The Ninth International Symposium on Operations Research and Its Applications (ISORA'10)
Chengdu-Jiuzhaigou, China, August 19–23, 2010

Composition Vector Method for Phylogenetics — A Review

Raymond H. Chan[1,*] Roger Wei Wang[2]
Hau Man Yeung[1]

[1]Department of Mathematics, The Chinese University of Hong Kong,
Shatin, N.T., Hong Kong
[2]CAS-MPG Partner Institute and Key Lab for Computational Biology,
Shanghai Institutes for Biological Sciences,
Chinese Academy of Sciences, Shanghai 200031, China

Abstract The composition vector (CV) method is an alignment-free method for phylogenetics. Because of its simplicity when compared with the alignment-based methods, the method has been widely discussed lately. There are mainly four steps in the CV method: (1) count the frequency of each k-string in the sequence; (2) construct the composition vector for the sequence; (3) compute the distance between every two composition vectors to form a distance matrix; and (4) construct the phylogenetic tree. In this paper, we review several developments of the CV method respectively.

1 Background

In the past few decades, a large volume of molecular sequences has been generated, and much information on the living organisms' evolutions and traits is thereby provided. These sequences all look very simple, for instance, the DNA sequence, no matter how long it is, only contains four different nucleotides `A`, `C`, `G` and `T`. On this account, these sequences alone cannot tell us too much information. In order to find more, sequence comparison becomes essential. The sequence comparison methods can be divided into two main categories: alignment-based [12, 15] and alignment-free [9, 11, 13, 20].

All alignment-based methods use the dynamic programming method to "align" the sequences and then calculate the similarity or dissimilarity scores after the alignment. To compare two sequences of length n by any alignment-based method, the computational cost and the memory requirement are both $\mathcal{O}(n^2)$ [12, 15]. Because of the accuracy of the dynamic programming method, currently, the alignment-based methods are widely used for analyzing the gene sequences. But different gene sequences may give different evolutionary results. For instance, based on the 18rRNA sequences, birds, which are more closely related to crocodilians, were grouped with mammals [25]. In addition, based on the gene sequences for MHG-CoA reductase, Archaeoglobus fulgidus, a definite archaean, was assigned into the Bacteria [5]. In the meantime, with the development of

*Corresponding author. Research was supported in part by HKRGC Grant 400708 and CUHK DAG 2060257. Email: rchan@math.cuhk.edu.hk

sequence techniques, more and more whole genome sequences are available and they have been generally accepted as excellent tools for the study of evolution [6]. But it is found that aligning the whole genomes is a great challenging problem, as every species has its own gene content and gene order, and we do not know which two genes can be aligned. Furthermore, as the genome sequences are usually very long, both of the computational cost and the memory requirement are expensive.

The alignment-free methods are thereby proposed for the whole genome phylogenetics. Among them, the composition vector method has drawn substantial attention recently [2, 7, 10, 11, 13, 27]. In this paper, we will review several developments of this method.

2 Introduction

The composition vector (CV) method was proposed by Hao *et al.* [9, 13] for the whole-genome-based prokaryotic phylogeny. Their phylogenetic tree provided a classification of the three domains of life which is consistent with those based on traditional analysis. Because of its success, quite a few models have been proposed along this direction. All models of the CV method consist mainly of the following four steps:

1. Construct the frequency vectors: two different methods for constructing the frequency vectors will be introduced based on different biological sequences in Section 3.
2. Construct the composition vectors: for each species, its corresponding composition vector is constructed, with each entry being a signal-to-noise ratio. Several models will be introduced respectively for estimating the noise in Section 4.
3. Compute the distance between every pair of composition vectors: several distance measures will be introduced and analyzed in Section 5. The distance represents the evolutionary distance between the corresponding species.
4. Build the phylogenetic trees: given the distance matrix obtained in Step 3, the neighbor-joining method [14, 18, 19] is used to build the phylogenetic tree afterwards.

