DNA Sequence Homology Recognition based on Similarity Measurement

ZHANG Junyan[1, a], YANG Chenhui[1, b] CHEN Xiaodan[1, c]

[1] Information Science and Technology College; Key Laboratory of Pattern Recognition and Intelligent Information Processing of Sichuan, Chengdu University, China

[a]67306683@qq.com, [b]123868907@qq.com, [c]94505078 @qq.com

Keywords: DNA Sequence; Homology Recognition; Similarity Measurement; Markov Model.

Abstract. DNA sequence homology recognition is a key problem in bioinformatics. In this paper, we solve this problem by use of the probability method instead of traditional sequence alignment because DNA character sequence satisfies the Markov properties. Hence, second order Markov model is used as the characteristic of DNA sequence. The similarity measurement is defined based on two-step transition probability. And then our SHR algorithm is put forward. The contrast experiments show that SHR algorithm can recognize DNA sequence homology correctly in higher processing speed.

Introduction

Generally, a DNA sequence is treated as a long string of characters with a four-character set $\sum$={A, C, G, T}. Thus, any one DNA sequence $S\in\sum^*$[1]. DNA sequence homology recognition is an important problem in bioinformatics, which refers that two or more DNA sequences are compared through some mathematical algorithms so as to determine homology on the basis of the similarity [2].

There are many ways to solve this problem, such as: (1) 2-D or 3-D graphics are employed to represent DNA sequences so as to analyze the relationship among DNA sequences [3], which has better visual effect but lower speed. (2) The four alphabets of DNA sequences in $\sum$ are mapped into numerical sequences and their features are compared by numerical analysis [4], which can make us obtain better predictive effect but lack of a unified measurement. (3) DNA sequences are regarded as character strings or texts, and the relative distances are adopted to analyze DNA sequences [5]. Thus, the methods of text compression can be introduced to improve speed, but some redundant sequences still exist. (4) Non-alignment methods are put to use for analyzing features of DNA sequences in order to improve efficiency though the segmentation and positioning are difficult to be achieved [6]. The number of DNA sequences is usually very large and their structures are very complicated. Therefore, the existed algorithms have their own advantages and disadvantages respectively.

In this paper, we concentrate on DNA sequence homology recognition based on similarity measurement by use of second order Markov Model [7]. The remainder of this paper is organized as follows. First of all, the relative concepts and definitions are presented. And then, the description of the problem-solving ideas and our SHR algorithm is put forward. After that, the contrast experiments and results are listed. Finally, we conclude this paper.

Concepts and Definitions

Second Order Markov Model. Markov model is one of the most important stochastic processes, and it is widely applied to modern biology, physics, business, geology, atmospherics, and so on.

Definition 1. Let $\rho(n)=\{x_n, n\in T\}$ be a stochastic process of discrete states with state space I and of non-negative integer parameters n. If $\rho(n)$ satisfies condition: $P\{x(n+1)=i_{n+1}|\ x(0)=i_0, x(1)=i_1, \ldots, x(n)=i_n\}$, $\rho(n)$ is called a Markov chain, where, $i_0, i_1, \ldots, i_n\in I$.

Definition 2. Conditional probability $p_{ij}(n)=p\{x_{n+1}=j|\ x_n=i\}$ $(i,j\in I,\ n\geq 0)$ is called one-step transition probability of Markov chain with state space I. Let P be one-step transition probability matrix which is made of p_{ij}, and $P=(p_{ij})$.

Definition 3. Conditional probability $p_{ij}^{(2)} = p\{x_{m+1}=k \mid x_m=g\}(k,g\in I,\ m\geq 0)$ is called two-step transition probability of Markov chain with the state space I. And $P^{(2)}=(p_{ij}^{(2)})$ is named two-step transition probability matrix, where $p_{ij}^{(2)}\geq 0$, and $\sum_{j\in I} p_{ij}^{(2)} = 1$.

Definition 4. Second order Markov model can be denoted by $\lambda=(\pi, P, P^{(2)})$, where $\pi=(\pi_1, \pi_2,\ldots \pi_k)$ is the initial state and k is the number of possible states of the sequences.

