A Novel Composition Coding Method of DNA Sequence and Its Application

Ronghui Wu\textsuperscript{1}\textsuperscript{*}, Qiguang Hu\textsuperscript{1,2}\textsuperscript{*}, Renfa Li\textsuperscript{1}\textsuperscript{*}, Guangxue Yue\textsuperscript{3}

\textsuperscript{1} School of Information Science and Engineering, Hunan University, Changsha Hunan, 410082, China
\textsuperscript{2} Computer Science Department of Hunan City University, Yiyang, Hunan 413000, China
\textsuperscript{3} College of Mathematics and Information Engineering, Jiaxing University, Jiaxing Zhejiang, 314001, China

(Received January 24, 2011)

Abstract: Mathematical analysis is very important for extracting information from the sequences. We propose a novel composition coding method for this sequence. This method is a combination of segmental probability of four nucleotides correlation factors and four components of the conventional nucleotides composition. Using the composition vectors, phylogeny analysis of species are given for the complete mtDNA sequences of eight Eutherian mammals.

1. Introduction

In recent years, effective representation of long DNA sequence has led to several innovative techniques to provide useful ways for viewing, sorting, and comparing various sequences. For example, Hamori and Ruskin [1], Randić[2,23-25], Nandy [3,22], Zhang [4], Liao [5-8,15-21], considered a real DNA primary sequence as a curve embedded in 2-D plane or 3-D space. The model essentially consists of plotting a point corresponding to a base by moving a step in a direction depending on the defined association of a base with the direction. The cumulative plot of such points produces a
graph that corresponds to the sequence of bases in the gene fragment under consideration.

Also, their researchers have introduced several methods for comparison of DNA sequences based on several mathematical invariants. For example, Randic et al[2,9] have considered kinds of condensed matrices. Bo Liao et al. [10] have presented three characteristic representations corresponding to the three classifications of the bases of DNA and constructed several sets of matrices to represent DNA primary sequences. Bailin Hao et al. [11-13] have presented a k-string composition vector approach without using sequence alignment to represent DNA primary. This method is based on counting the appearance frequency of nucleotides of a fixed length in the collection of DNA sequences of a species.

In this paper, we present a novel composition coding method to represent DNA primary sequences. This method is a combination of segmental probability of four nucleotides correlation factors and four components of the conventional nucleotides composition, which is based on counting appearance probability of nucleotides. The advantage of our approach is that it contains more sequence effects than the 4-D conventional nucleotide composition and reduces the time complexity of this method. We will analyze the phylogeny of the complete mtDNA sequences of eight Eutherian mammals to illustrate the utility of our new approach.

2. A novel composition coding method

The essence of our approach is, on one hand, to include the main feature of nucleotides composition, but on the other, to include information of segmental nucleotides composition. The conventional nucleotides composition contains four components, or discrete numbers, each reflecting the occurrence probability of one of the 4 native nucleotides in a DNA sequence. For the nucleotides composition, however, there are some other elements in addition to the four components. It is through these additional discrete numbers that the sequence effect of a DNA sequence is approximately reflected and improvements are made, as will be shown below.

Considering a DNA sequence chain of L nucleotides:
\[ R_1R_2R_3R_4R_5R_6R_7\ldots R_L \]  

(1)

We divide L bases into any m sections, which are as follows. m can take any value, and does not affect the calculation results. So each section includes \( \left\lfloor \frac{L}{m} \right\rfloor \) bases and the last section includes \( \{x+L-x*m\} \) bases.

\[ \frac{R_1R_2\ldots R_x}{1} \frac{R_{x+1}\ldots R_{x+x}}{2} \ldots \frac{R_{(m-1)x+1}\ldots R_L}{m} \]  

(2)

We count the probability of 4 bases in these m sections, and label them as \( P_i^A, P_i^T, P_i^G, P_i^C \), where i is the i section, and A,T,G,C are A,T,G,C bases. Effect can be approximately reflected with m sections of sequences correlated factors as defined below:

\[
\begin{align*}
\theta_1 &= \frac{1}{m} \sum_{i=1}^{m} \left| P_i^A - P_i^G \right| \\
\theta_2 &= \frac{1}{m} \sum_{i=1}^{m} \left| P_i^A - P_i^C \right| \\
\theta_3 &= \frac{1}{m} \sum_{i=1}^{m} \left| P_i^A - P_i^T \right| \\
\theta_4 &= \frac{1}{m} \sum_{i=1}^{m} \left| P_i^G - P_i^C \right| \\
\theta_5 &= \frac{1}{m} \sum_{i=1}^{m} \left| P_i^G - P_i^T \right| \\
\theta_6 &= \frac{1}{m} \sum_{i=1}^{m} \left| P_i^C - P_i^T \right|
\end{align*}
\]  

