Communications in Information and Systems
Volume 19, Number 1, 81–94, 2019

Virus classification based on Q-vectors

Hui Zheng, Jie Yang, Rong L. He, and Stephen S.-T. Yau

Based on a Markov model, we propose a new alignment-free method, Q-vector (QV), for sequence analysis. It incorporates the length information of viral sequences and could reflect the relationship between low mers and high mers. Compared with the k-mer and composition vector methods, QV method is significantly more efficient and accurate in classifying viral genomes. By incorporating the distance matrices derived by the QV and natural vector, respectively, we define a new distance matrix for classifying viral genomes and reduce the classification errors even further. We also construct the phylogenetic trees based on the new distance.

1. Introduction

Due to the diversity of viruses, their classification becomes crucial. The International Committee on Taxonomy of Viruses (ICTV) proposed a universal taxonomic scheme for all the viruses. Viral classification starts at the level of order and continues as follows: Order, Family, Subfamily, Genus, and Species. Correspondingly, the taxon suffixes are *-virales*, *-viridae*, *-virinae*, *-virus*, and [disease]*virus* for species. For example: Dolphin *morbillivirus* belongs to *Mononegavirales* (Order), *Paramyxoviri-dae* (Family), *Paramyxovirinae* (Subfamily), *Morbillivirus* (Genus) and Measles virus (Species). In this work, we focus more on Family and Genus labels due to the high missing rates of Order and Subfamily and Species labels in the records of ICTV. Another major scheme used for classifying viruses is the Baltimore classification system. Based on the combination of their nucleic acids, strandedness (single-stranded or double-stranded) and methods of replication, viruses are divided into the following seven groups: dsDNA, ssDNA, dsRNA, ssRNA-RT, (+)ssRNA, (-)ssRNA, and dsDNA-RT [1].

Sequence analysis is a tool to study the feature, function, structure, and the evolution of DNA, RNA, and proteins. Especially, it can be used to analyze the similarity between sequences and discover the evolutionary

This work is supported by National Natural Science Foundation of China grant (#91746119).

relationships of species. Sequence comparison methods can be divided into two categories: alignment-based and alignment-free. The alignment-based methods use dynamic programming to align the sequences and find the similarity or dissimilarity between sequences. It takes $O(n^2)$ time and memory for comparing two length-n sequences. Alignment-free methods overcome this disadvantage, which usually cost shorter time and lower memory to get highly accurate results. Just as its name implies, alignment-free methods do not rely on the alignments of sequences. They can be built up on k-mer frequencies, information theory or substrings. Among them, the approach based on k-mer frequencies is very popular and several algorithms have been developed for it, such as, feature frequency profile [10], composition vector [2], return time distribution [7], frequency chaos game representation [6], and spaced words [8]. Alignment-free methods typically start with a pairwise distance for measuring similarities between sequences. There usually exists an exact solution whose statistical significance could be readily assessed. Different from the alignment-based methods, the alignment-free methods depend less on evolutionary models and do not assume that the homologous regions are contiguous [3].

Based on a Markov model, we propose a new alignment-free sequence analysis method, called Q-vector. The Q-vector (QV) is inspired by the composition vector (CV) [2], which was applied in prokaryotic phylogeny [5], whole genome molecular phylogeny of large dsDNA viruses [4], and 16S and 18S rRNA sequences comparison based on maximum entropy principle [2]. The CV is defined on k-mer frequencies in a sequence. Compared with CV, our QV keeps the sequence length information and reflects the relationship of three conjoint mers.

The natural vector (NV) [13] method represents a viral nucleotide sequence by a 12-dimentional numerical vector based on the nucleotide positions, which does not rely on any model assumption. In later session, we will show the comparison work by comparing QV, CV and NV, also we developed a new distance by combined NV and QV.

Having applied QV, k-mer method, and CV to classify the viral reference sequences in seven Baltimore classes, QV shows significant advantage in both efficiency and accuracy. Phylogenetic analysis based on QV of viruses in Balti-more III is done through UPGMA (Unweighted Pair Group Method with Arithmetic Mean). By combining the distance matrix derived through QV and natural vector, we define a new distance matrix, which reduces the classification errors further.

2. Definitions

Most alignment-free sequence analysis methods work by converting sequences into vectors. Here we list some of them under consideration in this work.

