Molecular Phylogenetics and Evolution 59 (2011) 438–443

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

A new distribution vector and its application in genome clustering

Bo Zhao [a], Rong L. He [b], Stephen S.-T. Yau [c,*]

[a] Department of Mathematics, Statistics, and Computer Science, University of Illinois at Chicago, Chicago, IL, USA
[b] Department of Biological Sciences, Chicago State University, Chicago, IL, USA
[c] Department of Mathematical Sciences, Tsinghua University, Beijing 100084, PR China

ARTICLE INFO

Article history:
Received 1 October 2010
Revised 20 January 2011
Accepted 28 February 2011
Available online 6 March 2011

Keywords:
Distribution vectors
Clustering
Phylogenetic tree

ABSTRACT

In this paper we report a novel mathematical method to transform the DNA sequences into the distribution vectors which correspond to points in the sixty dimensional space. Each component of the distribution vector represents the distribution of one kind of nucleotide in k segments of the DNA sequences. The mathematical and statistical properties of the distribution vectors are demonstrated and examined with huge datasets of human DNA sequences and random sequences. The determined expectation and standard deviation can make the mapping stable and practicable. Moreover, we apply the distribution vectors to the clustering of the Haemagglutinin (HA) gene of 60 H1N1 viruses from Human, Swine and Avian, the complete mitochondrial genomes from 80 placental mammals and the complete genomes from 50 bacteria. The 60 H1N1 viruses, 80 placental mammals and 50 bacteria are classified accurately and rapidly compared to the multiple sequence alignment methods. The results indicate that the distribution vectors can reveal the similarity and evolutionary relationship among homologous DNA sequences based on the distances between any two of these distribution vectors. The advantage of fast computation offers the distribution vectors the opportunity to deal with a huge amount of DNA sequences efficiently.

1. Introduction

With the exponential growth of DNA sequences in the past twenty years, it is ineffective to analyze DNA sequences only through the traditional biological experiments. Various mathematical methods and computer algorithms are applied to sequence analysis and related research areas, which help the biological study to be upgraded into automatic programming from manual operation. Especially, the multiple sequence alignment is used to construct the phylogenetic tree based on homologous sequences. Moreover, some alignment-free sequence comparison methods are also introduced to cluster homologous sequences. For example, Woese and Fox (1977) defined the Archaea (a new domain or kingdom of life) in 1977 by phylogenetic taxonomy of 16S ribosomal RNA. Recently, Yau et al. (2008) developed the moment vectors to cluster the protein sequences in 2008. Moreover, Takahashi et al. (2009) estimated the phylogenetic tree of bacterial species with oligonucleotide frequency distances. In this paper, we introduce the distribution vectors to map each DNA sequence into a point in a sixty dimensional space. We also study the minimum, maximum, expectation and standard deviation of the distribution vectors. The distribution vector method is applied to build the phylogenetic trees of the Haemagglutinin (HA) gene of 60 H1N1 viruses from Human, Swine and Avian, the mitochondrial complete genomes from 80 placental mammals and the complete genomes from 50 bacteria. All the three trees show the similarity among the sequences in the three datasets and correspond to the evolutionary relationship of the the 60 H1N1 viruses, 80 placental mammals and 50 bacteria respectively. Moreover, it take much less time to build the phylogenetic tree by our the distribution vector method than the popular multiple sequence alignment methods, such as ClustalW (Brown et al., 2007), Muscle (Edgar, 2004), MAFFT (Katoh et al., 2009) and MISHIMA (Kryukov et al., 2010).

2. Methods

In the beginning, we define the indicator sequence $u_\alpha(n)$ of the DNA sequence.

$$u_\alpha(n) = \begin{cases} 1, & \text{if } \alpha \text{ appears at location } n \text{ of the DNA sequence,} \\ 0, & \text{otherwise,} \end{cases} \tag{1}$$

$\alpha \in I = \{A,T,C,G\}$, $n = 1, 2, \ldots, N$ and N is the length of the DNA sequence.

To construct the distribution vectors, we fix k, which is a preset integer much less than N. Then we define q as the quotient and r as the remainder in Eq. (2) when dividing N by k.

[*] Corresponding author.
E-mail address: yau@uic.edu (S.S.-T. Yau).

