Available online at www.sciencedirect.com
ScienceDirect
Biochemical and Biophysical Research Communications 368 (2008) 223–230
BBRC
www.elsevier.com/locate/ybbrc

A novel feature-based method for whole genome phylogenetic analysis without alignment: Application to HEV genotyping and subtyping

Zhihua Liu [a,b,c,*], Jihong Meng [d], Xiao Sun [a,*]

[a] State Key Laboratory of Bioelectronics, Southeast University, Nanjing 210096, PR China
[b] Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Harvard Medical School, 44 Binney Street, Boston, MA 02115, USA
[c] Harvard School of Public Health, 677 Huntington Avenue, Boston, MA 02115, USA
[d] Department of Microbiology and Immunology, School of Medicine, Southeast University, Nanjing 210009, PR China

Received 20 December 2007
Available online 28 January 2008

Abstract

Traditional phylogenetic analysis is based on multiple sequence alignment. With the development of worldwide genome sequencing project, more and more completely sequenced genomes become available. However, traditional sequence alignment tools are impossible to deal with large-scale genome sequence. So, the development of new algorithms to infer phylogenetic relationship without alignment from whole genome information represents a new direction of phylogenetic study in the post-genome era. In the present study, a novel algorithm based on BBC (base–base correlation) is proposed to analyze the phylogenetic relationships of HEV (Hepatitis E virus). When 48 HEV genome sequences are analyzed, the phylogenetic tree that is constructed based on BBC algorithm is well consistent with that of previous study. When compared with methods of sequence alignment, the merit of BBC algorithm appears to be more rapid in calculating evolutionary distances of whole genome sequence and not requires any human intervention, such as gene identification, parameter selection. BBC algorithm can serve as an alternative to rapidly construct phylogenetic trees and infer evolutionary relationships.

Keywords: Whole genome phylogeny; Alignment-free; Hepatitis E virus; Base–base correlation

With more and more genome sequences have been released, the genome sequence comparison and phylogenetic tree construction based on whole genomes attract more attention [1–6]. Most existing approaches for phylogenetic inference use multiple alignment of sequences and assume some regular evolutionary models [7]. In fact, it is computationally expensive in comparing relatively large genome sequences with the length in units of kilobases, megabases, or even gigabases [8,9]. Alternatively, genome phylogenetic tree construction is simply based on alignment of some special genes or conserved fragments [10]. However, it leads to the bias caused by using different genes or genome fragments and loss of whole genome information [11]. Moreover, in sequence alignment, insertions and deletions are poorly evaluated due to the assumption of regular evolutionary models [9,12–14]. Therefore, it is valuable and important to develop novel alignment-free methods for whole genome phylogenetic analysis at a fast rate.

In the present study, we developed a novel sequence feature, named as BBC (base–base correlation), for whole genome sequence analysis. This algorithm based on BBC feature is inspired from the method which analyzed the total information of all four bases between two positions with an interval of k using mutual information function in information theory [15]. We developed it and emphasized the information of different base pairs within the range of k. It improved the resolving power and provided a more appropriate description of sequence dissimilarity [16]. We evaluated BBC algorithm using the HEV (Hepatitis E virus) genome sequences.

* Corresponding author. Address: State Key Laboratory of Bioelectronics, Southeast University, Nanjing 210096, PR China.
E-mail addresses: zhliu@jimmy.harvard.edu, zhliu@seu.edu.cn (Z.H. Liu), xsun@seu.edu.cn (X. Sun).

0006-291X/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.bbrc.2008.01.070

HEV is a non-enveloped, positive-sense, single-stranded RNA virus and belongs to *Hepevirus* genus under the separate family of Hepeviridea [17]. The genome of HEV is approximately 7.2 kb in length and contains a short 5′ untranslated region (5′ UTR), three overlapped open reading frames (ORF1, ORF2, and ORF3) and a short 3′ UTR. By BBC algorithm, 48 HEV genomes were distinctly clustered into four genotypes identical to the traditional classification.

Materials and methods

Sequences. We retrieved a total of 48 full-length HEV genome sequences from NCBI (http://www.ncbi.nlm.nih.gov/). Abbreviation for the strains, accession number, nucleotide length, country, and genotype of all HEV genomes [17] are shown in Table 1.

Base–base correlation (BBC). DNA sequences can be viewed as symbolic strings composed of the four letters $(B_1,B_2,B_3,B_4) \equiv (A,C,G,T)$. The probability of finding the base B_i is denoted by p_i ($i = 1,2,3,4$). Then BBC feature is defined as the following:

$$T_{ij}(k) = \sum_{l=1}^{k} p_{ij}(l) \cdot \log_2\left(\frac{p_{ij}(l)}{p_i p_j}\right) \quad i,j \in \{1,2,3,4\} \tag{1}$$

Here, $p_{ij}(l)$ means the joint probabilities of bases i and j at a distance of l. $T_{ij}(k)$ represents the average relevance of the two-base combination with different gaps from 1 to k. It reflects a local feature of two bases with a range of k. For each genome sequences m, BBC feature has 16 parameters and constitutes a 16-dimensional feature vector $V_m^z (z = 1,2,\ldots,16)$.

Let L be a whole genome sequences length ($1 \leqslant k \leqslant L$). Thus, $T_{ij}(L)$ contains all base pairs information for this genome sequence. Theoretically, BBC feature extract more fully genome information when k is larger. Nucleosomal DNA contains a core DNA region with a stable length of 147 bp, which is relatively resistant to digestion by nucleases [18,19]. Moreover, we find that BBC feature curve has no considerable changes when $k > 147$. So, we take $k = 147$ in BBC feature calculation for genome sequence in the present study.