2.1. Vectors based on frequencies

a. Frequency vector. Frequency is the proportion of a k-string occurred in a sequence. For a nucleotide sequence with length L, we slide a width-k sliding window along the sequence, and count the number of times n of the k-string occurred. The frequency of k-string is then defined as $\frac{n}{L-k+1}$.

b. *k*-mer vector. k-mer vector is an alignment-free method which is applied for sequence analysis. It has 4^k dimensions for each sequence. Each element of the vector corresponds to the frequency of a k-mer.

Suppose we have a nucleotide sequence with length L. Then the k-mer vector is

$$\left(\frac{g(u_1)}{L-k+1}, \frac{g(u_2)}{L-k+1}, \ldots, \frac{g(u_{4^k})}{L-k+1}\right),$$

where $u_1, u_2, \ldots, u_{4^k}$ are the 4^k k-strings, $g(u)$ is the number of times that the k-string u occurs in the sequence.

For example, we have a sequence $S = ATGCCTG$, the 1-mer vector is $(A, T, G, C) = (\frac{1}{7}, \frac{2}{7}, \frac{2}{7}, \frac{2}{7})$ and the 2-mer vector is

$$\begin{aligned}&(AA, AT, AG, AC, TA, TT, TG, TC, GA, GT, GG, GC, CA, CT, CG, CC)\\ &= \left(0, \frac{1}{6}, 0, 0, 0, 0, \frac{2}{6}, 0, 0, 0, \frac{1}{6}, 0, 0, \frac{1}{6}, 0, \frac{1}{6}\right).\end{aligned}$$

c. Composition vector. For sequence S, $f(u) = \frac{f(u)-q(u)}{q(u)}$ where $f(u)$ is the frequency of k-string u, $q(u)$ is the estimated noise.

d. Q-frequency vector. In the definition of composition vector, $q(u)$ is the estimated noise in the phylogenetic signals [2]. It is based on a Markov model. Given a sequence $LwR = ATGCCTG$ where $L = A$, $R = G$, $w = TGCCT$, based on the joint probability and conditional probability formulae, we have

$$P(LwR) = P(Lw)P(R \,|\, Lw).$$

Assuming the Markov property, when the conditional probability $P(R \,|\, Lw)$ does not depend on L, we have

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

Then the estimated noise of the k-string LwR is

$$q(LwR) = \frac{f(Lw) * f(wR)}{f(w)},$$

where $f(u)$ is the frequency of u.

2.2. Vectors based on counts

a. *k*-mer count vector. Compared with the k-mer (frequency) vector, k-mer count vector replaces the frequencies with the corresponding counts of k-strings. For example, given a sequence $S = ATGCCTG$, the 1-mer count vector is $A, T, G, C = (1, 2, 2, 2)$, and the 2-mer count vector is

$$\begin{aligned}&(AA, AT, AG, AC, TA, TT, TG, TC, GA, GT, GG, GC, CA, CT, CG, CC)\\ &= (0, 1, 0, 0, 0, 0, 2, 0, 0, 0, 0, 1, 0, 1, 0, 1)\end{aligned}$$

b. Q-count vector. The Q-frequency vector is presented by the frequencies of the k-strings, which ignores the length of the sequence. The following Q-count vector uses the numbers of the k-strings instead which improves the classification accuracy dramatically in our experiments. For the sequence $LwR = ATCGCCTC$, where $L = A$, $R = G$, $w = TGCCT$ we define

$$q(LwR) = \frac{N(Lw) * N(wR)}{N(w)}$$

where $N(x)$ is the count of the string x. We use QV to represent Q-count vector.

2.3. Distance measure

(a) Angle distance. The cosine of the angle between the two vectors $a, b \in S$ is defined as:

$$\cos\theta = \frac{a^T b}{\|a\| \cdot \|b\|}.$$

For comparison purpose we use the same formula [2] as the angle distance between two vectors:

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

For frequency vectors, it is reasonable to apply angle distance on them. We apply it on k-mer frequency vector, composition vector and Q-frequency vector.

(b) Euclidean distance. The Euclidean distance between two vectors $X = (x_1, x_2, \ldots, x_n)$ and $Y = (y_1, y_2, \ldots, y_n)$ is defined as

$$d(X, Y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2}.$$

We apply Euclidean distance on k-mer count vector, Q-count vector in this work.