3 Frequency Vector

Consider a molecular sequence (nucleotide or amino acid sequence) of length N. Any consecutive k molecules within the sequence is called a *k-string*, where $1 \leq k \leq N$. We use a window of length k and slide it through the sequence by shifting one position at a time to determine the frequencies $f(\mathtt{u})$ of each k-string $\mathtt{u}$ in the sequence [2, 3, 13],

$$f(\mathtt{u}) = \frac{g(\mathtt{u})}{N-k+1}, \tag{1}$$

where $g(\mathtt{u})$ is the number of times that $\mathtt{u}$ appears in the sequence.

For the whole DNA sequence, the frequency vector is the vector with its entry given by (1). For the protein-coding DNA sequences, each entry of its frequency vector is given by:

$$f(\mathtt{u}) = \frac{\sum_{j=1}^{m} g_j(\mathtt{u})}{\sum_{j=1}^{m} (N_j - k + 1)}, \tag{2}$$

see [7, 13, 27]. Here m is the number of protein-coding DNA sequences from the whole genome, $g_j(\mathtt{u})$ is the number of times that $\mathtt{u}$ appears in the jth DNA sequence, and N_j is the length of the jth DNA sequence. We remark here that the usage of (2) avoids the problems from the gene order and the gene content in a genome sequence. For the amino acid sequences of all protein-coding sequences, the frequency vector can be constructed similarly, with each entry defined by (2). Therefore, the frequency vector is of length 4^k for the whole DNA sequence or protein-coding DNA sequence, and of length 20^k for the amino acid sequence of the protein-coding sequence.

4 Composition Vector

It is generally accepted that the phylogenetic signals in the biological data are often obscured by noise and bias [4]. Therefore, denoising is essential for the CV method. In the following, several models are introduced to estimate the noise, and the composition vector is then constructed with each entry being a signal-to-noise ratio. Specifically, for each $f(\mathtt{u})$, with the appearance frequency of the k-string $\mathtt{u}$ defined by (1) or (2), the estimated noise is denoted by $q(\mathtt{u})$. Then the composition vector of the species is the 4^k- or 20^k-vector, where each entry equals

$$\frac{f(\mathtt{u})-q(\mathtt{u})}{q(\mathtt{u})},$$

the signal-to-noise ratio of the k-string $\mathtt{u}$. Before introducing the noise-estimation models, for any k-string $\mathtt{u}$, let us write it as $\mathtt{LwR}$, where the characters "$\mathtt{L}$" and "$\mathtt{R}$" represent the first and the last nucleotide of $\mathtt{u}$ respectively, and "$\mathtt{w}$" represents the $(k-2)$-string in the middle. Moreover, we only consider DNA sequences in the followings. The amino acid sequences can be considered in a similar way.

4.1 Markov Model

The probability of the appearance of the k-string $\mathtt{LwR}$ in the molecular sequence can be estimated as,

$$\mathbb{P}(\mathtt{LwR}) = \mathbb{P}(\mathtt{Lw})\mathbb{P}(\mathtt{R}|\mathtt{Lw}) \approx \mathbb{P}(\mathtt{Lw})\mathbb{P}(\mathtt{R}|\mathtt{w}) = \frac{\mathbb{P}(\mathtt{Lw})\mathbb{P}(\mathtt{wR})}{\mathbb{P}(\mathtt{w})}.$$

Here the equalities hold on both sides by the relationship between the joint probability and the conditional probability. The approximation above assumes the Markov property. Hao *et al.* [9, 13] employed the following formula:

$$q^{\mathrm{Hao}}(\mathtt{LwR}) = \frac{f(\mathtt{Lw})f(\mathtt{wR})}{f(\mathtt{w})}, \tag{3}$$

to estimate the noise of the k-string $\mathtt{LwR}$ in the original sequence. If the denominator in (3), i.e. $f(\mathtt{w})$, is found to be zero, then it means that the $(k-2)$-string does not appear in the sequence. Obviously the $(k-1)$-strings $\mathtt{Lw}$ and $\mathtt{wR}$ will not appear in the sequence, and then

$$f(\mathtt{Lw}) = f(\mathtt{wR}) = 0.$$

When this degeneracy case happens, one can simply let $q^{\mathrm{Hao}}(\mathtt{LwR}) = 0$. Formula (3) has been found to be useful for the phylogenetic analysis of prokaryotes, chloroplasts, viruses etc. based on their whole genome sequences [3, 7, 8, 13, 23, 24].

