MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **66** (2011) 415-424

ISSN 0340 - 6253

A Discrete Measure for Phylogenetic Construction Based on Information Gain

Xinyuan Zhou[1,2*],Bo Liao[2*], Lijiao Liao[2], Xinguo Lu[2]

[1] *Department of Computer Science and Technology, Changsha University Changsha Hunan, 410003, China*

[2] *School of information science and technology, Hunan University, Changsha Hunan, 410082, China*

(Received September 20, 2010)

Abstract: Traditional methods of measuring the sequence distances require an alignment, which makes some subjective factors destroyed the original state of whole genome sequences. So this leads to constructing a poorly phylogenetic tree. This paper presents a new discrete measure based on information gain for phylogenetic construction, which works on sequences using the information gain and doesn't need aligning sequences to measure their distances and does not have subjective factors to interfere. Distance matrix of 10 mammals' whole mitochondrial genomes sequences is computed by this new measure. Compared with the proposed measures, the method of constructing phylogenetic trees based on new measure is feasible.

1 Introduction

It is an important topic in bioinformatics to study evolution relationship between different species, where the distance methods are the most common methods of constructing phylogeny. The sequence distances measure is roughly divided into two categories: alignment methods and non-alignment methods [1]. When the researchers use a larger set of species information (such as whole genome sequences) not

* Corresponding author. E-mail address: dragonbw@163.com(B. Liao)
* xyzhou@yeah.net (X.Zhou)

homologous sequences to study evolution and classification of species, multiple sequences alignment is very difficult. Because the base number of the whole genome sequences often reaches to million bp even billion bp, and genetic recombination is widespread in the whole genome sequences of different species. Thus, the users need to set parameters, penalty, and space inserting, so this interferes the subjective factors, destroys the original state of those data, and leads to bad results.

Therefore, some scholars have put forward some methods of calculating the sequence similarity between species without multiple sequence alignment. Those methods are known as non-alignment methods, which firstly put the DNA sequence into an object analyzed and processed by mathematical tools such as the existing linear algebra, the statistical theory, information theory and so on, then use the definition count vectors, the frequency vectors and others to analyze similarity or dissimilarity between vectors. Different measure methods build different similarity distance between sequences. At present, there have been some measure methods such as Eucidean distance [2], Angle distance [3], Kullback-Leiber entropy [4], Cross entropy [5], and FDOD (Fuction of Degree of Disa-greement) [6] based on Shannon entropy theory [7] and so on[8].

This paper presents a new discrete measure based on information gain for phylogenetic construction, which works on sequences using the information gain, doesn't need aligning sequences to measure their distances and does not have subjective factors to interfere. Because most measure methods are not very successful in the comparative analysis of the long sequences [9], so we select 10 mammals' whole mitochondrial genomes sequences as the experimental data, using our method to analysis similarity between species and construct their phylogenetic tree. Compared with the proposed measures, the method of constructing phylogenetic trees based on new measure is feasible.

2 Methods

2.1 sequences encoding

Similar with Yu's method [10], we also consider strings with fixed length K, called K-strings. There are a total of N= 4^K for DNA sequences possible types of K-strings. Assume the length of a DNA sequence is L. We use a window of length K and slide it through the sequences by shifting one position at a time to determine the frequencies of each of the N types of K-strings in this sequence. The observed frequency p ($\alpha_1\alpha_2...\alpha_K$) of a K-string $\alpha_1\alpha_2...\alpha_k$ is defined as $p(\alpha_1\alpha_2...\alpha_K)=f(\alpha_1\alpha_2...\alpha_K)/(L-K+1)$,where $f(\alpha_1\alpha_2...\alpha_K)$ is the number of times that $\alpha_1\alpha_2...\alpha_K$ appears in this sequence, and each α_i is one of the four nucleotides single-letter symbols. 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".

For all possible K-strings $\alpha_1\alpha_2...\alpha_K$, we use p ($\alpha_1\alpha_2...\alpha_K$) as components to form a composition vector for a genome. To further simplify the notation, we use P_i for i-th component corresponding to the string type i, i=1... N (the N strings are arranged in a fixed order as the alphabetical order). Hence we construct a composition vector P= ($P_1,P_2,...P_N$) for a genome.

2.2 A new discrete measure based on information gain

Information gain is an important concept in Shannon information theory [7], which has been widely used in machine learning and data mining areas. In the famous learning algorithm of decision trees such as ID3[11], C4.5[12], Quinlan has separately used the information gain and the information gain ratio as the choice standard of the node splitting property, which can quickly and accurately establish the corresponding decision tree to the sample data. In short, information gain is used to measure properties for distinguishing the ability of training the data sample.