(3)

Where \( \theta_1 \) is called the correlation factor of A and G, \( \theta_2 \) the correlation factor of A and C, \( \theta_3 \) the correlation factor of A and T, \( \theta_4 \) the correlation factor of G and C, \( \theta_5 \) the correlation factor of G and T, \( \theta_6 \) the correlation factor of C and T. The correlation factors between the four-letters alphabet \{A,C,G,T\} are 16. However, the correlation factor of A and G and G and A's are the same, and so A and C,A and T, G and C,G and T, C and T; and the correlation factor of A and A is 0, and so G and G,C and C, T and T.

As we can see from formula 3, the sequence effect of a DNA sequence can be, to
some extent, reflected through 6 sequence-correlation factors $\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6$. Now let us augment the formulation of nucleotides composition to include 6 discrete numbers. To realize this, instead of using a 4-D(dimensional) vector defined by 4 components, we use a (4+6)-D vector defined by 4+6 discrete numbers to represent a DNA sequence $X$; i.e.,

$$X = \begin{bmatrix} P_A \\ P_G \\ P_C \\ P_T \\ \theta_1 \\ \theta_2 \\ \vdots \\ \theta_6 \end{bmatrix},$$

Where $P_A$ is the occurrence probability of A in a DNA sequence, and so $P_G, P_C, P_T$. Such 10 components as formulated by Eqs.4 is our new composition coding for DNA sequence $X$. It has the following two advantages:(a) It contains more sequence effects than the 4-D conventional nucleotide composition as reflected by 6 sequence correlation factors with different correlation (see Eq.3). (b) we deal with a DNA sequence for segmentation way. It is a simple calculation, and reduces the time complexity of this method. $M$ value can be arbitrarily taken, and it will not affect the results.

3. Application

In order to facilitate the quantitative comparison of different species in terms of their composition vectors, we introduce a distance scale as defined below.

We suppose that there are two species $i$ and $j$, the composition vectors are $P_A^i, P_G^i, P_C^i, P_T^i, \theta_1^i, \theta_2^i, \ldots, \theta_6^i$ and $P_A^j, P_G^j, P_C^j, P_T^j, \theta_1^j, \theta_2^j, \ldots, \theta_6^j$, respectively. We will illustrate the use of the novel composition coding vectors of DNA sequences with an examination of similarities/ dissimilarities among the complete mtDNA sequences of eight Eutherian mammals of table 1.
Table 1 The complete mtDNA sequences of eight Eutherian mammals

<table>
<thead>
<tr>
<th>No</th>
<th>Species scientific name</th>
<th>abbreviation</th>
<th>accession</th>
<th>Length(bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Papio hamadryas baboon</td>
<td>Y18001</td>
<td>16521</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Hylobates lar gibbon</td>
<td>X99256</td>
<td>16472</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Pongo pygmaeus orangutan</td>
<td>D38115</td>
<td>16389</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Gorilla gorilla gorilla</td>
<td>D38114</td>
<td>16364</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Pan troglodytes c.chimpa</td>
<td>D38116</td>
<td>16563</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Pan paniscus p.chimpa</td>
<td>D38113</td>
<td>16554</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Equus caballus horse</td>
<td>X79547</td>
<td>16660</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Ceratotherium simum w.rhinoc</td>
<td>Y07726</td>
<td>16832</td>
<td></td>
</tr>
</tbody>
</table>

We construct the 10 composition vectors of every species in table 2, where m is 50.

In this paper, we compute the similarities among such vectors in this way: we calculate the Euclidean distance between species.

The distance $d_{ij}$ between the two species is

$$d_{ij} = \sqrt{[P_A^i - P_A^j]^2 + [P_G^i - P_G^j]^2 + [P_C^i - P_C^j]^2 + [P_T^i - P_T^j]^2 + \sum_{k=1}^{6} (\theta_k^i - \theta_k^j)^2} \tag{5}$$

Table 2 using our new coding method to construct the 10 composition vectors of 8 species