4.2 Dynamical Language Model

For the probability of the appearance of the k-string $\mathtt{LwR}$, if we assume the independence property, we have

$$\mathbb{P}(\mathtt{LwR}) = \mathbb{P}(\mathtt{Lw})\mathbb{P}(\mathtt{R}|\mathtt{Lw}) = \mathbb{P}(\mathtt{Lw})\mathbb{P}(\mathtt{R}),$$

and

$$\mathbb{P}(\mathtt{LwR}) = \mathbb{P}(\mathtt{L}|\mathtt{wR})\mathbb{P}(\mathtt{wR}) = \mathbb{P}(\mathtt{L})\mathbb{P}(\mathtt{wR}),$$

and then

$$\mathbb{P}(\mathtt{LwR}) = \frac{\mathbb{P}(\mathtt{Lw})\mathbb{P}(\mathtt{R}) + \mathbb{P}(\mathtt{L})\mathbb{P}(\mathtt{wR})}{2}.$$

Yu *et al.* [27] proposed the formula:

$$q^{\mathrm{Yu}}(\mathtt{LwR}) = \frac{f(\mathtt{L})f(\mathtt{wR}) + f(\mathtt{Lw})f(\mathtt{R})}{2}, \tag{4}$$

to estimate the noise of the k-string. As formula (4) also appeared in the theory of dynamical language, this model is called the *dynamical language model*. Formula (4) has been found to be useful and sometimes provide better results for the whole-genome-based phylogenetic analysis [26, 27].

4.3 Maximum Entropy Principle Model

Chan *et al.* [1] employed the maximum entropy principle to estimate the noise. They first assumed that the noise $q(\cdot)$ of the k-strings satisfies

$$\begin{cases} q(\mathtt{vA}) + q(\mathtt{vC}) + q(\mathtt{vG}) + q(\mathtt{vT}) = l(\mathtt{v}), \\ q(\mathtt{Av}) + q(\mathtt{Cv}) + q(\mathtt{Gv}) + q(\mathtt{Tv}) = r(\mathtt{v}), \end{cases} \tag{5}$$

where $l(\mathtt{v})$ and $r(\mathtt{v})$ are given non-negative numbers for each $(k-1)$-string $\mathtt{v}$. Entropy is widely used in information theory as a numerical measure of the missed information content. The larger the value of entropy is, the more information is missing, or say, the variable is associated with more randomness. Based on this, Chan *et al.* maximized the entropy to find the most uninformative noise. More precisely, let $q_i \equiv q(\mathtt{u}_i)$ be the noise of the k-string $\mathtt{u}_i$, q_i is then obtained by solving the optimization problem:

$$\begin{aligned} &\text{maximize} && -\sum_{i=1}^{4^k} q_i \log q_i \\ &\text{subject to} && \begin{cases} q_i \text{ satisfies (5)}, \\ q_i \geq 0 \text{ for all } i. \end{cases} \end{aligned} \tag{6}$$

We note that $-q_i \log q_i$ is the entropy of q_i.