Characteristic matrix of DNA Sequence. For all the DNA sequences, we have $I=\sum$. The next base has nothing to do with the last one. Therefore, DNA sequence can be treated as Markov chain and it also can be described by two-step transition probability. Hence, each DNA sequence is corresponding to one $P^{(2)}$ which is looked on as its characteristic matrix. So we have:

$$P^{(2)}=(p_{ij}^{(2)})=\begin{bmatrix} p_{11}^{(2)} & p_{12}^{(2)} & \cdots & p_{1n}^{(2)} \\ p_{21}^{(2)} & p_{22}^{(2)} & \cdots & p_{2n}^{(2)} \\ \cdots & \cdots & \cdots & \cdots \\ p_{n1}^{(2)} & p_{n2}^{(2)} & \cdots & p_{nn}^{(2)} \end{bmatrix}=\begin{bmatrix} p_{AA}^{(2)} & p_{AC}^{(2)} & p_{AG}^{(2)} & p_{AT}^{(2)} \\ p_{CA}^{(2)} & p_{CC}^{(2)} & p_{CG}^{(2)} & p_{CT}^{(2)} \\ p_{GA}^{(2)} & p_{GC}^{(2)} & p_{GG}^{(2)} & p_{GG}^{(2)} \\ p_{TA}^{(2)} & p_{TC}^{(2)} & p_{TG}^{(2)} & p_{TT}^{(2)} \end{bmatrix}. \quad (1)$$

In term of *Definition 1*, we have $p_{ij}^{(2)} = P_{iAj}+P_{iCj}+P_{iGj}+P_{iTj}$, where $i,\ j\in\sum$. Here, $P_{iAj}=\frac{N_{iAj}}{N_{iA}+N_{iC}+N_{iG}+N_{iT}}$. P_{iCj}, P_{iGj} and P_{iTj} can be deduced from this. Hence $p_{ij}^{(2)}$ can be represented as follow:

$$p_{ij}^{(2)}=\frac{Nij}{N_{iA}+N_{iC}+N_{iG}+N_{iT}}. \quad (2)$$

In Eq. 2, $N_{ij}=N_{iAj}+N_{iCj}+N_{iGj}+N_{iT}$. N_{ij} indicates how many times the base pair ij can be present. Meanwhile, N_{iAj} shows how many times three successive bases iAj can be present.

DNA Sequence Similarity Measurement. Given two DNA sequences D_1 and D_2, let $P_{D1}^{(2)}=(p1_{ij}^{(2)})$ and $P_{D2}^{(2)}=(p2_{ij}^{(2)})$ be their characteristic matrixes respectively. In order to achieve DNA sequence homology recognition, we come up with DNA sequence similarity measurement ξ as follows:

$$\xi=\sum_{i\in I}\sqrt{\sum_{j\in I}(p1_{ij}^{(2)}-p2_{ij}^{(2)})^2}. \quad (3)$$

Definition 5. DNA sequence homology recognition: Given an unidentified sequence S and arbitrary number of identified sequence $D_1, D_2,\ldots, D_n$, let $\xi_1, \xi_2,\ldots, \xi_n$ be the similarities between S and each D_k ($1\leq k\leq n$). Suppose $\xi_k=min\{\xi_1, \xi_2,\ldots, \xi_n\}$, we can determine that S is homologous with D_k.

Example and Algorithm

Example. Given two identified DNA sequences D_1 and D_2, determine an unidentified DNA sequence S comes from D_1 or D_2. Here S=GATCACAGGTCT, and:

D_1=ATGGTTGCACCTGACTGATGCTGAGAAGTCTGCTGTCTCTTGCCTGTGGGCAAAGGTGAACCCCGATGAAGTTGGTGGTGAGGCCCTGGGCAGG.

D_2=ATGGTGCACCTAACTGATGCTGAGAAGGCTACTGTTAGTGGCCTGTGGGGAAGGTGAACCCTGATAATGTTGGCGCTGAGGCCCTGGGCAG.