In this paper, we propose a new discrete measure based on information gain for

the phylogeny construction, which is used information gain to measure the distance between sequences, and construct the phylogenetic tree. The measurement process of the similarity between sequences is as follow: firstly, we put the DNA sequences into objects such as above defined count vectors, the frequency vectors and so on, which are analyzed and processed by mathematical tools such as the existing linear algebra, the statistical theory, information theory and so on. Then we use information discrete measure to calculate the similarity or dissimilarity between vectors. A fundamental point of this idea is that the similar sequences have the similar field in common, in a way.

This paper proposes a new discrete measure based on information gain as follows:

Giving a composition vector P= $(P_1,P_2,...P_N)$ for a genome, where N is 4^K, we can get expectation information of a genome:

$$I(A)=-\sum_{i=1}^{4^K}(P_i^A)\log_2(P_i^A), \text{ Where A is a genome} \quad (1)$$

We calculate the expectation information of all species genomes. Assuming any two species genomes respectively are A species and B species, we can get the condition entropy of A and B:

$$E(A,B)=\sum_{i=1}^{4^K}(\frac{P_i^A}{P_i^B}\log_2(\frac{P_i^A}{P_i^B})+\frac{P_i^B}{P_i^A}\log_2(\frac{P_i^B}{P_i^A})) \quad (2)$$

At last, we define the information gain of A and B is as follow:

$$IG(A,B)=|I(A)+I(B)-E(A,B)| \quad (3)$$

The information gain of A and B reflects the similarity of A and B, The smaller IG (A, B), the higher similarity of A and B. In addition, in order to normalize the information gain of A and B, we can use formula (4):

$$SU(A,B)=\left[\frac{I(A)+I(B)}{IG(A,B)}\right] \quad (4)$$

We can know that the larger SU (A, B), the higher similarity of A and B by formula (4).

3 Experiments and Analysis

3.1 Experimental data

From the molecular level, we analyze the mammals' phylogeny, which is a controversial issue in molecular systematic. In this paper, we select 10 mammals' whole mitochondrial genome sequences as the experimental data, which are divided into four categories: primates, rodents, ferungulates and non-placental. All the data comes from the Genbank database of NCBI (http: //www.ncbi. nlm.nih.gov/). Species name and serial number are as shown in table 1**:**

Table 1: The complete mitochondrial genome sequences of 10 mammals

No	Species Scientific Name	abbreviation	Accession	category	Length(nt)
1	Homo Sapiens	human	V00662	Primates	16569
2	Pan Troglodytes	chimpanzee	D38116	Primates	16563
3	Macaca Mulatta	monkey	AY612638	Primates	16564
4	Mus Musculus	mouse	V00711	Rodents	16295
5	Rattus Norvegicus	rat	X14848	Rodents	16300
6	Canis Lupus Familiaris	dog	U96639	Ferungulates	16727
7	Equus Caballus	horse	X79547	Ferungulates	16554
8	Bos Taurus	cow	V00654	Ferungulates	16338
9	Monodelphis Domestica	opossum	AJ508398	Non-placental	17079
10	Ornithorhynchus Anatinus	platypus	X83427	Non-placental	17019

3.2 Research ideas and results

Firstly, we calculate frequency vectors of each sequence by 2.1 sections and get frequency vectors of each sequence $P= (P_1^X ,P_2^X ,...,P_N^X)$, where X is a sequence, and N is 4^7. In this paper, we select K=7. Then we use a new discrete measure based on information gain to calculate each distance between sequences, get the following distance matrix, which are as shown in table 2, and get the normalized distance matrix using formula 4, which are as shown in table 3 .The smaller the distance, the higher similarity between sequences. At last, we use the vertical and horizontal method [13]

to construct the phylogenetic tree by getting the distance matrix, and get the phylogenetic tree to be shown in figure 1.

Table 2: The familiarity matrix based on K=7 and by the information gain between sequences' frequency vectors

