A burrows-wheeler transform based method for DNA sequence comparison

Chun Li1, 2, *, Huan Liu1, Junhong Liu1, Yuping Qin3, Zhifu Wang1

1College of Mathematics and Physics, Bohai University, Jinzhou 121013 China
2Food Safety Key Lab of Liaoning Province, Bohai University, Jinzhou 121013 China
3College of Engineering, Bohai University, Jinzhou 121013 China

Email address:
lchlmb@163.com (Chun Li)

To cite this article:

Abstract: Burrows-Wheeler Transform (BWT) is an extremely useful tool for textual lossless data compression. Recently, it has found many applications to bioinformatics. In this paper, BWT is introduced from the view of combinatorics, and then an equivalence relation on words is proposed which shows that the transformation captures some common features of equivalent words. Based on the rationale that to what extent two words differ can be evaluated by the factors excluding their common features, a matrix representation for a DNA sequence is defined by means of a “subtraction operation” between the original word and its BWT word, thus a DNA sequence is converted into a 24-D vector whose components are the spectral norms of such matrices. To illustrate the use of the quantitative characterization of DNA sequences, phylogenetic trees of the full β-globin genes of 15 species and the S segments of 13 hantaviruses are constructed. The resulting monophyletic clusters agree well with the established taxonomic groups.

Keywords: BWT, Elementary Row Operations, Equivalence, k-Word, Matrix Representation, Permutation, Phylogenetic Analysis

1. Introduction

Sequence comparison is often viewed as a fundamental precondition for further study, for instance, for the identification and quantification of conserved regions or functional motifs, for profiling of genetic disease, for phylogenetic analysis, and for sequence profiling and prediction. One of the most important problems in this area is how to measure the similarity/dissimilarity between two biological sequences. The first widely accepted solution to this problem is based on sequence alignments, in which a distance function or a score function is used to represent insertion, deletion, and substitution of letters in the compared sequences. However, the computational complexity and the inherent ambiguity of the alignment cost criteria are still the bottleneck problems. In addition, as pointed out by Yang et al. [1], there are many integral properties lost if we use the alignment methods only. Therefore, the emergence of research into alignment-free measure is apparent and necessary to overcome the limitation of alignment-based measure.

The graphical representation of DNA is one kind of alignment-free methods of sequences analysis [2-15], which facilitates visual recognition of differences among related DNA sequences by inspection. Meanwhile, such representations can lead to numerical characterizations of the sequence, which is accomplished by associating with the graphical representation of DNA a corresponding mathematical object such as a matrix, and then using various properties of the mathematical object as sequence descriptors [6-10, 13-17]. Besides these, there are some alignment-free methods based on the rationale that functionally similar sequences must share some common words. Within these methods each sequence is associated with a vector whose components are related to the k-mer. A distance function for these vectors is then defined [18-22]. There are also some other important methods such as Lemple-Ziv (LZ) complexity, Burrows-Wheeler (BW) transform [1, 23-28] which are based on compression algorithm, but do not actually apply the compression. The Burrows-Wheeler Transform (BWT) was introduced by Burrows and Wheeler in 1994, and is recently studied also from a combinatorial point of view [24-26,29-32]. Loosely speaking, BWT can map any finite string (word) over an
ordered alphabet to another one which can be compressed easier. To compare the similarity of two sequences, Mantaci et al. [25,30] introduced an extension of the Burrows-Wheeler Transform (EBWT) and defined a class of dissimilarity measures. While Yang et al. [1] used a Burrows-Wheeler similarity distribution (BWSD) based on Burrows-Wheeler transform to express the similarity between two protein sequences. In this paper, we first introduce the algorithm of BWT from the view of the linear permutation and the circular permutation. Then we construct matrix representations for a DNA sequence by means of a “subtraction” between the sequence and its BWT sequence, on the basis of which we characterize a DNA sequence by a 24-D vector whose entries are the spectral norms of these matrices. The proposed method is tested by phylogenetic analysis on two different data sets: one is composed of full β-globin genes of 15 species, and the other is composed of S segments of 13 hantaviruses. The results show that the approach proposed in this paper is a powerful and useful tool for the comparison of DNA sequences.

