MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **58** (2007) 559-568

ISSN 0340 - 6253

Comparison for DNA Primary Sequence

Bao-Hua Zhang, Hai-Shui Wang, Lu Xu*

Changchun Institute of Applied Chemistry, Chinese Academy of Sciences, Changchun 130022, Jilin, China

(Received March 19, 2007)

Abstract: In this article, two schemes are suggested based on three exons of β-globin gene belonging to 10 species for comparison of DNA primary sequences. At first, the positions of four nucleic acid bases were extracted, and then based on the information, as the numerical characterization of DNA sequences, the sequence invariants were derived. Sequences comparisons of 10 species selected in this work by using these invariants were performed. The results, especially with scheme 2, are quite satisfactory.

1. Introduction

Deoxyribonucleic acid (DNA) is a long polymer of nucleotides. It is responsible for the genetic propagation of most inherited traits, so comparison of primary sequences of different DNA strands remains one of the important aspects of analysis of DNA data.

There are many species in the world, the differences between their biological traits are enormous, but the differences between their DNA sequences are not enormous as we imagined, so it is difficult to describe effectively the different DNA sequences.

*Corresponding author. Fax:86-431-85685653

E-mail: luxu@ciac.jl.cn

Tel: 86-431-85262239

Postal address: 5625 Renmin Street,Changchun,China.

Postcode: 130022

At present, many schemes such as methods based on matrix [1~2], and methods related to graphs [3~7] can be used for the comparison of DNA primary sequences.

Formerly, most of the comparisons are based on the first exon of β-globin gene[8~14]. The first exon of β-globin gene contains 86~94 bases, but there are millions of bases contained in the DNA sequence, so only one exon may be not to give us enough information about a DNA sequence. Thus, many of the results [1, 3] in the literatures did not agree very well with the phylogenetic tree [15].

In this article, we also follow the rule: important information is contained in the exons. However, our studies are based on three exons of β-globin gene belonging to the 10 species. The suitable invariants were extracted, and the sequence comparisons were made among the 10 species.

1.1 The selected species

We selected 10 species (shown in table1) from phylogenetic tree (figure 1).

Table1. Coding domain sequence (CDS) of ten species

	Exon1		Exon2		Exon3	
species	bases	Region	bases	Region	bases	Region
bovine	86	278-363	223	492-714	129	1613-1741
goat	86	279-364	223	493-715	129	1621-1749
pig	92	871-962	223	1080-1302	129	1944-2072
rabbit	92	480-571	223	698-920	129	1494-1622
rat	92	310-401	223	517-739	129	1377-1505
mouse	92	2718-2809	223	2926-3148	129	3802-3930
gallus	92	465-556	223	649-871	129	1682-1810
geochelone	89	1-89	223	190-412	126	513-638
chimpanzee	105	4189-4293	222	4412-4633	49	5484-5532
gorilla	93	4538-4630	222	4761-4982	49	5833-5881

For facilities to observe, the 10 species are grouped roughly into four classes. The first class includes bovine, goat, rabbit, pig; the second class includes mouse, rat; the third class includes gallus and geochelone(Tortoise), the fourth class includes chimpanzee and gorilla.

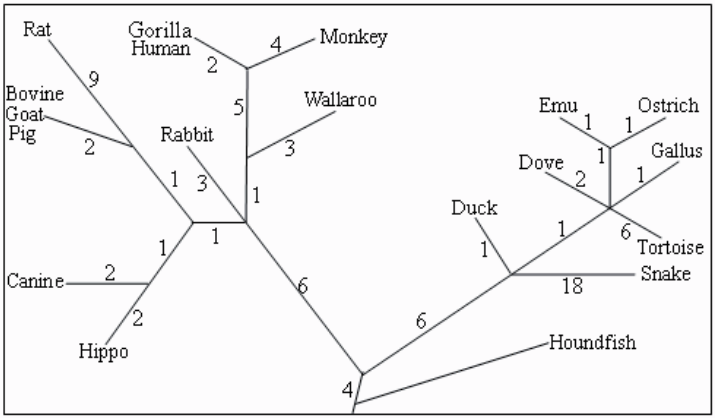

Figure 1.Phylogenetic tree (part)

2. Theory and experimental results

In this article, we proposed two coding schemes.