1055-7903/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.ympev.2011.02.020

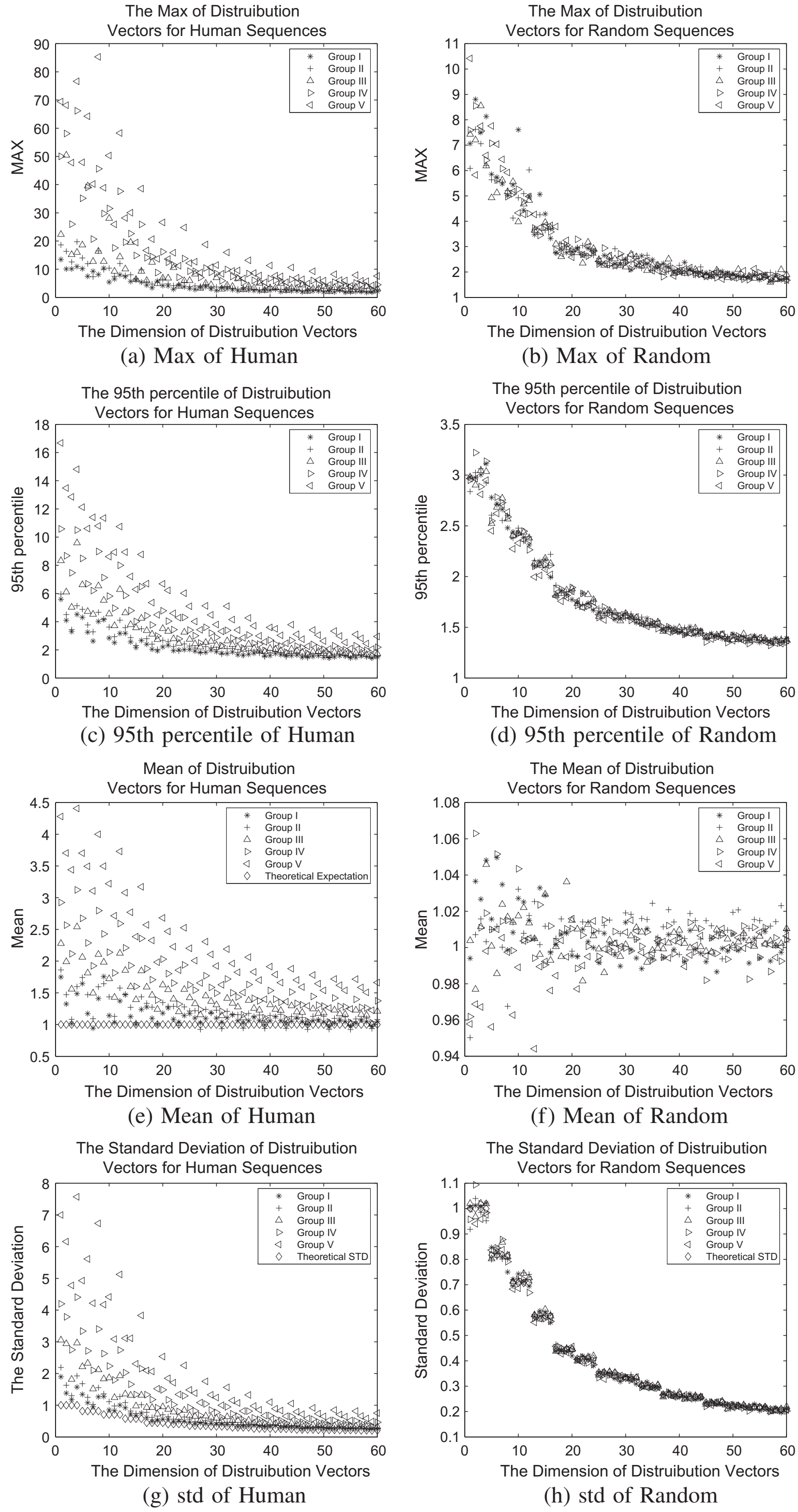
Table 1
The grouping of 5000 human DNA sequences.

	Number	Range of length
Group I	996	<384
Group II	1001	⩾384 and <651
Group III	999	⩾651 and <1053
Group IV	1500	⩾1053 and <2265
Group V	504	⩾2265

$$q = \left\lfloor \frac{N}{k} \right\rfloor, \quad r = N - k \times q \tag{2}$$

It is clear that $0 \leqslant r < k$. Therefore, we divide the DNA sequence into k segments with almost equal lengths: The first r segments possess $q + 1$ nucleotides and the remaining k–r segments hold q nucleotides. Eq. (3) explains the partition clearly.

$$N = k \times q + r = r(q + 1) + (k - r)q \tag{3}$$



(a) Max of Human (b) Max of Random

(c) 95th percentile of Human (d) 95th percentile of Random

(e) Mean of Human (f) Mean of Random

(g) std of Human (h) std of Random

Fig. 1. The experiment of distribution vectors.