2. Burrows-Wheeler Transform

Let \( \Omega \) be a finite ordered alphabet. A \( k \)-word \( s \) over the alphabet \( \Omega \) is an ordered \( k \)-tuple \( s = s_1s_2\ldots s_k \) of symbols from \( \Omega \). The set of \( k \)-words over \( \Omega \) is denoted by \( \Omega^k \). Obviously, a \( k \)-word can be treated as a linear \( k \)-permutation, and thus one can naturally obtain a circular \( k \)-permutation from the \( k \)-word. Let \( u,v \in \Omega^k \), and define a relation \( \sim \) on \( \Omega^k \) by \( u \sim v \) provided that \( u \) and \( v \) have the same circular permutation. Then it is easy to check that \( \sim \) is an equivalence relation, and the equivalence class of \( u \) is a subset of \( \Omega^k \) consisting of all words with the same circular permutation as \( u \). We denote by \( CL(u) \) the equivalence class of \( k \)-word \( u \): \( CL(u) = \{ v \in \Omega^k : u \sim v \} \).

On the other hand, once a circular \( k \)-permutation is obtained from \( k \)-word \( u \), one can easily list \( k \) linear \( k \)-permutations (repetition is allowed) corresponding to the circular permutation. For convenience, we denote the multiset of all these linear \( k \)-permutations by \( C_pL(u) \). Obviously, the underlying set of \( C_pL(u) \) is the equivalence class \( CL(u) \).

The Burrows Wheeler Transform (BWT), introduced by Burrows and Wheeler in 1994, is a well founded mathematical transformation on sequences, and is considered as an extremely useful tool in the field of lossless textual data compression. The transform does not perform any compression but modifies the string in a way to make it easy to compress with a secondary algorithm. Now let’s explain this transformation from the view of combinatorics: given an input \( k \)-word \( u \), BWT first forms the multiset \( C_pL(u) \) and sorts its members lexicographically, and then extracts the last character of each word in the sorted multiset \( C_pL(u) \). BWT\((u)\), the output of BWT is a string \( L \) consisted of these extracted characters. This transformation also computes the index \( I \) that stands for the position of the original word \( u \) in the sorted multiset \( C_pL(u) \). With only BWT\((u)\) and the index \( I \) there is an efficient algorithm to recover the original word.

For example, let \( u=ATGGTGCACCTGACT \), then the corresponding circular permutation is shown in Fig. 1, while the multiset \( C_pL(u) \) and the sorted \( C_pL(u) \) are as follows.

\[
C_pL(u) = \begin{bmatrix}
ATGGTGCACCTGACT \\
TATGTGCACCTGAC \\
CTATGTCGACCTGA \\
ACTATGTGCACTTG \\
GACTATGTGCACTCT \\
TGACTATGTGCAACC \\
CTGACTATGTGCAAC \\
CTGACTATGTGCAAC \\
CTGACTATGTGCAAC \\
ACCTGACTATGTGCT \\
\vdots \\
\end{bmatrix}
\]

the sorted \( C_pL(u) \):

\[
C_pL(u) = \begin{bmatrix}
ACCTGACTATGTGCT \\
ACTATGTGCACTTG \\
ATGGTGCACCTGACT \\
CACCCTGACTATGTTG \\
CCTGACTATGTGCCA \\
CTATGGTGCACTCG \\
CTGACTATGTGCAAC \\
GACTATGTGCACTCT \\
GCACCTGACTATGCT \\
\vdots \\
\end{bmatrix}
\]

So, BWT\((u)\)=CGTGAACCTTTGCGCA, and \( I=3 \).

From the discussion above, one can obtain the following proposition.

Proposition 1. \( u \sim v \) if and only if BWT\((u)\)= BWT\((v)\).