2.4. Phylogenetic tree

For a given virus sequence dataset, UPGMA first regard each virus sequence as a cluster, then groups two smaller clusters of nodes to build up the phylogenetic tree until there is only one tree left that contains all the virus sequences. It is a clustering method based on a distance matrix and has been widely used in sequence similarity analysis [9].

3. Dataset and methods

Totally 1,988 single-segmented reference viruses up to Feb. 14, 2014 were downloaded from NCBI. For the purpose of method comparison, we removed the viruses with missing family or genus labels. Among the 1,988 viruses, there are 1,752 viruses left after removing those missing family labels, and 1,422 viruses left after removing those without genus labels. Note that each virus belongs to one of the seven Baltimore classes. For further study, we divide the whole data set into seven groups according to their Baltimore labels.

The parameter k is critical on computational complexity and the result of sequence comparison for k-mer method, CV, and QV. However, some researchers showed that there are some relationships between suitable k and sequences' lengths. For example, an optimal k for dissimilarity measurement should be increased when the sequence length increases [12] and the optimal word length lies within an approximate range with a lower bound $\log_4(L)$, where L is the length of sequence, and an upper bound given by the criterion that the phylogenetic tree topology for length k must be parallel to that of $k + 1$ [10]. So far, there is no recognized criterion on choosing the optimal k for k-mer models [11]. By combining the previous research and

	All	Balt I	Balt II	Balt III	Balt IV	Balt V	Balt VI	Balt VII
Total	1833	758	323	45	539	67	39	62
LF	1752	732	310	36	509	64	37	62
LG	1422	415	302	30	503	56	37	55
k	6	7	5	6	6	6	6	6

Table I: The dataset and options of each Baltimore group.

our experimental results, we propose

$$k = floor(\log_4(median(s(L))))$$

where $s(L)$ is the set of sequence lengths. In other words, we choose k to be the largest integer no greater than $\log_4(median(s(L)))$.

Table I shows the detailed dataset information considered in this work. The row of "Total" lists the numbers of viruses after removing those without Baltimore information and the numbers of viruses in each Baltimore group. "LF" means "left Family", listing the numbers of viruses after removing those without family labels and the numbers of viruses in each Baltimore groups. "LG" means "left Genus", lists the numbers of viruses after removing those without genus labels and the numbers of viruses in each Baltimore groups. The last row, "k", lists the option of k according to the median of sequence lengths for that Baltimore group. Besides Balt I and Balt II, other five Baltimore groups all choose $k = 6$ according to our criterion.

4. Results

We extracted the Baltimore information, family label and genus label of each virus downloaded from NCBI. In the dataset, we calculated the distance matrix based on QV, CV, k-mer methods, and NV. We assigned those viruses with their smallest distance to the same group and predict their family labels and genus labels. If the predicted label is not consistent with the dataset, we regard it as an error. Table II shows the error counts of each method when we predict family labels. A smaller number indicates the better model. We compare QV.counts and QV.frequency, Mer.counts and Mer.frequency, respectively. Count vectors show significant advantages according to Table II.

Although composition vector performed well on other dataset such as [14], it is the worst one on our virus dataset. We also find out that QV.count vector has much less errors compared with mer.count vector. Note that the error counts are NA for Balt VI, because there is only one family for this Baltimore group and thus no need for classification.

	Total	QV.count	Mer.count	QV.frequency	Mer.frequency	CV
ALL	1752	143	216	365	754	677
Balt I	732	87	116	177	234	452
Balt II	310	8	9	9	14	30
Balt III	36	4	3	12	17	7
Balt IV	509	27	42	86	178	348
Balt V	64	0	2	4	9	17
Balt VI	37	NA	NA	NA	NA	NA
Balt VII	62	3	4	2	1	17

Table II: Classification comparison through counts and frequencies methods (Family). Note: *QV.count stands for Q-vector based on counts, QV.frequency stands for Q-vector based on frequency, CV stands for composition vector.*

Table III shows the error counts of each method when we classify genus labels. There is a similar pattern as in Table II when we classify the family labels. Besides the error counts of Balt VI that count vectors have 2 or 3 more error counts than the frequency vectors, count vector performs pretty well on all other groups. Also, QV.count vector is the best one among the five methods.