Then we define $Q_\alpha(m,k)$ as the number of the nucleotides α in the mth segment of the DNA sequence in Eq. (4).

$$Q_\alpha(m,k) = \begin{cases} \sum\limits_{i=(m-1)q+m}^{m(q+1)} u_\alpha(i), & m = 1,1,2,\dots,r \\ \sum\limits_{i=(m-1)q+r+1}^{m\times q+r} u_\alpha(i), & m = r+1, r+2,\dots,k \end{cases} \tag{4}$$

For each k, we define the $DV_\alpha(k)$ in terms of $Q_\alpha(m,k)$ to describe the variability between any two of $Q_\alpha(m,k)$ for the particular nucleotide α in one DNA sequence.

$$DV_\alpha(k) = \frac{8}{3N(k-1)}\left(\sum_{\substack{i=1\\ i\neq j}}^{k}\sum_{j=1}^{k}(Q_\alpha(i,k) - Q_\alpha(j,k))^2\right) \tag{5}$$

The intention for choosing the coefficient $\frac{8}{3N(k-1)}$ is to simplify the expectation to be a constant. The explanation will be given later.

For each $k \in K = \{3,4,5,7,11,13,17,19,23,29,31,37,41,43,47\}$, we compute the $DV_A(k)$, $DV_C(k)$, $DV_G(k)$ and $DV_T(k)$ and put these together to obtain the sixty dimensional distribution vector $\overline{DV}$. It is clear there is no common factor except 1 among the numbers in the set K including 14 small odd prime numbers and 4, which makes the elements in the distribution vector more independent. We do not choose 2 because the definition of $DV_\alpha(2)$ is a little simple and the value $DV_\alpha(2)$ is more unstable than $DV_\alpha(4)$. The selection of the size of the set K is crucial. The distribution vectors can map the sequences more precisely when the size of K is large. On the other hand for the short sequences, each segment is too short to provide the information if the k is too large. In addition, the larger the size of the set K, the longer the computation time. All of the above reasons should be considered in the selection of the set K.

$$\begin{aligned}\overline{DV} = \{&DV_A(3), DV_C(3), DV_G(3), DV_T(3),\\ &DV_A(4), DV_C(4), DV_G(4), DV_T(4),\\ &\dots\\ &DV_A(47), DV_C(47), DV_G(47), DV_T(47)\}\end{aligned} \tag{6}$$

3. Mathematical and statistical properties

According the definition of $Q_\alpha(m,k)$ and $DV_\alpha(k)$, We derive the minimum, maximum, expectation and standard deviation of $DV_\alpha(k)$ when we consider the DNA sequence as a random sequence, which means every position in the DNA sequence can be A, C, G or T with the same probability $\frac{1}{4}$ independently. The proof is available in the Supplement.

$$Min(DV_\alpha(k)) = 0, \quad \text{if } Q_\alpha(1,k) = Q_\alpha(2,k) = \cdots = Q_\alpha(k,k) \tag{7}$$

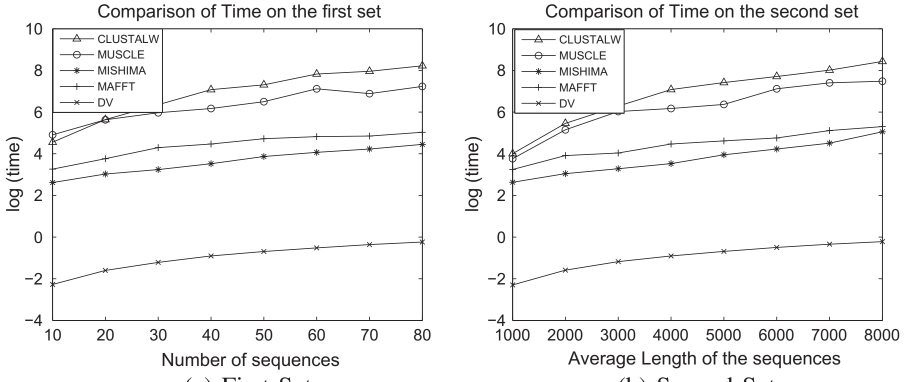
$$\begin{aligned} Max(DV_\alpha(k)) &= \frac{8}{3N(k-1)}\left(\sum_{\substack{i=0\\ i\neq j}}^{k-1}\sum_{j=0}^{k-1}(Q_\alpha(i,k) - Q_\alpha(j,k))^2\right)\\ &= \begin{cases} \frac{2N}{3(k-1)} & \text{if } k \text{ is even}\\ \frac{2N}{3k} & \text{if } k \text{ is odd}\end{cases}\end{aligned} \tag{8}$$