Now, $P_{D1}^{(2)}$, $P_{D2}^{(2)}$ and $P_S^{(2)}$ are calculated according to Eq. 2 and ξ_1, ξ_2 are calculated according to Eq. 3. All the results are listed as follows:

$$P_{D1}^{(2)}=\begin{bmatrix}0.1176 & 0.1765 & 0.5294 & 0.1765\\ 0.1000 & 0.2500 & 0.4500 & 0.2000\\ 0.1875 & 0.2500 & 0.2500 & 0.3125\\ 0.2609 & 0.1739 & 0.3478 & 0.2174\end{bmatrix}, P_{D2}^{(2)}=\begin{bmatrix}0.1579 & 0.2105 & 0.4211 & 0.2105\\ 0.1111 & 0.2222 & 0.4444 & 0.2222\\ 0.1875 & 0.2188 & 0.3125 & 0.2813\\ 0.3810 & 0.1429 & 0.3333 & 0.1429\end{bmatrix},$$

$$P_{S}^{(2)}=\begin{bmatrix}0.3333 & 0.3333 & 0.3334 & 0.0000\\ 0.0000 & 0.5000 & 0.5000 & 0.0000\\ 0.0000 & 0.3333 & 0.0000 & 0.6667\\ 0.5000 & 0.0000 & 0.0000 & 0.5000\end{bmatrix}, \xi_1=0.7679, \text{ and } \xi_2=0.9256.$$

Obviously, $\xi_1<\xi_2$. So we can determine that S is homologous with D_1 in term of *Definition 5*.

Algorithm. In this subsection, our DNA sequence homology recognition algorithm named SHR is put forward which is based on similarity measurement ξ and employs second order Markov model. SHR algorithm consists of two sub-algorithms which are listed as follows:

Algorithm 1: calculate two-step transition probability matrix.

Input: DNA sequence D.

Output: $P_D^{(2)}$

```
i, j∈∑={A,C,G,T}; l=1;
for(i='A'; i≤'T', l≤D.length; i++, l++){
   if(regionmatches(D, "iA", l, 2)) N_iA++;  if(regionmatches(D, "iC", l, 2)) N_iC++;
   if(regionmatches(D, "iG", l, 2)) N_iG++;  if(regionmatches(D, "iT", l, 2)) N_iT ++;}
for(i='A'; i≤'T'; i++)
   for(j='A'; j≤'T'; j++){
      if(regionmatches(D, "iAj", l, 3)&& l≤D.length) {N_iAj++; l++;}
      if(regionmatches(D, "iCj", l, 3)&& l≤D.length) {N_iCj++; l++;}
      if(regionmatches(D, "iGj", l, 3)&& l≤D.length) {N_iGj++; l++;}
      if(regionmatches(D, "iTj", l, 3)&& l≤D.length)  {N_iTj++; l++;} }
for(i='A'; i≤'T'; i++)
   for(j='A'; j≤'T'; j++)
      calculate each p_ij^(2) = Nij / (N_iA+N_iC+N_iG+N_iT) and obtain P_D^(2);
```

Algorithm 2: calculate similarities between each D_i and S.

Input: identified DNA sequences $D_1, D_2,\ldots, D_n$, unidentified DNA sequence S.

Output: $\xi_{minimum}$.

```
i, j∈∑={A,C,G,T};
Calculate P_S^(2) by use of Algorithm 1;
for(k=1; k≤n; k++) {
   Calculate P_Dk^(2) by use of Algorithm 1; ξ_k=0;
   for(i='A'; i≤'T'; i++){
      for(j='A'; j≤'T'; j++)    temp_k=( pDk_ij^(2) – pS_ij^(2) )^2;}
   ξ_k +=temp_k;}
ξ_minimum= min{ξ_1, ξ_2,…, ξ_n}; //It means D_minimum has the homologous with S
```

Contrast Experiments. In this section, the contrast experiments are done between our SHR algorithm and FPM algorithm [8]. Both of them are implemented by use of Java in our experiments.

We select six identified subsequences of DNA sequences. D_1=V0062, D_2=U20753, D_3=D38114, D_4=V00711, D_5=AJ001588, and D_6=AJ002189 which are the mitochondrion DNA sequences of human, cat, gorilla, mouse, rabbit, and pig respectively. All of them can be downloaded from the National Center for Biotechnology Information website (See: http://www.ncbi.nlm.nih.gov).