The reason that count vectors perform better than frequency vectors is that, the length information does matter for virus classification. However, frequency vectors use the k-string count divided by the sequence length, which scales the vector and ignores the length information. Count vectors use the original count of the k-string and keep the length information, which reflects the relationship of viruses better. For example, Family *Phycodnaviridae* has 10 members and the sequence lengths are between 185,373 and 407,339; family *Adenoviridae* has 24 members and the sequence lengths are between 26,263 and 45,063; family *Hepadnaviridae* has 8 members and the sequence lengths are between 3,027 and 3,323, etc.. If we keep the length

	Total	QV.count	Mer.count	Qv.frequency	Mer.frequency	CV
ALL	1422	146	211	287	654	547
Balt I	415	50	72	123	149	259
Balt II	302	11	12	12	18	38
Balt III	30	5	7	10	18	19
Balt IV	503	53	71	118	191	358
Balt V	56	2	2	8	11	33
Balt VI	37	5	6	2	4	14
Balt VII	55	6	8	6	8	3

Table III: Classification comparison through counts and frequency (Genus). Note: *QV.count stands for Q-vector based on counts, QV.frequency stands for Q-vector based on frequency, CV stands for composition vector.*

information, we may easily separate these three families by count vector, while it is not guaranteed if we use frequency vectors. Similar story happens for genus labels. For examples, family *Alpha exiviridae* has 8 genera, the sequence lengths of genus *Allexivirus* are between 8,106 and 8,660; the sequence lengths of *Batrachovirus* are between 220,859 and 231,801; the lengths of *Cyprinivirus* are between 248,526 and 295,146; the lengths of *Potexvirus* are between 5,816 and 7,212, etc.. Another good example, family *Astroviridae* has 3 genera, genus *Ascovirus* has sequence lengths between 119,343 and 186,262, genus *Avastrovirus* has sequence lengths between 6,927 and 7,722, genus *Mamastrovirus* has sequence lengths between 6,440 and 6,813. We conclude that count vectors are more suitable for virus classification.

Note that NV keeps the position information of the sequences, and QV reflects the relationship of previous segment and last segment of the sequences. In order to utilize their advantages, we define a new distance:

$$d^{new}(v_i, v_j) = \frac{d^{qv}(v_i, v_j)}{\max(D_{ij}^{qv})} + \frac{d^{nv}(v_i, v_j)}{\max(D_{ij}^{nv})}$$

where v_i is the i^{th} virus, $d^{qv}(v_i, v_j)$ is the distance of v_i and v_j based on QV when $k = 6$ or $k = 7$, D_{ij}^{qv} is the distance matrix of all viruses based on QV,

$d^{nv}(v_i, v_j)$ is the distance of v_i and v_j based on NV, and D_{ij}^{nv} is the distance matrix of all viruses based on NV.

The reason that we use the distances between v_i and v_j divided by the maximum of the distance matrix is to make a consistent scale of the QV distance and the NV distance. By adding them together, we avoid seeing one is too big while the other is too small.

We apply this new distance to the above classification job. Table IV shows the comparison results. Among the whole dataset and the seven Baltimore groups, the new distance achieves the minimum error count in six Baltimore groups and only 2 more error counts than QV and NV methods, and 1 more error count than QV method. In addition, the new distance performs excellent for Baltimore group I and only has 1/2 error counts of QV method, and 1/4 error counts of NV method. Table V shows the genus label classification results of the new distance. The new distance performs almost same as well as QV method. Phylogenetic tree of Baltimore group III is shown in Figure 1. Baltimore III has 36 viruses belonging to three families: *Endornaviridae*, *Hypoviridae* and *Totiviridae*. These three families are well separated in the tree.

	Total	QV	NV	Newdist
ALL	1752	143	316	145
Balt I	732	87	167	39
Balt II	310	8	9	5
Balt III	36	4	3	2
Balt IV	509	27	47	17
Balt V	64	0	7	1
Balt VI	37	NA	NA	NA
Balt VII	62	3	4	2

Table IV: Classification comparison through different distance (Family).

	Total	QV	NV	Newdist
ALL	1422	142	296	158
Balt I	415	50	167	61
Balt II	302	11	9	9
Balt III	30	5	8	7
Balt IV	503	53	47	47
Balt V	56	2	6	4
Balt VI	37	5	12	6
Balt VII	55	6	6	6

Table V: Classification comparison through different distance (Genus).