$$E[DV_\alpha(k)] = 1 \tag{9}$$

And

$$std[DV_\alpha(k)] = \sqrt{Var[DV_\alpha(k)]} \approx \sqrt{\frac{2}{k-1}} \tag{10}$$

Since the distribution of the four nucleotides for authentic DNA sequences is not same as the random DNA sequences. we examine these properties with two large datasets. One is 5000 Human DNA sequences from the NCBI database, which are divided into five groups by the respective lengths of DNA sequences. The detail of grouping is provided in Table 1. Another dataset is 5000 random DNA sequences which are divided into five groups also. Each group consists of 1000 random sequences each with a fixed length. The lengths of these groups are 200, 400, 800, 1500 and 3000 corresponding to the Group I, II, III, IV and V respectively. We compare the maximum, 95th percentile, mean and standard deviation for the ten groups in Fig. 1. The mean and standard deviation of random sequences are close to theoretical expectation and standard deviation, which do not depend on the length of sequences. The maximum of random sequences is much smaller than the theoretical maximum because the probability of a high value of f $DV_\alpha(k)$ is very small. On the other hand, the maximum of human sequences is bigger than the maximum of random sequences but still much smaller than the theoretical maximum. Moreover, the maximum of human sequences increases when the length of sequences increases, but the acceleration is much slower. We also locate the 95th percentile for the human sequences. Fig. 1c shows that 95% of $DV_\alpha(k)$ of human sequences is smaller than 18. Furthermore, The mean and standard deviation within the five human DNA sequence groups also converge to the theoretical expectation and standard deviation when we increase the dimension, even though the convergence is not as good as those of the random sequences. Therefore, the distribution vectors of authentic DNA sequences are also bounded and each



(a) First Set

(b) Second Set

Fig. 2. The time comparison of four methods.

component plays same important role when we compute the distance matrix and cluster the sequences.

4. Application

We apply the distribution vector method to three datasets. the first data set includes the Haemagglutinin(HA) gene of 60 H1N1 viruses from Influenza Virus Sequence Database. The average length of the gene is only around 1600 bp. We know Many viruses have short generation times and relatively high mutation rates, such as Influenza Virus. It is very useful to analyze the genes of these viruses to find the origin and transmission of the viruses, Especially, the research became crucial in the outbreak of Swine H1N1 in 2009. Secondly, We collect 80 mitochondrial complete genomes of placental mammals from NCBI database. The average length of the genomes is around 16,000 bp. It is useful for studying the evolutionary relationships based on Mitochondrial genome because it is inherited from the mother (maternally inherited) in most multicellular organisms and is not highly conserved and has a rapid mutation rate. Thirdly, we test ourthe distribution vector method on the complete genomes from 50 bacteria. Because the average length of the bacteria genome is around 4,000,000 bp, this application can verify the high efficiency of our method on the large dataset.

We calculate the distribution vectors of these sequences and the distances between any two of these distribution vectors for each datasets. The phylogenetic trees are built based on the distance matrix by using the function hclust from the R program (R Development Core Team, 2008), where the average linkage method is used in the clustering. The three trees are plotted in Figs. 3–5. Moreover, we apply the multiple alignment on the same three