<table>
<thead>
<tr>
<th>species</th>
<th>(P_A, P_G, P_C, P_T, \theta_1, \theta_2, \cdots, \theta_6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>baboon</td>
<td>(0.3144,0.1313,0.3055,0.2488,0.0379,0.0043,0.0084,0.0359,0.0162,0.0069)</td>
</tr>
<tr>
<td>gibbon</td>
<td>(0.3059,0.1370,0.3176,0.2396,0.0326,0.0037,0.0079,0.0378,0.0132,0.0087)</td>
</tr>
<tr>
<td>orangutan</td>
<td>(0.3055,0.1323,0.3244,0.2378,0.0336,0.0039,0.0076,0.0430,0.0133,0.0104)</td>
</tr>
<tr>
<td>gorilla</td>
<td>(0.3092,0.1320,0.3069,0.2520,0.0353,0.0036,0.0069,0.0357,0.0170,0.0062)</td>
</tr>
<tr>
<td>c.chimpa</td>
<td>(0.3133,0.1270,0.3069,0.2527,0.0381,0.0041,0.0071,0.0379,0.0183,0.0066)</td>
</tr>
<tr>
<td>p.chimpa</td>
<td>(0.3113,0.1289,0.3080,0.2518,0.0365,0.0040,0.0071,0.0371,0.0175,0.0066)</td>
</tr>
<tr>
<td>horse</td>
<td>(0.3216,0.1342,0.2854,0.2588,0.0398,0.0069,0.0074,0.0272,0.0174,0.0035)</td>
</tr>
<tr>
<td>w.rhinoc</td>
<td>(0.3341,0.1289,0.2796,0.2574,0.0472,0.0070,0.0106,0.0275,0.0183,0.0038)</td>
</tr>
</tbody>
</table>

The smaller Euclidean distance, the more similar the DNA sequence. That is to say, the distances between evolutionary closely related species are smaller, while those between evolutionary disparate species are larger.

Using the Euclidean distance measure, we obtain the similarity/dissimilarity matrix(see table 3) for the 8 complete mtDNA sequences belonging to different species, and the elements of the similarity matrix are used to construct phylogenic tree using the Neighbor.exe program in PHYLIP software. In fig.1, we show the phylogenic tree belonging to 8 species, and in fig.2, we use the vertical and horizontal method [14] to construct the phylogenetic tree of the 8 species.
Observing fig.1 and 2, we find the more similar species pairs are p.chimp-c.chimp, gibbon-orangutan, and horse-w.rhino. Similar results have been obtained by Liao[14]. Our approach is possible and available.

**Table 3** The similarity/dissimilarity matrix for Table 1 based on the Euclidean distances between the 10 composition vectors

<table>
<thead>
<tr>
<th>species</th>
<th>baboon</th>
<th>gibbon</th>
<th>orangutan</th>
<th>gorilla</th>
<th>c.chimpa</th>
<th>p.chimpa</th>
<th>horse</th>
<th>w.rhinoc</th>
</tr>
</thead>
<tbody>
<tr>
<td>baboon</td>
<td>0</td>
<td>0.0195</td>
<td>0.0255</td>
<td>0.0071</td>
<td>0.0069</td>
<td>0.0062</td>
<td>0.0258</td>
<td>0.0363</td>
</tr>
<tr>
<td>gibbon</td>
<td>0</td>
<td>0.0102</td>
<td>0.0183</td>
<td>0.0224</td>
<td>0.0194</td>
<td>0.0434</td>
<td>0.0548</td>
<td></td>
</tr>
<tr>
<td>orangutan</td>
<td>0</td>
<td>0.0247</td>
<td>0.0265</td>
<td>0.0242</td>
<td>0.0509</td>
<td>0.0611</td>
<td></td>
<td></td>
</tr>
<tr>
<td>gorilla</td>
<td>0</td>
<td>0.0075</td>
<td>0.0044</td>
<td>0.028</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>c.chimpa</td>
<td>0</td>
<td>0.0036</td>
<td>0.0276</td>
<td>0.0377</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p.chimpa</td>
<td>0</td>
<td>0.0287</td>
<td>0.0399</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>horse</td>
<td>0</td>
<td>0.0169</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>w.rhinoc</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Fig. 1.** Using the Neighbor.exe program in PHYLIP software to construct the phylogenic tree

**Fig. 2.** Using the vertical and horizontal method [14] to construct the phylogenic tree
4. Conclusion

High complexity and loss of the sequence information are major problems in previous DNA sequence coding. Our new composition coding method for DNA sequence can solve these two problems. This method is a combination of segmental probability of four nucleotides correlation factors and four components of the conventional nucleotides composition. The similarity/dissimilarity matrix for the 8 complete mtDNA sequences belonging to different species is built, and the elements of the similarity matrix are used to construct phylogenetic tree.

Acknowledgment

This work is supported by the National Nature Science Foundation of China (Grant 60973082), the National Nature Science Foundation of Hunan province (Grant 07JJ5080), the Planned Science and Technology Project of Hunan Province (Grant 2009FJ3195) and the National Nature Science Foundation of Zhejiang province (Grant Y1090264).

References


