AGP: A Multimethods Web Server for Alignment-Free Genome Phylogeny

Jinkui Cheng,[†,1,2] Fuliang Cao,[†,1] and Zhihua Liu*,[1,2]

[1]Nanjing Forestry University, Nanjing, China

[2]Department of Computational Biology and Bioinformatics, Institute of Medicinal Plant Development, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing, China

[†]These authors contributed equally to this work.

***Corresponding author:** E-mail: zhliu.liulab@foxmail.com, zhliu@implad.ac.cn.

Associate editor: Koichiro Tamura

Abstract

Phylogenetic analysis based on alignment method meets huge challenges when dealing with whole-genome sequences, for example, recombination, shuffling, and rearrangement of sequences. Thus, various alignment-free methods for phylogeny construction have been proposed. However, most of these methods have not been implemented as tools or web servers. Researchers cannot use these methods easily with their data sets. To facilitate the usage of various alignment-free methods, we implemented most of the popular alignment-free methods and constructed a user-friendly web server for alignment-free genome phylogeny (AGP). AGP integrated the phylogenetic tree construction, visualization, and comparison functions together. Both AGP and all source code of the methods are available at http://www.herbbol.org:8000/agp (last accessed February 26, 2013). AGP will facilitate research in the field of whole-genome phylogeny and comparison.

Key words: alignment-free method, genome phylogeny, tree comparison.

Introduction

Phylogenetic analysis reveals the evolutionary derivation of species. A phylogenetic tree is traditionally inferred from multiple sequence alignment of conservative proteins or genes. Alignment methods and tools have been widely used for the construction of phylogenetic trees. However, with the development of various genome sequencing projects, alignment methods meet huge challenges when dealing with whole-genome sequences. Alignment methods cannot evaluate the recombination, shuffling, and rearrangement events of the whole genomes, and whole-genome multiple alignments are computationally intensive. These obstacles for alignment-based phylogenetic reconstruction motivated several alignment-free methods in recent years. Two main categories of alignment-free methods have been proposed including methods based on word frequency and methods that do not require resolving the sequence with fixed length word. The first category includes feature frequency profile (FFP), composition vector (CV), return time distribution (RTD), and frequency chaos game representation (FCGR). FFP assembles the frequency information for all the possible words of fixed length *k* (k-mers) into an FFP vector, and the selection of word length is critical in the method (Jun et al. 2009; Sims et al. 2009; Sims and Kim 2011). CV method subtracts the random background of these frequencies using a Markov model to diminish the influence of random neutral mutations to highlight the shaping role of selective evolution and then puts these normalized frequencies in a fixed order into a CV (Qi et al. 2004; Gao and Qi 2007; Xu and Hao 2009; Yu, Liang, et al. 2010). Chaos game representation (CGR) was proposed as a scale-independent representation for genomic sequences (Jeffrey 1990). Each CGR is a unique fingerprint of the underlying sequence. However, the CGRs are not directly comparable. If the CGRs are divided by grid lines, each grid square denotes the occurrence of one pattern of k-mers in the sequence (Deschavanne et al. 1999; Almeida et al. 2001). These frequencies can be represented as FCGRs. FCGRs are numerical matrices and can be used to infer phylogenetic trees (Wang et al. 2005; Hatje and Kollmar 2012). RTD was defined as the time required for the reappearance of particular k-mers. Two statistical parameters (μ and σ) of each RTD were used to derive a feature vector (Kolekar et al. 2011, 2012). Methods that are not based on k-mers are rather heterogeneous, for example, average common subsequence (ACS), graph-based methods, the Kr estimator, and methods based on information correlation or compress. ACS calculates the pairwise genome sequence distances using average common substring at every site of each sequence (Cohen and Chor 2012). The Kr estimator designed by Haubold et al. is closely related to the ACS, which calculates the number of substitutions per site between two unaligned DNA sequences using the shortest absent substring (Domazet-Loso and Haubold 2009; Haubold et al. 2009). Graph-based methods have been used for graphical representations of DNA sequences (Gates 1985; Nandy et al. 2006). The features from graphical representations of DNA sequences have been developed to capture the essence of the base composition and distribution of the sequences in a quantitative manner. Deng et al. (2011) and Huang et al. (2011) used natural vector and a 10-dimensional statistical vector to characterize the two-dimensional (2D) graphical DNA curve, they were called two-dimensional natural vector (2DNV) and two-dimensional statistical vector (2DSV), respectively. Yu, Chu, et al. (2010) converted the 2D genome space to an *N*-dimensional moment vector (*N* equals to the length of genome, and this method was called 2DMV) and used the first *n* components of the vectors to construct phylogenetic tree for 34 lentiviruses based on whole-genome sequences. Methods based on information correlation emphasized the base correlation property of DNA sequence. Information correlation and partial information correlation (IC-PIC) and base–base correlation (BBC) were proposed to analyze the phylogenetic relationships among species (Liu et al. 2005; Liu and Sun 2008; Liu et al. 2008; Gao and Luo 2011; Liu, Zeng, Yang, Chu, et al. 2012; Liu, Zeng, Yang, Ren, et al. 2012; Zeng et al. 2012). Sequence distance measures based on information compress were proposed using Lempel–Ziv and Kolmogorov complexity (Li et al. 2001; Otu and Sayood 2003). Lempel–Ziv complexity uses the relative information between the sequences and is computationally intensive. Kolmogorov complexity can be regarded as the ultimate lower bound of all measures of information, it is a theoretical limit and cannot be computed in the general case (Li et al. 2001).