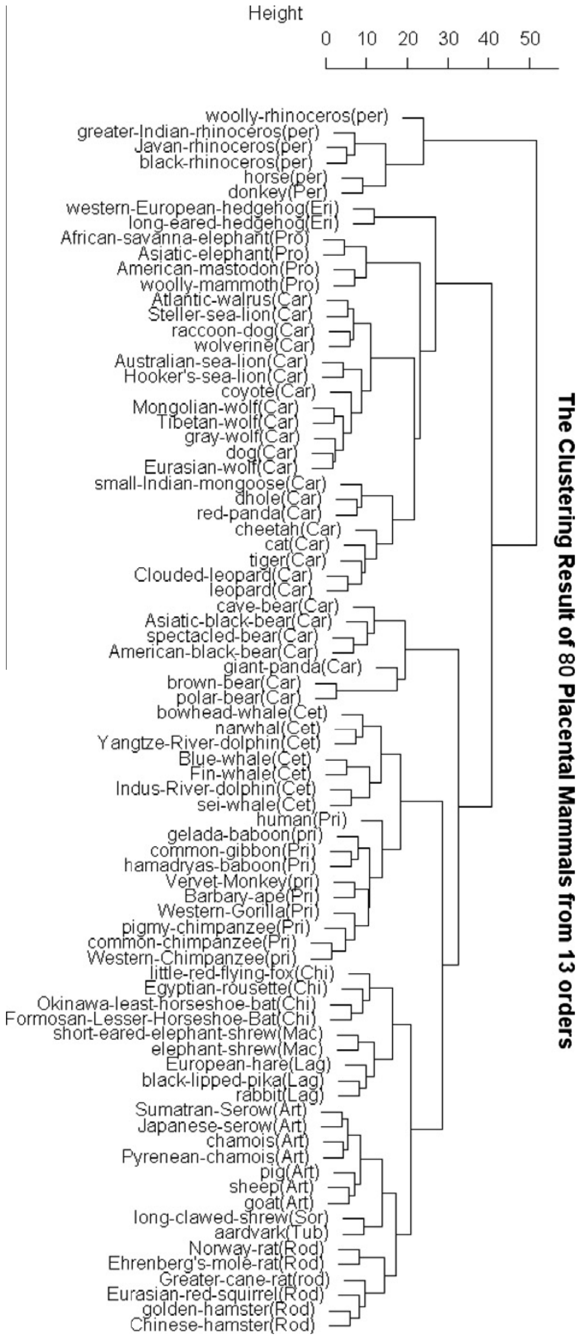


Fig. 3. The clustering result of 80 mitochondrial genomes.

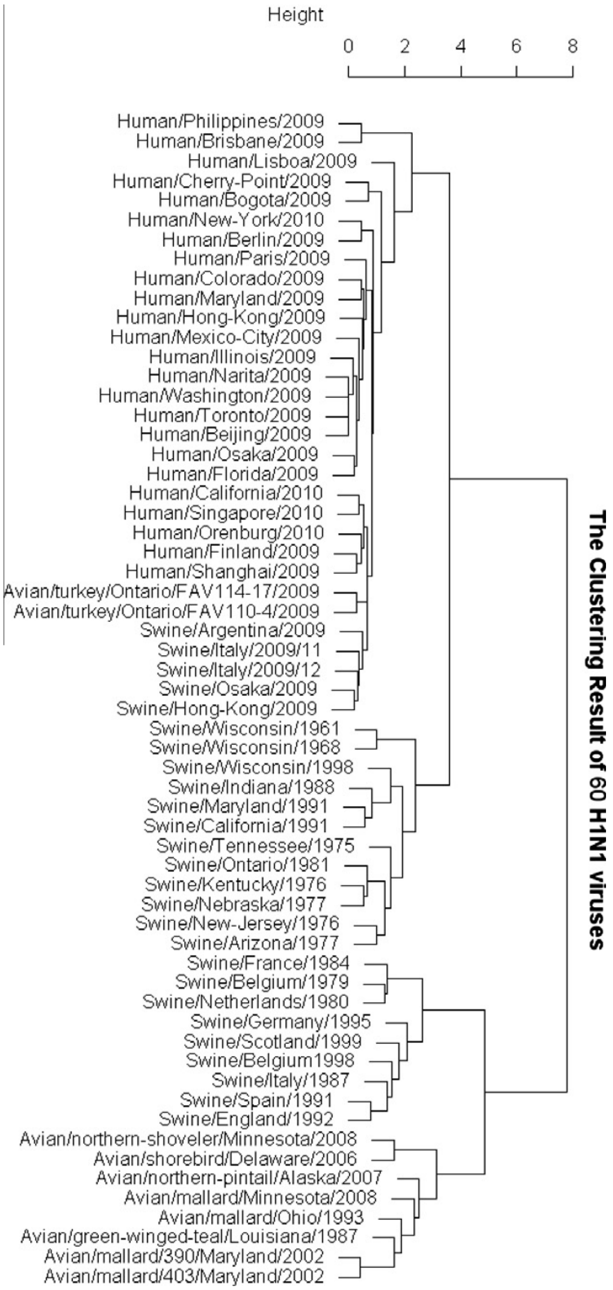


Fig. 4. The clustering result of 60 H1N1 viruses.