2.1 Coding scheme 1

（1） Sequence invariants $^{m}H_{t}$.

At first, we extract invariants of 4 kinds of bases from three exons of β-globin genes. The approach can be illustrated by using a portion of DNA sequence: atggtgcacctgactcctgaggagaagtctgcc. In this segment to consider the position of adenine leads to the following numerical sequence: 1，8，13，20，23，25 , 26. Similarly, the numerical sequences can be got for the remaining nucleic acids, i.e. guanine (g)：3，4，6，12，19，21，22，24，27，31；cytosine(c)：7，9，10，14，16，17，32，33；thymine (t)：2，5，11，15，18，28，30.

Then, the elements of such sequences are divided by the total number of the bases in a DNA sequence, thus the relative position sequence of a base can be obtained. For example, the sequence for adenine is: 1/33, 8/33, 13/33, 20/33….

Obviously, the members included in a relative position sequence for different kinds of bases are generally not the same, thus, it can not be used directly to DNA sequence comparison. For this question, molecular connectivity index [16] was applied. The power of equation used in this article is 0.5. It differs from molecular connectivity

index, -0.5. So, we use ${}^{m}H_t$ to represent our index.

$$ {}^{m}H_t = \sum (\delta_1.\delta_2...\delta_k)^{0.5} $$

In this equation, δ_k represents the members included in a relative position sequence, so ${}^{m}H_t$ of adenine is:

$$ {}^{1}H_p = (1/33x8/33)^{0.5} + (8/33x13/33)^{0.5} + (13/33x20/33)^{0.5} + \cdots\cdots $$

The similar indices (${}^{1}H_p \sim {}^{n}H_p$) can be got for the remaining nucleic acids g, c, t. In this article, the indices of ${}^{1}H_p \sim {}^{5}H_p$ were calculated as invariants for DNA sequence comparison.

（2） Invariants *Z*

According to step (1), we get 20 invariants belonging to each exon, then, we make quotients between indices of exon 1 and exon 2, as well as exon 2 and exon 3, thus, we got 40 invariants. These are invariants *Z*, proposed in this paper.

2.2 Coding scheme 2

(1) Graphical representation of DNA sequences

In this part, we consider graphical representation of DNA primary sequences. At first we transform a sequence into a graph, then, extract invariants from this graph. The approach is illustrated on coding domain sequence of goat β-globin gene. The coding domain sequence(CDS) of goat β-globin gene is:

atgctgactgctgaggagaaggctgccgtcaccggcttctggggcaaggtgaaagtggatgaagttg
gtgctgaggccctgggcaggctgctggttgtctacccctggactcagaggttctttgagcactttgggg
acttgtcctctgctgatgctgttatgaacaatgctaaggtgaaggcccatggcaagaaggtgctagactc
ctttagtaacggcatgaagcatcttgacgacctcaagggcacctttgctcagctgagtgagctgcactgt
gataagctgcacgtggatcctgagaacttcaagctcctgggcaacgtgctggtggttgtgctggctcgc
caccatggcagtgaattcaccccgctgctgcaggctgagtttcagaaggtggtggctggtgttgccaat
gccctggcccacagatatcactaa

We assign to each nucleic acid base two number sequences [17]: one numerical sequence is the position of a base in the DNA sequence, we use ***m*** to represent it; the

other is the position of a base in the subsequence of nucleic acid bases of the same kind, we use ***n*** to represent it. Then, the CDS sequence mentioned above corresponding to adenine(a) leads to the following two numerical sequences: ***m***=[1，7，14，17，19，20，31，59，.....434，437，438]; ***n***=[1，2，3，4，5，6，7，…88，89].