Although many alignment-free methods have been proposed, only the CV and Kr methods have been implemented as web servers, source code for FFP, and Lempel–Ziv complexity can be obtained from authors who proposed the methods. Thus, we implemented 12 popular alignment-free methods and constructed a user-friendly web platform for alignment-free genome phylogeny (AGP) research. AGP also implemented methods for phylogenetic tree visualization and comparison. AGP will facilitate the phylogenetic researches in the field of whole-genome phylogeny and comparison.

New Approaches

AGP implemented 12 alignment-free methods for the construction of phylogeny trees using whole genomes and four methods for phylogenetic tree comparison. AGP constructed the first user-friendly multimethods web server for the phylogeny analysis using alignment-free methods and whole genomes. AGP integrated functions including phylogenetic tree construction, visualization, and comparison, which is a comprehensive multimethods platform for the phylogenetic analysis of whole genomes.

Results and Discussion

There are total six pages in AGP. The home page is shown in figure 1. "METHODS" and "TREECOMPARE" pages perform the web server functions, including phylogenetic tree construction, visualization, and comparison.

Tree Construction and Visualization

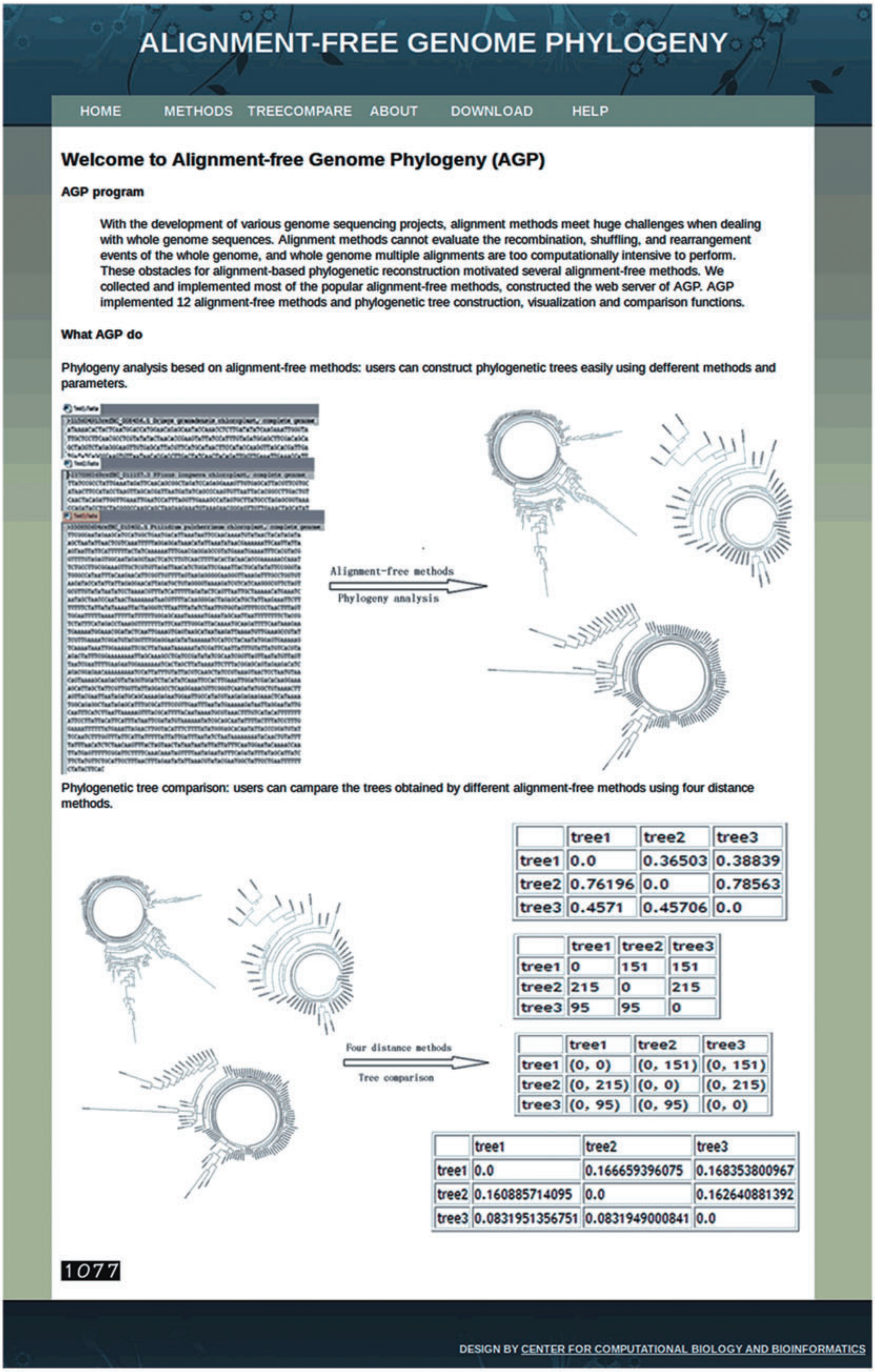
"METHODS" page gives a simple introduction of all alignment-free methods and lists every method with a hyperlink, which will lead you to its input page. Detailed information about the method and input data is described on the input page. For all methods, the input genome file must be in multi-FASTA format. For methods based on k-mers (e.g., FFP, CV, FCGR, and RTD), you need to supply the *k* value, which indicates the fixed length of word. The *k* value has influence on the results of sequence comparisons, which is determined by the length of genome sequences. For BBC, IC-PIC, and 2DMV, you also need to set the *k* parameter, which indicates the max