datasets with ClustalW2, Muscle, MAFFT and MISHIMA and do the clustering with the average linkage method also. The results are provided in the Supplement. For the dataset of 60 H1N1 (HA) viruses, the distribution vector method classifies these viruses into four groups correctly. The four groups include the avian older than 2009, European swine older than 2009, American swine older than 2009 and the new 2009 viruses from human, swine and avian. The result shows the 2009 human H1N1 viruses have closer relationship with old American swine than old avian and European swine. ClustalW2, Muscle, MAFFT and MISHIMA also classify the 60 H1N1 viruses into the four groups except that the virus swine/wisconsin/1961 is not classified well by ClustalW2, Muscle and MAFFT. Moreover, only the distribution vector method put the 2009 swine and 2009 avain together in the group of new 2009 viruses from human, swine and avian. Secondly, all the five methods classify most of the 80 animals correctly by the respective orders they belong. Ourthe distribution vector method divides the animals in the order of Carnivora into two groups: bears and non-bears, while other four methods make more errors with the order of Carnivora. Moreover, only the distribution vector method puts pig in to the order of Artiodactyla successfully. Thirdly, only the distribution vector method classify 50 bacteria correctly by the respective families they belong. However, the other four alignment methods are unable to process the 50 bacteria genomes in our personal computer (3G CPU and 2G memory). In general, all the five methods can do the clustering with the viruses, animals and bacteria corresponding to the evolution relationship. But the distribution vector method obtains the more accurate results in the clustering. Furthermore, the distribution vector method is much faster than the other methods. We record the time each method takes on each dataset and list them in Table 2.

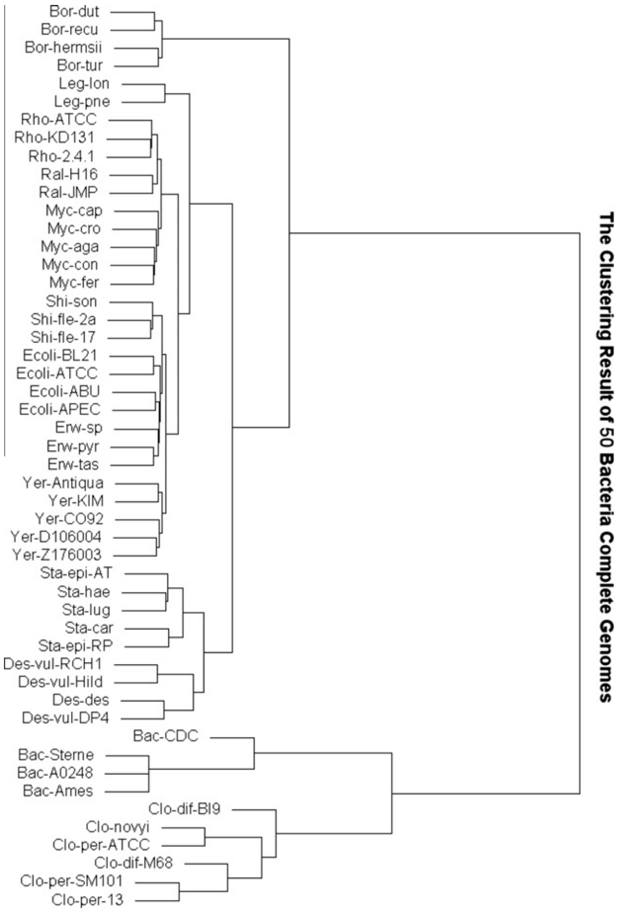


Fig. 5. The clustering result of 50 bacteria genomes.

Table 2
The clustering time (s).