Statistical independence of two bases in a distance l is defined by $p_{ij}(l) = p_i p_j$. Thus, deviations from statistical independence is defined by

$$D_{ij}(l) = p_{ij}(l) - p_i p_j \tag{2}$$

We expand BBC feature $T_{ij}(k)$ using a Taylor series in terms of Eq. (2)

$$\begin{aligned} T_{ij}(k) &= \sum_{l=1}^{k} p_{ij}(l) \cdot \log_2\left(\frac{p_{ij}(l)}{p_i p_j}\right) \\ &= \sum_{l=1}^{k} \left[D_{ij}(l) + p_i p_j\right] \cdot \ln\left[1 + \frac{D_{ij}(l)}{p_i p_j}\right] \\ &= \sum_{l=1}^{k} \left[D_{ij}(l) + p_i p_j\right] \cdot \left[\frac{D_{ij}(l)}{p_i p_j} - \frac{D_{ij}^2(l)}{2 p_i p_j} + \cdots\right] \\ &= \sum_{l=1}^{k} D_{ij}(l) + \frac{D_{ij}^2(l)}{2 p_i^2 p_j^2} + o\left[D_{ij}^3(l)\right] \end{aligned} \tag{3}$$

This mathematical transformation further increases the calculation speed and solves effectively the problem of $0 \cdot \log_2 0$ (i.e. $p_{ij}(l) = 0$ in Eq. (1)).

The distance matrix. Given two sequences m and n, the distance H_{mn} between two sequences m and n is defined as the following:

$$H_{mn} = \sqrt{\sum_{z=1}^{16} \left(V_m^z - V_n^z\right)^2} \quad m,n = 1,2,\cdots,N \tag{4}$$

Here, V_m and V_n represent the 16-dimensional feature vectors of sequences m and n. N is the total number of all sequences analyzed. According to Eq. (4), H_{mn} obviously satisfies the definition of distance: (i) $H_{mn} > 0$ for $m \neq n$; (ii) $H_{mm} = 0$; (iii) $H_{mn} = H_{nm}$ (symmetric); (iv) $H_{mn} \leqslant H_{mq} + H_{nq}$ (triangle inequality). For N sequences, a real symmetric $N \times N$ distance matrix is then obtained.

Clustering. Accordingly, a real symmetric $N \times N$ matrix is used to reflect the evolutionary distance between N sequences. The clustering tree is constructed using neighbor-joining method. The reliability of the branches is assessed by performing 1000 resamplings. Bootstrap values, greater than 900, are shown on nodes.

Phylogenetic analysis based on sequence alignment. Multiple sequence alignment was performed by ClustalX [22]. Genetic distances were calculated using the Kimura two-parameter method. Phylogenetic tree was constructed using neighbor-joining method. Bootstrap (1000 replicates) values were indicated on the nodes. Calculation of genetic distances, construction of phylogenetic tree and assessment of the reliability of the tree were all performed by using MEGA software [24].

Results

Genotyping of HEV based on BBC feature for whole genome sequences

The genotypes of HEV based on BBC feature in whole genome sequences are shown in Fig. 1. Forty-eight HEV genomes were distinctly divided into four genotypes. This was in very good agreement with that of previous study [17]. Genotype I included 16 HEV strains. Seven strains, B1 (Bur-82), B2 (Bur-86), I2 [Mad-93], I3, Np1 (TK15/92), P2 [Abb-2B], and Yam-67 were classified into subtype Ia. They were derived from Burma (B1, B2), India (I2, I3, Yam-67), Nepal (Np1) and Pakistan (P2). Subtype Ib contained only I1 (FHF) strain that was isolated from India. Six strains, C1, C2, C3, C4, China Hebei and P1 (Sar-55) were clustered together and classified into subtype Ic. They were isolated from China (C1, C2, C3, C4, China Hebei) and Pakistan (P1), respectively. Except China Hebei strain, the other Chinese strains in subtype Ic were isolated from Xinjiang Ughur Autonomous Region during 1986–1988 [25]. P1 (Sar-55) was isolated from an outbreak in Sargodha in 1987 [26]. It is similar to the four Xinjiang HEV sequences and is attributed to close religious and business ties between Pakistan and Xinjiang. Subtype Id and subtype Ie were represented by Morocco strain, and T3 strain from Chad, respectively. In some studies, Morocco and T3 were categorized into a single African cluster, namely genotype V [27,28]. However, many studies thought Morocco and T3 were more closer to the other subtypes in genotype I and represented two different subtypes of genotype I [17,29–31]. Our study found that Morocco and T3 were more genetically closer to the other subtypes in genotype I, supporting that these African HEV strains (Morocco, T3) were clustered into genotype I and represented two different genotypes. Genotype II contained only a complete genome M1, which was isolated from Mexico. Genotype III included 17 HEV strains. Among these strains, seven strains (HE-JA10, JMY-HAW, JKN-Sap, swUS1, US1, US2, Arkell) were classified into subtype IIIa. They were derived from Japan (HE-JA10, JMY-HAW, JKN-Sap), USA (swUS1, US1, US2), and Canada (Arkell), respectively. JSO-Hyo03L, JMO-Hyo03L, JYO-Hyo03L, JTH-Hyo03L,