species	human	chimpanzee	monkey	mouse	rat	dog	horse	cow	opossum	platypus
human	0.0000	3652.0743	5375.0317	6640.2926	6341.7778	6664.7750	6091.8138	6304.9589	7835.7856	7643.5088
chimpanzee	3652.0743	0.0000	5513.7112	6790.4209	6119.6036	6791.0757	5980.5854	6161.6261	7766.1585	7478.5720
monkey	5375.0317	5513.7112	0.0000	6786.3086	6382.4791	6666.3865	6591.7408	6539.5533	8125.5850	8050.7383
mouse	6640.2926	6790.4209	6786.3086	0.0000	5147.2600	6333.3414	6342.6551	5618.1167	6214.7678	6743.4895
rat	6341.7778	6119.6036	6382.4791	5147.2600	0.0000	6297.8412	5813.2500	5737.2376	6452.9890	6805.6801
dog	6664.7750	6791.0757	6666.3865	6333.3414	6297.8412	0.0000	6256.7246	5769.6010	6534.9052	6572.4853
horse	6091.8138	5980.5854	6591.7408	6342.6551	5813.2500	6256.7246	0.0000	5985.6399	8052.0729	8002.7007
cow	6304.9589	6161.6261	6539.5533	5618.1167	5737.2376	5769.6010	5985.6399	0.0000	6455.6033	6636.2713
opossum	7835.7856	7766.1585	8125.5850	6214.7678	6452.9890	6534.9052	8052.0729	6455.6033	0.0000	6546.6711
platypus	7643.5088	7478.5720	8050.7383	6743.4895	6805.6801	6572.4853	8002.7007	6636.2713	6546.6711	0.0000

Table 3: The normalized familiarity matrix of table 2 using formula 4

species	human	chimpanzee	monkey	mouse	rat	dog	horse	cow	opossum	platypus
human	0.000000	0.995225	0.996750	0.997376	0.997250	0.997370	0.997130	0.997223	0.997777	0.997712
chimpanzee	0.995225	0.000000	0.996833	0.997435	0.997152	0.997420	0.997079	0.997160	0.997759	0.997663
monkey	0.996750	0.996833	0.000000	0.997433	0.997268	0.997371	0.997348	0.997323	0.997857	0.997828
mouse	0.997376	0.997435	0.997433	0.000000	0.996626	0.997242	0.997253	0.996895	0.997209	0.997416
rat	0.997250	0.997152	0.997268	0.996626	0.000000	0.997224	0.997000	0.996956	0.997309	0.997437
dog	0.997	0.9974	0.997	0.997	0.997	0.000	0.997	0.996	0.997	0.997

	370	20	371	242	224	000	199	958	330	333
horse	0.997 130	0.9970 79	0.997 348	0.997 253	0.997 000	0.997 199	0.000 000	0.997 075	0.997 837	0.997 814
cow	0.997 223	0.9971 60	0.997 323	0.996 895	0.996 956	0.996 958	0.997 075	0.000 000	0.997 299	0.997 361
opossu m	0.997 777	0.9977 59	0.997 857	0.997 209	0.997 309	0.997 330	0.997 837	0.997 299	0.000 000	0.997 341
platypu s	0.997 712	0.9976 63	0.997 828	0.997 416	0.997 437	0.997 333	0.997 814	0.997 361	0.997 341	0.000 000

The Constructed phylogenetic tree by this new method compares with Sims et al [14] which is shown in figure 2. They are very familiar and all separated out four categories: primates, rodents, ferungulates and non-placental. This result proves that new measure is feasible and valid. But, we carefully observe two trees; it is not difficult to discover that the branch situation of each item actually has the big difference. Constructed tree by new method agrees with the closer genetic relationship of Ferungulates and Rodents, but Constructed tree by Sims et al agrees with the closer genetic relationship of Primates and Rodents. Researchers [15, 16] analyze the mammals' phylogeny, which is a controversial issue in molecular systematic.