After solving (6), a system of noise estimation formulae is provided by:

$$q^{\mathrm{MEP}}(\mathtt{LwR}) = \frac{l(\mathtt{Lw})r(\mathtt{wR})}{\sigma},$$

where

$$\sigma = \sum_{\mathtt{L}\in\{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\}} l(\mathtt{Lw}) = \sum_{\mathtt{R}\in\{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\}} r(\mathtt{wR}).$$

We remark that $l(\cdot)$ and $r(\cdot)$ are parametric functions, and different $l(\cdot)$ and $r(\cdot)$ will give different estimation formulae. If Hao's formula (3) is used for the $q(\mathtt{LwR})$ in (5), we will obtain a formula which is exactly the same as (3). If Yu's formula (4) is used in (5), we will have a new formula:

$$q(\mathtt{YwZ}) = \frac{1}{4\sigma}\left[f(\mathtt{Yw}) + f(\mathtt{Y})\sum_{\mathtt{R}} f(\mathtt{wR})\right]\left[f(\mathtt{wZ}) + f(\mathtt{Z})\sum_{\mathtt{L}} f(\mathtt{Lw})\right], \tag{7}$$

where

$$\sigma = \frac{1}{2}\left[\sum_{\mathtt{L}} f(\mathtt{Lw}) + \sum_{\mathtt{R}} f(\mathtt{wR})\right].$$

According to the traditional classification and the results derived from a large amount of molecular, morphological and paleontological data, birds are thought to be first grouped with crocodilians, and then as a whole grouped with mammals [2]. However, many studies based on 18S rRNA sequences supported the grouping of birds and mammals first. The CV method with formula (7) provided a tree which supports the traditional and widely accepted classification [1].

5 Distance Measure

Denote the set of composition vectors by $\mathbb{S}$. We define our distance measure as follows:

(1) (Non-negativity) $0 \le d(\mathbf{a},\mathbf{b}) < +\infty$ for all $\mathbf{a}$ and $\mathbf{b} \in \mathbb{S}$.
(2) (Identity of indiscernibles) $d(\mathbf{a},\mathbf{b}) = 0$ if and only if $\mathbf{a} = \mathbf{b}$.
(3) (Symmetry) $d(\mathbf{a},\mathbf{b}) = d(\mathbf{b},\mathbf{a})$ for all $\mathbf{a}$ and $\mathbf{b} \in \mathbb{S}$.

Note that the "triangle inequality" of the "metric distance"

$$d(\mathbf{a},\mathbf{b}) \le d(\mathbf{a},\mathbf{c}) + d(\mathbf{c},\mathbf{b}), \qquad \forall \mathbf{a},\mathbf{b},\mathbf{c} \in \mathbb{S},$$

is not included in our definition.

5.1 Angle-based distance

To measure the distance between the composition vectors $\mathbf{a}$, $\mathbf{b} \in \mathbb{S}$, it is common to employ the cosine of their angle defined as:

$$\cos\theta = \frac{\mathbf{a}^T\mathbf{b}}{\|\mathbf{a}\| \cdot \|\mathbf{b}\|}, \tag{8}$$

where $\|\cdot\|$ represents Euclidean distance.

Stuart *et al.* [16, 17] was the first to introduce the angle distance for the phylogenetic analysis. Their formula is as follow:

$$d^{\text{Stuart}}(\mathbf{a},\mathbf{b}) = -\log\left(\frac{1+\cos\theta}{2}\right) = -\log\left[\frac{1}{2}\left(1 + \frac{\mathbf{a}^T\mathbf{b}}{\|\mathbf{a}\| \cdot \|\mathbf{b}\|}\right)\right].$$

It is easy to check that this formula satisfies the three conditions of a distance mentioned above.

Hao *et al.* [9, 13] proposed the formula:

$$d^{\text{Hao}}(\mathbf{a},\mathbf{b}) = \frac{1-\cos\theta}{2} = \frac{1}{2}\Big(1 - \frac{\mathbf{a}^T\mathbf{b}}{\|\mathbf{a}\|\cdot\|\mathbf{b}\|}\Big) \tag{9}$$

for measuring the difference between two composition vectors $\mathbf{a}$ and $\mathbf{b}$. We can verify that this measure is a distance satisfying the three conditions. Since the cosine value computed by (8) varies between -1 and 1, the function value of $d^{\text{Hao}}(\mathbf{a},\mathbf{b})$ is normalized to the interval $[0,\ 1]$. Till now distance (9) is widely used and achieved a great success in the phylogenetic analysis of whole genomes of bacteria, viruses, and vertebrates [2, 3, 7, 8, 13, 27].