distance between bases and the number of components of the moment vector used in the analysis. The reference range of each *k* value is also supplied to users on the input page of the method. For FCGR, RTD, 2DNV, 2DSV, 2DMV, IC-PIC, and BBC, you could select one of the 10 distance methods to calculate the distance matrix among genome sequences. To compare with the traditional alignment-based method, AGP also provided functions for the phylogeny construction based on whole-genome alignment, which was implemented using MUMmer (Kurtz et al. 2004).

When all input data have been successfully submitted, the web server will return you back the computing results. The results of phylogeny analysis contain distance matrices, phylogenetic trees, and tree maps. Two kinds of distance matrices were provided in Phylip and Nexus formats. Phylogenetic trees were formatted into standard Newick and Nexus files, which can be used for editing and viewing in other tools directly (e.g., Mega, Phylip, and TreeView). Circular and rectangular tree maps were rendered into five types of figures including TIFF, GIF, JPG, PS, and PNG (Felsenstein 1989; Page 1996; Tamura et al. 2011). All results can be viewed and downloaded online. Result page for the phylogenetic analysis of 155 complete chloroplast genomes using FFP is shown in figure 2.

Tree Comparison

When you obtain phylogenetic trees using different methods with various parameters, you can compare the differences between these trees. You need to put these trees in a plain text file. Each tree starts with the symbol ">" and the name of the tree at a new line; the following lines describe the structure of the tree. Then you can submit the file to "TREECOMPARE" page. The web server will return back the comparison results in an all-by-all distance matrix. Optionally, you can choose the distance method used for the comparison. These four distance methods were implemented for tree comparison in AGP, including RobinsonFoulds, Symmetric, FalsePositivesAndNegatives, and Euclidean.