Based on ***m, n,*** i.e., to take ***n*** as the x-coordinate, ***m*** as y-coordinate, we can drive a curve in 2-D plane (see Figure 2).

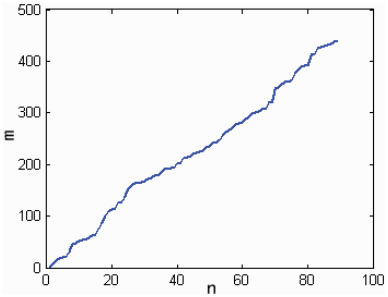

Figure 2 Graphical representation of DNA sequence

Similarly, graphic representation can be constructed for the remaining nucleic acids, guanine (g), cytosine(c), thymine (t).

(2) Extracting invariant from the graph

We also take base adenine in the domain sequence mentioned above as an example to explain the extraction of the sequence invariants. From 0 to 89, we equally take k points from the x-coordinate, that the values of y-coordinate corresponding to the points of x-coordinate are the invariants of base adenine. The magnitude of k is arbitrary, but it should be large enough. Because of only when the k value is large enough, the DNA sequence can be described sufficiently.

(3) Proportional characterization of exons

According to step (2) above, we get （4 x k） invariants for each exon, then, we make quotients between invariants of exon1 and exon2, as well as exon2 and exon3. thus, we got （8 x k）invariants. Using such invariants, comparisons were made for coding domain sequence of the 10 species.

3. Similarity comparison

The analysis of similarities/dissimilarities among the DNA sequences was performed by calculating the Euclidean distance between two species. Clearly, the smaller is the Euclidean distance the more similar are the two DNA sequences.

3.1 Results with schemes 1.

Table 2 is the result of sequences comparison based on the invariant Z. The smallest entries in the rows of the matrix are shown in bolt letters.

Table 2 Similarity/dissimilarity matrix for the 10 species

	bovine	goat	pig	rabbit	rat	mouse	gallus	geochelone	chimpanzee	gorilla
bovine	0	**0.31353**	1.2667	0.36201	0.62711	0.38675	1.3952	3.6672	30.512	31.993
goat	**0.31353**	0	1.2858	0.45854	0.7019	0.51438	1.2876	3.41	30.393	31.871
pig	1.2667	1.2858	0	**1.0825**	1.8056	1.532	2.1851	3.7939	30.503	31.984
rabbit	**0.36201**	0.45854	1.0825	0	0.79965	0.5542	1.5147	3.6203	30.44	31.925
rat	0.62711	0.7019	1.8056	0.79965	0	**0.34407**	1.3536	3.7824	30.438	31.93
mouse	0.38675	0.51438	1.532	0.5542	**0.34407**	0	1.3907	3.7564	30.444	31.934
gallus	1.3952	**1.2876**	2.1851	1.5147	1.3536	1.3907	0	2.8638	29.852	31.276
geochelone	3.6672	3.41	3.7939	3.6203	3.7824	3.7564	**2.8638**	0	28.51	29.871
chimpanzee	30.512	30.393	30.503	30.44	30.438	30.444	29.852	28.51	0	**2.4275**
gorilla	31.993	31.871	31.984	31.925	31.93	31.934	31.276	29.871	**2.4275**	0

From Table 2, we can see that the most similarity pairs are:

Row 1(bovine, goat) row 6 (mouse, rat)

Row 2(goat, bovine) row7 (gallus, goat)

Row3 (pig, rabbit) row8 (geochelone, gallus)

Row4 (rabbit, bovine) row9 (chimpanzee, gorilla)

Row5 (rat, mouse) row10 (gorilla, chimpanzee)

Obviously, most of the similarity pairs agree with the phylogenetic tree. There are two exceptional entries in this table: (1) in row 3, the smallest entry corresponding to rabbit and pig, moreover, the phylogenetic tree shows that goat and bovine are the most similar species to pig; (2) in row 7, the smallest entry shows that rabbit is most

similar to gallus, however, phylogenetic tree shows that geochelone is the most similar to gallus. This means that it is necessary to further improve our approach. **3.2 Results with scheme 2**

In this case, let k=300, the results are shown in Table 3.

