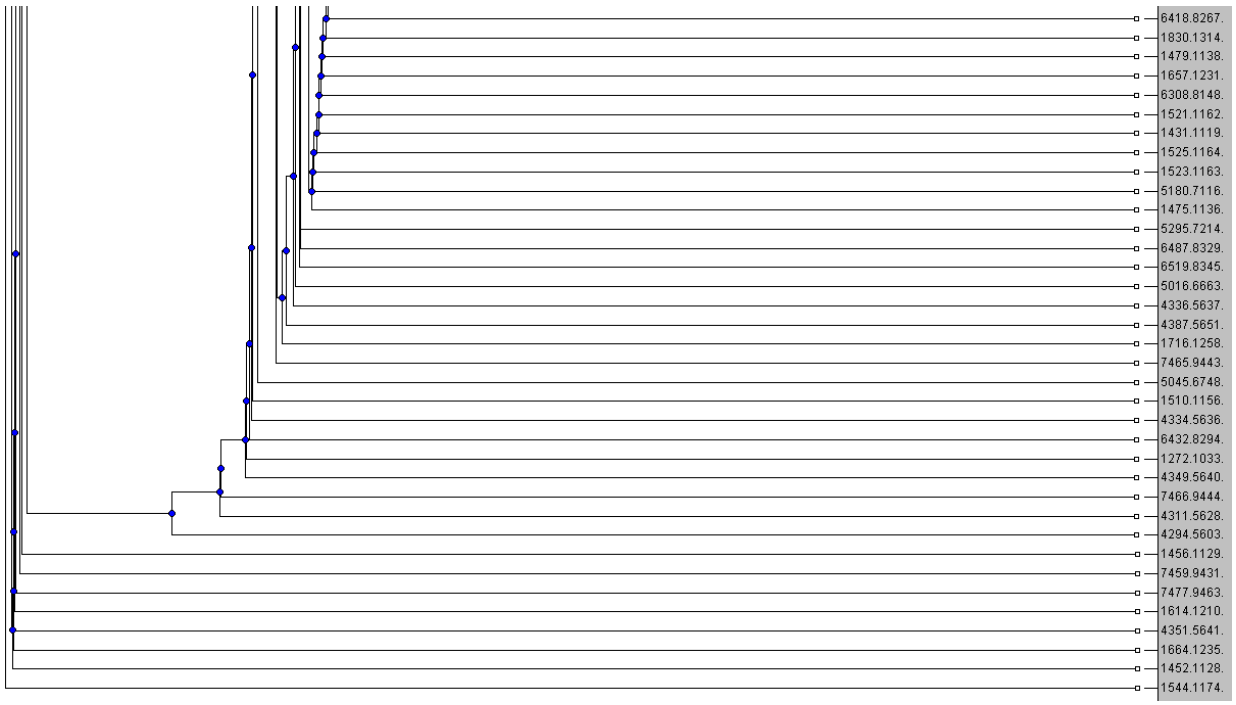

Figure 1. One part of the phylogeny tree of the 305 HIV sequences. It's constructed by the kitsch program in the package of PHYLIP using the distance matrix calculating through the component vector method

IV. DISCUSSION

A. Component vector method can be used in measurement of the shared information between two sequences. It is useful because biological sequences encode information and the occurrence of evolutionary events (such as insertions, deletions, point mutations, rearrangements, and inversions) separating two sequences sharing a common ancestor will result in the loss of their shared information. Regions of sequences， which do not share a common ancestor， will not share more information than would be expected at random.

B. The setting of the K-value in this method is very important. In principle, the bigger the value is, the better the effect. Considering other issues such as the time and the space of the computation, the value cannot be set at a very high level. As long as it can solve the problem, the value can be set at any appropriate number. Some suggestions related to the setting of the K-value could be obtained from the previous work [3].

C. In the study of detecting the similarities between sequences using the component vector method, it was found that statistical correlation was obvious when the k-value became bigger. Whether it can be used to analyze remote homologous sequences is an issue, which is worth considering. The data suggest that the method can successfully solve some questions.

D. The traditional tool used to find the similar sequences is BLAST [11], from the result of our work it may be used as another tool like BLAST.

REFERENCES

[1] Julie D. Thompson, Frederic Plewniak and Oliver Poch. A comprehensive comparison of multiple sequence alignment programs [J]. Nucleic Acids research, 1999, 27(13):2682~2690.

[2] MA McClure, TK Vasi and WM Fitch. Comparative analysis of multiple protein sequence alignment methods [J]. Mol Biol Evol, 1994, 11(5):811.

[3] Qi J, Wang B, Hao B L. Whole genome prokaryote phylogeny without sequence alignment: a K-string composition approach [J]. J. Mol. Evol, 2004, 58: 1~11.

[4] Luo L F, et al. Statistical correlation of nucleotides in a DNA sequence [J]. Phys. Rev. 1998, E58:861~871.

[5] Lobzin V V, et al. Order and correlation in genomic DNA sequences[J]. Physics Uspekhi. 2000, 43:55~78.

[6] Trifonov E N, Brendel. Genomic-a dictinary of genetic code [J]. Balaban, Philadelphia.1986.

[7] Hu R, Wang B. Statistically significant strings are related to regulatory elements in the promoter regions of Saccharomyces cerevisiae [J]. Physica A, 2001, 290(3): 464-474.

[8] Brendel V, Beckmann JS, Trifonov EN. Linguistics of nucleotide sequences: morphology and comparison of vocabularies[J]. J Biomol Struct Dyn. 1986, 4(1): 11–21.

[9] Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG.. Clustal W and Clustal X version 2.0[J]. Bioinformatics, (2007), 23, 2947-2948.

[10] Felsenstein, J. 2004. PHYLIP(Phylogeny Inference Package) version 3.6. Distributed by the author. Department of genome sciences, University of Washington, Seattle.

[11] SF Altschul, TL Madden, AA Schaffer, J Zhang, Z Zhang, W Miller and DJ Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids research, 1997, 25(17):2389~3402.