Although distance (9) is defined based on the cosine of the angle, it is the same as the square of the Euclidean distance of the normalized vectors:

$$d^{\text{Hao}}(\mathbf{a},\mathbf{b}) = \frac{1}{4}\left\| \frac{\mathbf{a}}{\|\mathbf{a}\|} - \frac{\mathbf{b}}{\|\mathbf{b}\|} \right\|^2.$$

5.2 Information theory-based distance

In probability and information theory, the Kullback-Leibler divergence, also called respectively relative entropy, is widely employed to measure the difference between two probability distributions. It is defined as follow,

$$KL(\mathbf{a},\mathbf{b}) = \sum_{i=1}^{n} a_i \log\Big(\frac{a_i}{b_i}\Big), \tag{10}$$

where $\mathbf{a} = (a_i)_{i=1}^n$ and $\mathbf{b} = (b_i)_{i=1}^n$ are distribution vectors. Several attempts have been made to introduce this concept into the area of alignment-free methods [10, 21, 22]. For instance, Wu *et al.* [22] considered the following formula

$$KL^{\text{Wu}}(\mathbf{a},\mathbf{b}) = \sum_{\alpha\in D} a_\alpha \log\Big(\frac{a_\alpha}{b_\alpha}\Big), \tag{11}$$

where D is the domain such that $b_\alpha > 0$. However, it is obvious that both of formulae (10) and (11) do not fulfill the symmetric property. Wang and Zheng [21] introduced the Jensen-Shannon divergence:

$$d^{\text{JS}}(\mathbf{a},\mathbf{b}) = \sum_{i=1}^{n}\Big[a_i\log(a_i) + b_i\log(b_i) - (a_i+b_i)\log\big(\frac{a_i+b_i}{2}\big)\Big] \tag{12}$$

into the composition vector approach. We can check that formula (12) satisfies all the three conditions of a distance.

6 Conclusions

As different genes may provide different phylogenies, people have considered to do phylogenetic analysis based on the whole genome sequences. Hao *et al.* proposed the composition vector (CV) method for the whole-genome-based prokaryotic phylogeny, and obtained a three domains of the tree of life. Since then, the CV method has drawn

people's substantial attention and thereby been widely discussed. The CV method has several advantages. For instance, it is a systematic method that requires no scoring matrix or gap penalty. Moreover, for computing the distance between two taxa, its operation cost is $\mathcal{O}(N \log N)$ and the memory requirement is $\mathcal{O}(N)$, where N is the length of the longer sequence. We remark that the fast-computing property is essential for the methods of the whole genome data analysis. As more whole genome sequences are available, phylogenetic analysis is entering a new era. The CV method will be an alterative and faster method than sequence alignment methods in handling these problems.

References

[1] Chan, R.H., Chan, T.H., Yeung, H.M. and Wang, R.W. Composition vector method based on maximum entropy principle for sequence comparison, submitted.

[2] Chu, K.H., Li, C.P. and Qi, J. (2006) Ribosomal RNA as molecular barcodes: a simple correlation analysis without sequence alignment. *Bioinformatics*, **22**, 1690–1710.

[3] Chu, K.H., Qi, J., Yu, Z.G. and Anh, V. (2004) Origin and phylogeny of chloroplasts: A simple correlation analysis of complete genomes. *Molecular Biology and Evolution*, **21**, 200–206.

[4] Charlebois, R.L., Beiko, R.G. and Ragan, M.A. (2003) Microbial phylogenomics: Branching out. *Nature*, **421**, 217–217.

[5] Doolittle, W. F. (1999) *Phylogenetic classification and the universal tree*, Science, **284**, 2124–2128.

[6] Eisen, J. A. and Fraser, C. M. (2003) *Phylogenomics: intersection of evolution and genomics*, Science, **300**, 1706–1707.

[7] Gao, L. and Qi, J. (2007) Whole genome molecular phylogeny of large dsDNA viruses using composition vector method. *BMC Evolutionary Biology*, **7**, 1–7.