Table3 Similarity/dissimilarity matrix for the 10 species

	bovine	goat	pig	rabbit	rat	mouse	gallus	geochelone	chimpanzee	gorilla
bovine	0	**6.671**	8.0517	6.882	10.263	9.4482	13.736	13.071	108.93	109.05
goat	**6.671**	0	7.5542	10.465	10.712	8.4406	14.108	12.667	108.50	109.14
pig	8.0517	**7.5542**	0	8.5538	11.506	9.6565	14.52	14.267	108.04	108.55
rabbit	**6.882**	10.465	8.5538	0	10.715	10.415	13.773	15.34	108.81	108.70
rat	10.263	10.712	11.506	10.715	0	**5.5797**	13.001	14.103	112.04	112.18
mouse	9.4482	8.4406	9.6565	10.415	**5.5797**	0	11.93	12.295	110.86	111.21
gallus	13.736	14.108	14.52	13.773	13.001	11.93	0	**10.453**	110.99	111.09
geochelone	13.071	12.667	14.267	15.34	14.103	12.295	**10.453**	0	112.10	112.46
chimpanzee	108.93	108.5	108.04	108.81	112.04	110.86	110.99	112.10	0	**13.504**
gorilla	109.05	109.14	108.55	108.7	112.18	111.21	111.09	112.46	**13.504**	0

Table 3 shows that the most similarity pairs are:

Row 1(bovine, goat) row 6 (mouse, rat)

Row 2(goat, bovine) row7 (gallus, geochelone)

Row3 (pig, goat) row8 (geochelone, gallus)

Row4 (rabbit, bovine) row9 (chimpanzee, gorilla)

Row5 (rat, mouse) row10 (gorilla, chimpanzee)

Obviously, the results are improved greatly. They agree with the phylogenetic tree very well.

Based on invariants of 10 species, the principal component analysis was performed. These10 species were projected onto 2-D plane shown in Figue3. The 10 species are separated clearly into three areas.

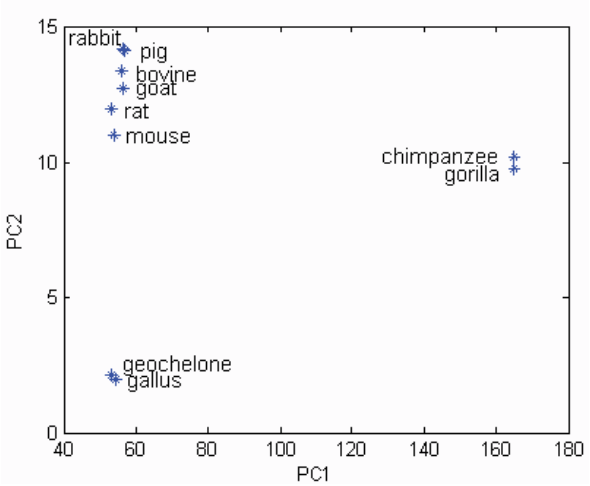

Figure 3 Projected graph of the 10 species with principal component analysis

4. Concluding remarks

All the invariants proposed in this article contain proportional relationships of three exons in the coding domains of the 10 species. In which, based on the scheme1, the results are basically agree with phylogenetic tree, but there are exceptions. Whereas, based on the scheme2, the results agree with phylogenetic tree very well. Therefore, the following conclusions could be given out:

(1) The primary sequence of a DNA looks to be simple, but it is difficult to describe it effectively. We would say that among many methods those that indicate great dissimilarity among two species (while other approaches may not show such big difference) are to be more trusted even if other methods do not show such difference. Converse, of course is not true: two species showing apparent similarity (having small difference in their selected invariants used for similarity analysis) need not be similar at all - but if they show great similarity within a number of different representations - they are likely to be similar.