AGP implemented 12 alignment-free methods and 10 distance methods for the construction of phylogenetic trees. The phylogenetic trees constructed were outputted as standard Newick and Nexus files and visualized as circular and traditional rectangular tree maps. To compare with the traditional alignment-based method, AGP also implemented functions for the phylogeny construction based on whole-genome alignment. Furthermore, AGP implemented four methods for the comparison of phylogenetic trees constructed in the analysis. All results can be viewed and downloaded online. AGP is the first multimethods platform for alignment-free phylogeny analysis, and it will help researchers



Welcome to Alignment-free Genome Phylogeny (AGP)

AGP program

With the development of various genome sequencing projects, alignment methods meet huge challenges when dealing with whole genome sequences. Alignment methods cannot evaluate the recombination, shuffling, and rearrangement events of the whole genome, and whole genome multiple alignments are too computationally intensive to perform. These obstacles for alignment-based phylogenetic reconstruction motivated several alignment-free methods. We collected and implemented most of the popular alignment-free methods, constructed the web server of AGP. AGP implemented 12 alignment-free methods and phylogenetic tree construction, visualization and comparison functions.

What AGP do

Phylogeny analysis besed on alignment-free methods: users can construct phylogenetic trees easily using defferent methods and parameters.

Phylogenetic tree comparison: users can campare the trees obtained by different alignment-free methods using four distance methods.

	tree1	tree2	tree3
tree1	0.0	0.36503	0.38839
tree2	0.76196	0.0	0.78563
tree3	0.4571	0.45706	0.0

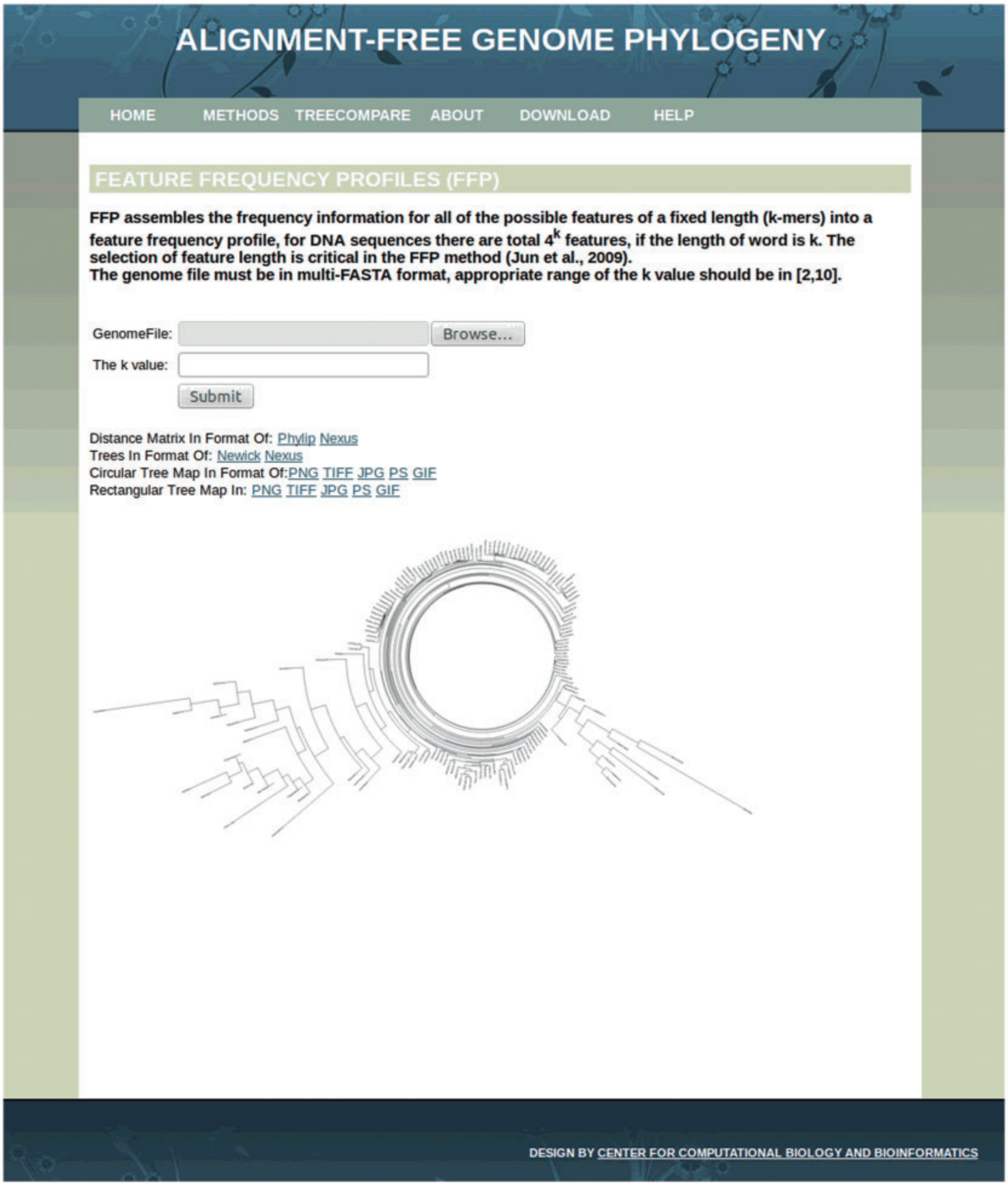
	tree1	tree2	tree3
tree1	0	151	151
tree2	215	0	215
tree3	95	95	0