Proposition 1 implies that BWT captures some common features of equivalent words. As we know, how different two words can be evaluated by the factors excluding their common features. On the basis of this idea, in what follows, we will show a matrix representation for a DNA sequence by means of a subtraction between the original word and its BWT word.
3. The Matrix Representation of a DNA Sequence

Usual representation of a DNA primary sequence is that of a string of letters A, G, C, and T, which signify the four nucleic acid bases adenine, guanine, cytosine, and thymine, respectively. In other words, a DNA sequence \( x = x_1 x_2 \ldots x_n \) is an \( n \)-word over the alphabet \( \Omega = \{ A, C, G, T \} \). In order to numerically characterize a DNA sequence, we assign the following fractions to the four bases:

- A: 1/3
- C: 1/5
- G: 1/7
- T: 1/11

Then DNA sequence \( x \) can be transformed into a sequence of real numbers. Here, the 3, 5, 7, 11 are different prime numbers that could make the mapping is one-to-one. Take \( x = \text{ATGGTCACC TGA} \) and \( \text{BWT}(x) = \text{CGTGAACCTTTGCGA} \) as examples, the corresponding sequences of real numbers are

\[
\phi(x) = \frac{1}{3} x_1 + \frac{1}{5} x_2 + \frac{1}{7} x_3 + \frac{1}{11} x_4, \quad (j = 1, 2, \ldots, n)
\]

From the two sequences of real numbers, a “subtraction matrix” can be constructed by the formula:

\[
M(x) = (a_{ij}) \quad a_{ij} = \phi(x_i) - \phi(\text{BWT}(x_j))
\]

For instance, the matrix corresponding to \( x = \text{ATGGTCACC TGA} \) is

\[
M(x) = \begin{bmatrix}
1/3 & 1/5 & 1/7 & 1/11 & 1/3 & 1/5 & 1/7 & 1/11 & 1/3 & 1/5 \\
3 & 5 & 7 & 11 & 3 & 5 & 7 & 11 & 3 & 5 \\
11 & 5 & 7 & 11 & 11 & 7 & 11 & 11 & 7 & 11 \\
7/5 & 7/7 & 7/11 & 7/7 & 7/5 & 7/7 & 7/11 & 7/7 & 7/5 & 7/7 \\
... & ... & ... & ... & ... & ... & ... & ... & ... & ...
\end{bmatrix}
\]

Based on the knowledge of algebra, we have

Proposition 2. Suppose \( u \sim v \), and their subtraction matrices are \( M(u) \) and \( M(v) \), respectively. Then \( M(u) \sim M(v) \), that is, the two matrices can be converted into one another by a series of the first type of elementary row operations.

From proposition 2, we immediately get the following conclusion.

Proposition 3. For \( \forall x, y \in \text{CL}(u) \), \( \|M(x)\|_2 = \|M(y)\|_2 \), where \( \|M\|_2 \) is the spectral norm of matrix \( M \).
The data files can be retrieved from GenBank (see Table 2). China, one from Finland, which is used as a reference. And hantaviruses analysed in this study, twelve are isolated in other literature [7, 14, 16, 17, 33-35].

This result is similar to that reported in other literature [7, 14, 16, 17, 33-35].

As another application, we choose the S segments of 13 hantaviruses to construct the phylogenetic tree. Hantaviruses (HV) are negative sense RNA viruses in the Bunyaviridae family. The name hantavirus is derived from the Hantan River area in South Korea. Humans may be infected with hantaviruses through rodent bites, urine, saliva or contact with rodent waste products. Some individuals HV types. This result is in accordance with that reported by Ref. Yao et al. [37].

In the same way, we obtain a 13×13 distance matrix, and then construct the phylogenetic tree of the 13 hantaviruses (see Fig. 3). From Fig. 3, we find that Sotkamo can be distinguished easily from other two groups of hantaviruses. The 5 Hantaan (HTN) viruses, Z10, Z5, Z251, ZLS611, ZLS-12, form a separate branch, while the 7 Seoul (SEO) viruses, ZT71, ZT10, Z37, Gou3, Z15, K24-e7, K24-v2, form another branch. A closer look at the subtree of 12 HV strains from Zhejiang Province shows that the phylogeny was closely related to the isolated regions, but had no distinct relationship with the host. In other words, HV distributions showed geographical clustering within individual HV types. This result is in accordance with that reported by Ref. Yao et al. [37].

Acknowledgements

This work was partially supported by the National Natural Science Foundation of China (No. 11171042), the Program for Liaoning Excellent Talents in University (No. LJQ2011122), the Liaoning BaiQianWan Talents Program (No. 2012921060), and the Open Project Program of Food Safety Key Lab of Liaoning Province (No. LNSAKF2011034).

References