	DV	MISHIMA	MAFFT	Muscle	ClustalW2
H1N1 viruses	<1	67	283	407	463
Mitochondrial	1	1128	2026	73,190	78,211
Bacteria	43	NA	NA	NA	NA

In order to compare the speed of our method and the other four methods generally, we do the test on two sets of sequences. The first set consists of eight datasets. The number of sequences in each dataset is 10, 20, 30, 40, 50, 60, 70 an 80 respectively where the lengths of all the sequences are around 4000. Another set also consists of eight datasets. All the eight datsets include 40 sequences. The lengths of all sequences in the eight datasets are around 1000, 2000, 3000, 4000, 5000, 6000, 7000 and 8000 respectively. We build the phylogenetic tree on each dataset of the two sets by the four methods and record the time each method takes. The results in Fig. 2 shows that our method is much faster than the other three methods. The time of our method increases linearly when the number of sequences or the length of sequences increases, whereas the acceleration of the time for the other four methods is much higher. The actual time differences are much higher than the visual differences in the figure since we are using the log(time) as the label of y-axis.

5. Conclusion

This paper introduces the distribution vectors to map the DNA sequences into the sixty dimensional Euclidean space. We prove that expectation and standard deviation of the distribution vectors do not depend on the length of the sequences. The experiments on the human DNA sequences and random sequences confirm the result. The determined expectation and standard deviation show that the distribution vector mapping is bounded and stable. Each component of the distribution vectors represents the distribution of one kind of nucleotide in k segments of the DNA sequence and plays the same important role in the mapping and clustering. Furthermore, we do the clustering on the Haemagglutinin (HA) gene of 60 H1N1 viruses, 80 mitochondrial complete genomes and 50 complete bacteria genomes with the distribution vector method and other four methods. The phylogenetic trees we obtain show that the distances between the distribution vectors correspond to the evolutionary relationships between these sequences. Our method works for a set of genome sequences or a set of gene sequences. Most importantly, the distribution vector method is much faster than the other methods. Hence our method is more efficient to deal with huge datasets than the other methods. Especially, the distribution vector method only needs to compute the distribution vector of a new sequence when it is put in the dataset, while those multiple sequence alignment methods have to do the multiple sequence alignment on the new dataset when a new sequence is added. It will be more practical to find the closest sequence to the new sequence in a huge dataset with the distribution vector method. Our method may help to discover the functionality or the evolution of the new sequence.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ympev.2011.02.020.

References

Brown, N.P., Larkin, M.A., Blackshields, G., 2007. Clustal w and clustal x version 2.0. Bioinformatics 23 (21), 2947–2948.

Edgar, Robert C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 32 (5), 1792–1797.

Katoh, Kazutaka, Asimenos, George, Toh, Hiroyuki, 2009. Multiple alignment of dna sequences with mafft. Methods in Molecular Biology 537, 39–64.

Kryukov, K., Saitou, N., 2010. MISHIMA – a new method for high speed multiple alignment of nucleotide sequences of bacterial genome scale data. BMC Bioinformatics 11 (142).

R Development Core Team. 2008. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Takahashi, M., Krukov, K., Saitou, N., 2009. Estimation of bacterial species phylogeny through oligonucleotide frequency distances. Genomics 93, 525–533.

Woese, C.R., Fox, G.E., 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proceedings of the National Academy of Sciences 74 (11), 5088–5090.

Yau, Stephen S.-T., Yu, Chenglong, He, Rong, 2008. A protein map and its application. DNA and Cell Biology 27 (5), 241–250.

Further Reading

Iwabe, N., Kuma, K., Hasegawa, M., Osawa, S., Miyata, T., 1989. Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. Proceedings of the National Academy of Sciences of the United States of America 86 (23), 9355–9359.

Raina, Sameer Z., Faith, Jeremiah J., Disotell, Todd R., Seligmann, Herv, Stewart, Caro-Beth, Pollock, David D., 2005. Evolution of base-substitution gradients in primate mitochondrial genomes. Genome Research 15 (5), 665–673.

Smith, Gavin J.D., Vijaykrishna, Dhanasekaran, Bahl, Justin, Lycett, Samantha J., Worobey, Michael, Pybus, Oliver G., Ma, Siu K., Cheung, Chung L., Raghwani, Jayna, Bhatt, Samir, Malik Peiris, J.S., Guan, Yi, Rambaut, Andrew, 2009. Origins and evolutionary genomics of the 2009 swine-origin h1n1 influenza a epidemic. Nature 459 (7250), 1122–1125.

Waterhouse, Andrew M., Procter, James B., Martin, David M.A., Clamp, Michele, Barton, Geoffrey J., 2005. Jalview version 2 - a multiple sequence alignment editor and analysis workbench. Bioinformatics 25 (9), 1189–1191.