4.1. Minor finding

In order to find the optimal k to reduce the classification errors, we check the performance when $k = 3, 4, 5, 6, 7, 8, 9, 10$. The computational time and storage space increase dramatically when k increases. For example, when $k = 9$, the data size of 9-mer vectors for whole dataset reached 3.9G and the computational time is 9.3 hours; when $k = 10$, the data size of 10-mer vectors is 15.5G and it spends 25 hours to finish the computation. Note that the computational time includes both the calculation from sequences to vectors and the calculation of the distance matrix.

In order to reduce the computational cost, we employ a dictionary method. The dictionary method is a method to simplify the computation and decrease the computing time. However, this method does not work here due to the large size of our dataset. We sum up all the mer counts of each virus for $k = 6, 7, 8, 9$ and find that all mers appear in the dataset. When $k = 10$, 1,988 viruses with 4^{10} has 15.5G. Adding up the mers of the first 500 viruses, we find out that only 2,616 mers never appear in the dataset. The rate is $2616/4^{10} = 0.25\%$. We conclude that this method does not work for this dataset.

In order to test the robustness of NV, k-mer and QV, we select a sequence with length 11,965, which is close to the mean of the lengths of all the sequences in this dataset. We delete one letter from the start to the end of

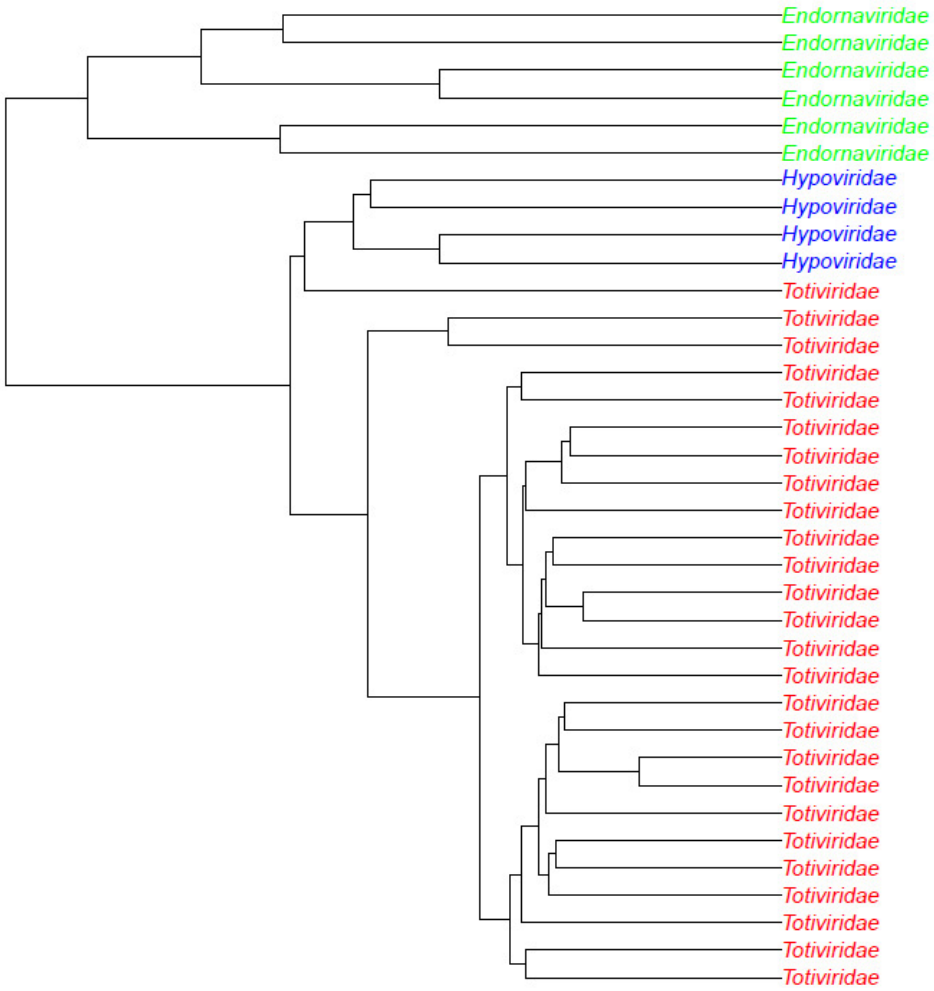

Figure 1: Phylogenetic tree of Baltimore group III based on QV-method through UPGMA.

the sequence and get 11,965 new sequences with length 11,964. We calculate the QV vectors for these sequences and denote them as $q_1, q_2, q_3, \ldots, q_{11965}$. We denote the QV vector of the original sequence as q_0. Then

$$diff = \sum_{i=1}^{11965} \frac{dist(q_i, q_0)}{\|q_0\| \cdot 11965}.$$

Doing the same calculation for NV and k-mer ($k = 6$), we get the following results:

Method	NV	k-mer	QV
diff	0.0001295576	0.005023519	0.006834223

Table VI: Robustness comparison.