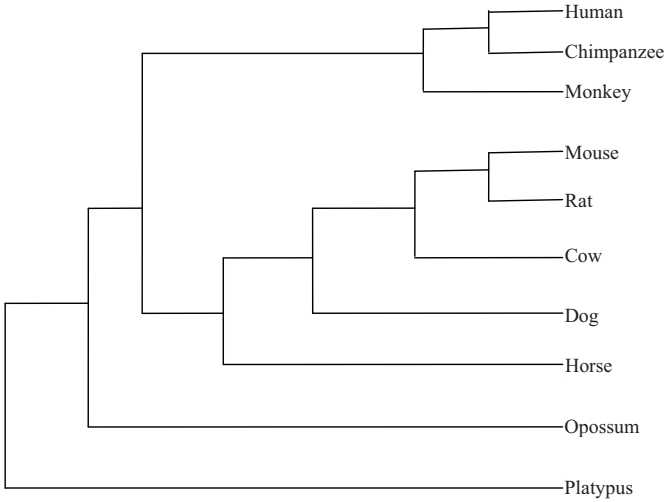

Figure 1: Constructed by new method

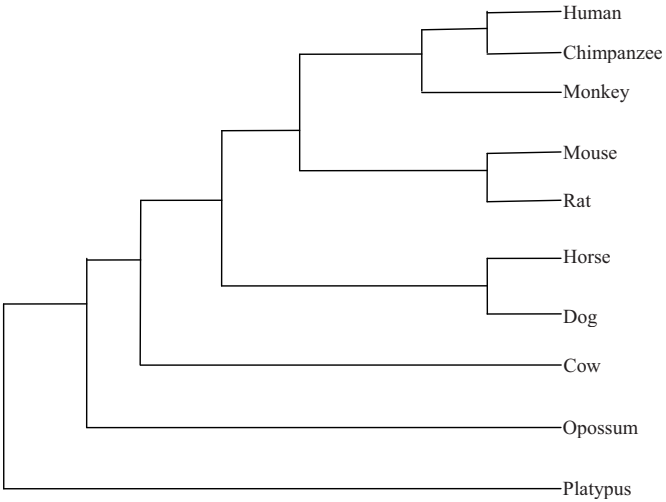

Figure 2: Constructed by Sims

Conclusion

In this paper, we propose a new discrete measure based on information gain for phylogenetic construction, which builds distance matrix between sequences using the information gain and doesn't need aligning sequences to measure their distances. New method analyses the similarity of sequences, which puts molecule sequences into mathematics implement of the information theory without the subjective interference factor. In fact, it simply uses the original genetic information automatic analysis. This paper select 10 mammals' whole mitochondrial genome sequences as the experimental data, the experiment results show that the phylogenetic tree by new measure is feasibility and validity. Although analyzing the mammals' phylogeny, is a controversial issue in molecular systematic, the new proposed method will provide a favorable method for study the difference of biological sequences.

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 Science and Technology Planning Project of Hunan Province (Grant 2009FJ3195) and supported by China Postdoctoral Science Foundation(Grant 20100471790).

References

[1] S. Vinga, J. Almeida, Alignment-free sequence comparison a review, Bioinformatics **9** (2003) 513–523.

[2] B. Blaisdell, A measure of the similarity of sets of sequences not requiring sequence alignment, Proc. Nail. Acad. Sci. **83** (1986) 5155–5159.

[3] G. Stuart, K. Moffer, J. Leader, A comprehensive vertebrate phylogeny using vector Representations of protein sequences from whole genomes, Mol. Biol. Evol. **19** (2002) 554–562.

[4] T. Wu, Y. Hsieh, L. Li, Statistical measure of DNA sequence dissimilarity under Markov chain models of base composition, Biometrics **57** (2001) 441–443.

[5] S. Kulback, *Information Theory and Statistics*, Wiley, New York, 1959.

[6] W. Fang, Disagreement degree of multi-person judgments in an additive structure, Math. Soc. Sci. **28** (1994) 85–111.

[7] C. Shannon, A mathematical theory of communication, Bell Sys. Techn. **27** (1948) 379–423.

[8] W. Zhu, B.Liao, R. Li, A novel method for constructing phylogenetic tree based on a dissimilarity matrix, MATCH Commun. Math. Comput. Chem. **63** (2010) 483–492.

[9] A. Hariri, B. Weber, J. Olmsted, On the validity of Shannon-information calculations for molecular biological sequences, J. Theo. Bio. **147** (1990) 235–254.

[10] Z. Yu, L. Zhou, V. Anh, K. Chu, C. Li, Y. Chen, Distance-based analyses to reveal vertebrate phylogeny without sequence alignment using complete mitochondrial genomes. Proceedings 11th World Multi-Conference on Systemics, Cybernetics and Informatics, Florida, USA, 2007, pp. 206–211.

[11] J. Quinlan, Induction of decision trees, Machine Larning **1** (1986) 81–106.

[12] J. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufman Publishers, San Mateo, 1993.

[13] B. Liao, L. Liao, G. Yue, R. Wu, W. Zhu, A vertical and horizontal method for constructing phylogenetic tree, MATCH Commun. Math. Comput. Chem. **63** (2010) 691–700.

[14] G. Sims, S. Jun, G. Wu, S. Kima, Whole-genome phylogeny of mammals: evolutionary information in genic and nongenic regions, PNAS **106** (2009) 17077–17082.

[15] S. Nikolaev, J. Burgos, E. Margulies, J. Rougemont, B. Nyffeler, S. Antonarakis, Early history of mammals is elucidated with the encode multiple species sequencing data, Plos Genetics **3** (2007) e2.

[16] A. Prasad, M. Allard, Confirming the phylogeny of mammals by use of large comparative sequence data sets, *Mol. Biol. Evol*. **25** (2008) 1795–1808.