(2) In appearance, schem2 also is a method of graph transformation, but it has no the disadvantages of the type approaches, such as those methods can not be used to the longer sequences; loss of information associated with repeating moves that overlap; the choice of axes for various bases is arbitrary and so on. Theoretically, the scheme in this research can be used for any lengthy sequences.

(3) It is expected that those invariants can be applied to similarity analysis of RNA

sequences.

REFERENCES

(1) Randić, M. Condensed Representation of DNA Primary Sequences. *J.Chem.Inf. Comput. Sci.* **2000**, 40, 50-56.

(2) Randić, M. On Characterization of DNA Primary Sequences by Condensed Matrix. *Chem. Phys. Lett.* **2000**, 317, 29-34.

(3) Nandy, A. A New Graphical Representation and Analysis of DNA Sequence Structure. I. Methodology and Application to Globin Genes. *Curr. Sci*. **1994**, 66, 309-314.

(4) Randić, M.; Balaban, A. T. On a Four-Dimensional Representation of DNA Primary Sequences. *J. Chem. Inf. Comput. Sci.* **2003**, 43, 532-539.

(5) Randić, M.; Vračko, M.; Lerš, N.; Plavšić, D. Novel 2-D graphical representation of DNA sequences and their numerical characterization. *Chem. Phys. Lett.* **2003**, 368, 1-6.

(6) Randić, M.; Vračko, M.; Lerš, N.; Plavšić, D. Analysis of Similarity/Dissimilarity of DNA Sequences Based on Novel 2-D Graphical Representation. *Chem. Phys. Lett.* **2003**, 371, 202-207.

(7) Zupan, J.; Randić, M. Algorithm for DNA Sequences into “Spectrum-like” and “Zigzag” Representations. *J. Chem. Inf. Comput. Sci.* **2005**, 45, 309-313.

(8) Randić, M.; Vračko, M. On Similarity of DNA Primary Sequences. *J.Chem.Inf. Comput. Sci.* **2000**, 40, 599-606.

(9) Randić, M.; Balaban, A. T.; Basak，S. C. On Structural Interpretation of Several Distance Related Topological Indices, *J. Chem. Inf. Comput. Sci.* **2001**, 41, 593-601.

(10) Randić, M.; Guo, X. F.; Basak, S. C. On the Characterization of DNA Primary Sequences by Triplet of Nucleic Acid Bases *J. Chem. Inf. Comput. Sci.* **2001**, 41, 619-626.

(11) Randić, M.; Vračko, M.; Nandy, A.; Basak, S. C. On 3-D Graphical Representation of DNA Primary Sequences and Their Numerical Characterization. *J. Chem. Inf. Comput. Sci.* **2001**, 41, 1235-1244.

(12) Guo, X. F.; Randić, M.; Basak, S. C. A Novel 2-D Graphical Representation of DNA Sequences of Low Degeneracy. *Chem. Phys. Lett*. **2002**, 350, 106-112.

(13) Randić, M.; Zupan, J. Highly Compact 2-D Graphical Representation of DNA Sequences. *SAR QSAR Environ Res*. **2004**, 15, 147-157.

(14) Randić, M. 2-D Graphical Representation of Proteins Based on Virtual Genetic Code. *SAR QSAR Environ.* Res. **2004**, 15, 191-205.

(15) Wang, J. Y.; Zhu, S. G,; Xu, C. F. (Eds.), Biochemistry, Chapter 4, High Education, Beijing, **2003**, p. 183.

(16) Randić, M.; Characterization of Molecular Branching. *J Am Chem Soc*. **1975**, 97, 6609-6615.

(17) Randić, M.; Basak, S. C. Characterization of DNA Primary Sequences Based on the Average Distances between Bases. *J. Chem. Inf. Comput. Sci.* **2001**, 41, 561-568.