From the above Table VI, NV shows the smallest distance between before and after the letter deletion. It indicates that NV has the best robustness when deletion happens in the sequence.

References

[1] D. Baltimore, *Expression of animal virus genomes*, Bacteriological Reviews **35** (1971), no. 3, 235.

[2] R. H. Chan, T. H. Chan, H. M. Yeung, and R. W. Wang, *Composition vector method based on maximum entropy principle for sequence comparison*, Com-putational Biology and Bioinformatics, IEEE/ACM Transactions on **9** (2012), no. 1, 79–87.

[3] M. Deng, C. Yu, Q. Liang, R. L. He, and S. S.-T. Yau, *A novel method of characterizing genetic sequences: genome space with biological distance and applications*, PloS one **6** (2011), no. 3, e17293.

[4] L. Gao and J. Qi, *Whole genome molecular phylogeny of large dsdna viruses using composition vector method*, BMC evolutionary biology **7** (2007), no. 1, 41.

[5] B. Hao, J. Qi, and B. Wang, *Prokaryotic phylogeny based on complete genomes without sequence alignment*, Modern Physics Letters B **17** (2003), no. 03, 91–94.

[6] K. Hatje and M. Kollmar, *A phylogenetic analysis of the brassicales clade based on an alignment-free sequence comparison method*, Frontiers in Plant Science **3** (2012).

[7] P. Kolekar, M. Kale, and U. Kulkarni-Kale, *Alignment-free distance measure based on return time distribution for sequence analysis: applications to clustering, molecular phylogeny and subtyping*, Molecular Phylogenetics and Evolution **65** (2012), no. 2, 510–522.

[8] C.-A. Leimeister, M. Boden, S. Horwege, S. Lindner, and B. Morgenstern, *Fast alignment-free sequence comparison using spaced-word frequencies*, Bioinformatics (2014), btu177.

[9] J. Sourdis and C. Krimbas, *Accuracy of phylogenetic trees estimated from dna sequence data*, Molecular Biology and Evolution **4** (1987), no. 2, 159–166.

[10] G. E. Sims, S.-R. Jun, G. A. Wu, and S.-H. Kim, *Whole-genome phylogeny of mammals: evolutionary information in genic and nongenic regions*, Proceedings of the National Academy of Sciences **106** (2009), no. 40, 17077–17082.

[11] J. Wen, Y. Zhang, and S. S. Yau, *k-mer sparse matrix model for genetic sequence and its applications in sequence comparison*, Journal of Theoretical Biology **363** (2014), 145–150.

[12] T.-J. Wu, Y.-H. Huang, and L.-A. Li, *Optimal word sizes for dissimilarity measures and estimation of the degree of dissimilarity between dna sequences*, Bioinformatics **21** (2005), no. 22, 4125–4132.

[13] C. Yu, T. Hernandez, H. Zheng, S.-C. Yau, H.-H. Huang, R. L. He, J. Yang, and S. S.-T. Yau, *Real time classification of viruses in 12 dimensions*, PloS one **8** (2013), no. 5, e64328.

[14] Z.-G. Yu, L.-Q. Zhou, V. V. Anh, K.-H. Chu, S.-C. Long, and J.-Q. Deng, *Phylogeny of prokaryotes and chloroplasts revealed by a simple composition approach on all protein sequences from complete genomes without sequence alignment*, Journal of Molecular Evolution **60** (2005) no. 4, 538–545.

Hui Zheng[1,4,†] Jie Yang[1] Rong L. He[2], and Stephen S.-T. Yau[3,‡]

1. Department of Mathematics, Statistics, and Computer Science
University of Illinois at Chicago, Chicago, Illinois, USA

2. Department of Biological Sciences, Chicago State University
Chicago, Illinois, USA

3. Department of Mathematical Sciences, Tsinghua University
Beijing 100084, China

4. AbbVie Inc.

† First author and co-corresponding author.
E-mail address: huizhenguic@gmail.com
‡ Corresponding author. *E-mail address*: yau@uic.edu