Subsequently, we randomly select three unidentified DNA sequences as follows: S_1=GATCACAGGTCT, S_2=CTGAGATCTGA, and S_3=TCGATCTGACTTT.

In the respect of DNA sequence homology recognition, ξ is used as similarity measurement in SHR algorithm (Eq. 3), but frequency f is used as similarity measurement in FPM algorithm [8]. Therefore, ξ_k^i is employed to denote the value of similarity between S_i and D_k. At the same time, f_k^i is used to denote the frequency that S_i appears in D_k. The results of contrast experiments are shown in Table 1, where t indicate the time to compute each ξ_k^i or f_k^i, and its unit is second.

Table 1. The results of contrast experiments

D_k & S_i	D_1		D_2		D_3		D_4		D_5		D_6	
	ξ_1^i	f_1^i	ξ_2^i	f_2^i	ξ_3^i	f_3^i	ξ_4^i	f_4^i	ξ_5^i	f_5^i	ξ_6^i	f_6^i
S_1	0.8297	7	0.7438	19	0.8054	13	0.6137	32	0.8092	10	0.7628	21
t	0.790	0.811	0.903	0.970	0.840	0.892	0.780	0.793	0.621	0.756	0.940	1.200
S_2	0.6839	35	0.8124	3	0.8346	6	0.7960	17	0.7391	23	0.8625	10
t	0.762	0.780	0.965	1.020	0.823	0.947	0.740	0.841	0.632	0.779	0.920	0.991
S_3	0.9530	6	0.6320	39	0.7245	27	0.8102	18	0.8236	13	0.7520	20
t	0.757	0.792	0.912	1.140	0.862	0.956	0.773	0.890	0.655	0.740	0.961	1.302

From the Table 1, we can find that: (1) Both SHR and FPM can obtain the right results. That is to say, S_1 and D_4 have the homology, S_2 and D_1 have the homology, and S_3 and D_2 have the homology. (2) In the same experimental environment, processing the same data set, SHR is better than FPM in term of the computing time.

Summary

DNA sequence homology recognition is a key problem in bioinformatics. We solve this problem by use of probability method instead of traditional sequence alignment because the character set of DNA sequence satisfies the Markov properties. In this paper, we first present the relative concepts and definitions. And then second order Markov model is used to describe characteristic of DNA sequence. We also define similarity measurement by utilizing two-step transition probability so as to realize DNA sequence homology recognition. Our DNA sequence homology recognition algorithm SHR is put forward after an example. In succession, the contrast experiments are done between SHR and FPM. The results show that SHR algorithm can correctly recognize the DNA sequence homology at an even faster processing speed.

References

[1] Junyan Zhang, Chenhui Yang. DNA Sequence Recognition Based on the Markov Model. Proceedings of the 6th International Conference on BioMedical Engineering and Informatics, 2013, p. 538-542

[2] Thompson D., Higgins J. Improving the sensitivity of progressive multiple sequence alignment through weight matrix choices. Nuclei Acids Research. Vol. 22(2011), p. 73-80.

[3] Randic M. Graphical representations of DNA as 2D map. Chem. Phys. Vol. 386(2013), p. 468-437.

[4] Voss R. Evolution of long range fractal correlation and 1/f noise in DNA base sequences. Phys Rev Lett, Vol. 25(2012), p. 35-36.

[5] Li M., Chen X. An information-based sequence distance and its application to whole mitochondrial genome phylogeny. Bioinformatics, Vol. 17(2011), p. 149-154.

[6] Haubold B., Pfaffelhuber P. Estimating mutual distances from unaligned genomes. J Compt Biol, Vol. 16(2010), p. 87-100.

[7] W. Y. Liu, L. J. Liu, C. W. Wang. Accessible miRNA target gene prediction based on second-order Markov model. Vol. 36(2012), p. 334-338.

[8] J. Y. Zhang and F. Min. Frequent Patterns Mining with Inflexible Wildcard Gaps. Proceedings of IEEE International Conference on Oxide Materials for Electronic Engineering, 2012, p. 539–543.