[8] Gao, L., Qi, J., Wei, H., Sun, Y. and Hao, B. L. (2003) Molecular phylogeny of coronaviruses including human SARS-CoV. *Chinese Science Bulletin*, **48**, 1170–1174.

[9] Hao, B.L., Qi, J. and Wang, B. (2003) Prokaryotic phylogeny based on complete genomes without sequence alignment. *Modern Physics Letters B*, **2**, 1–4.

[10] Sims, F.E., Jun, S.-R., Wu, G.A. and Kim, S.-H. (2009) Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 2677–2682.

[11] Lu, G., Zhang, S. and Fang, X. (2008) An improved string composition method for sequence comparison. *BMC Bioinformatics*, **9 (Suppl 6)**, S15.

[12] Needleman, S. B., and Wunsch C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, **48**, 443–453.

[13] Qi, J., Wang, B. and Hao, B.L. (2004) Whole proteome prokaryote phylogeny without sequence alignment: A k-string composition approach. *Journal of Molecular Evolution*, **58(1)**, 1–11.

[14] Saitou, N. and Nei, M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, **4**, 406–425.

[15] Smith, T.T. and Waterman M.S. (1981) Identification of common molecular subsequences. *Journal of Molecular Biology*, **147**, 195–197.

[16] Stuart, G.W., Moffett, K. and Baker, S. (2002) Integrated gene and species phylogenies from unaligned whole genome protein sequences. *Bioinformatics*, **62**, 100–108.

[17] Stuart, G.W., Moffett, K. and Leader, J.J. (2002) A comprehensive vertebrate phylogeny using vector representations of protein sequences from whole genomes. *Molecular Biology and Evolution*, **19**, 554–562.

[18] Studier, J.A. and Keppler, K.J. (1988) A note of the neighbor-joining algorithm of Saitou and Nei. *Molecular Biology and Evolution*, **5**, 729–731.

[19] Tamura, K., Dudley, J., Nei, M. and Kumar, S. (2007) MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0. *Molecular Biology and Evolution*, **24**, 1596–1599.

[20] Vinga, S. and Almeida, J. (2003) Alignment free sequence comparison-a review. *Bioinformatics*, **19**, 513–523.

[21] Wang, J. and Zheng, X. *WSE, a new sequence distance measure based on word frequencies*, Mathematical Biosciences, **215** (2008), 78–83.

[22] Wu, T. J., Hsieh, Y. C. and Li, L. A. (2001) *Statistical measures of DNA sequence dissimilarity under Markov chain models of base composition*, **57**, 441–448.

[23] Wu, X., Cai, Z., Wan, X.-F., Hoang, T., Goebel, R. and Liu, G. (2007) Nucleotide composition string selection in HIV-1 subtyping using whole genomes. *Bioinformatics*, **23**, 1744–1752.

[24] Wu, X., Wan, X.-F., Wu, G., Xu, D. and Lin, G. (2006) Phylogenetic analysis using complete signature information of whole genomes and clustered Neighbor-Joining method. *International Journal of Bioinformatics Research and Applications*, **2**, 219–248.

[25] Xia, X., Xie, Z. and Kjer, K.M. (2003) 18S ribosomal RNA and tetrapod phylogeny. *Systematic Biology*, **52**, 283–295.

[26] Yu, Z.G., Chu, K.H., Li, C.P., Anh, V., Zhou, L.Q. and Wang, R.W. (2010) Whole-proteome phylogeny of large dsDNA viruses and parvoviruses through a composition vector method related to dynamical language model. *BMC Evolutionary Biology*, **10**:192.

[27] Yu, Z.G., Zhou, L.Q., Anh, V., Chu, K.H., Long, S.C. and Deng, J.Q. (2005) Phylogeny of prokaryotes and chloroplasts revealed by a simple composition approach on all protein sequences from whole genome without sequence alignment. *Journal of Molecular Evolution*, **60**, 538–545.