	tree1	tree2	tree3
tree1	(0, 0)	(0, 151)	(0, 151)
tree2	(0, 215)	(0, 0)	(0, 215)
tree3	(0, 95)	(0, 95)	(0, 0)

	tree1	tree2	tree3
tree1	0.0	0.166659396075	0.168353800967
tree2	0.160885714095	0.0	0.162640881392
tree3	0.0831951356751	0.0831949000841	0.0

1077
DESIGN BY CENTER FOR COMPUTATIONAL BIOLOGY AND BIOINFORMATICS

FIG. 1. The home page of the AGP platform.



FEATURE FREQUENCY PROFILES (FFP)

FFP assembles the frequency information for all of the possible features of a fixed length (k-mers) into a feature frequency profile, for DNA sequences there are total 4^k features, if the length of word is k. The selection of feature length is critical in the FFP method (Jun et al., 2009).
The genome file must be in multi-FASTA format, appropriate range of the k value should be in [2,10].

GenomeFile: Browse...

The k value:

Submit

Distance Matrix In Format Of: Phylip Nexus
Trees In Format Of: Newick Nexus
Circular Tree Map In Format Of:PNG TIFF JPG PS GIF
Rectangular Tree Map In: PNG TIFF JPG PS GIF

DESIGN BY CENTER FOR COMPUTATIONAL BIOLOGY AND BIOINFORMATICS

FIG. 2. Result page for the phylogenetic analysis of 155 chloroplast genomes in AGP.

perform whole-genome phylogeny analysis and compare the results of various methods.

Methods and Implementation

AGP implemented 12 alignment-free methods including FFP, CV, FCGR, RTD, ACS, Kr, 2DNV, 2DSV, 2DMV, IC-PIC, BBC, and Lempel–Ziv. FCGR, RTD, 2DNV, 2DSV, 2DMV, IC-PIC, and BBC converted genome sequences into numerical multidimensional vectors and then used 10 kinds of methods to compute the distance matrix among the vectors, including Euclidean, Braycurtis, Canberra, Chebyshev, Cityblock, Correlation, Cosine, Minkowski, Seuclidean, and Sqeuclidean. FFP and CV represented genome sequence as a 4^k-dimension frequency vector of k-mers and calculated the distance matrix using the distance formula published in the articles (Qi et al. 2004; Sims et al. 2009). ACS, Kr, and Lempel–Ziv did not convert genome sequences into vectors, they computed pairwise genome distances directly. When we got the distance matrix among genomes analyzed, we used neighbor-joining method to construct the phylogenetic trees (Saitou and Nei 1987). AGP implemented four methods for the comparison of phylogeny trees obtained, including the following:

RobinsonFoulds: This method returns the Robinsons–Foulds distance between two trees, the sum of the square of differences in branch lengths for equivalent splits between two trees (Robinson and Foulds 1981).

Symmetric: The symmetric distance between two trees is the sum of the number of splits found in one of the trees but not the other.

FalsePositivesAndNegatives: This method returns a tuple pair, with the first element is the number of splits in the first tree but not found in the second tree compared, whereas the second element is the number of splits in the second tree, which are not in the first tree.

Euclidean: This method returns the "branch length distance" of Felsenstein (2004), the sum of absolute differences in branch lengths for equivalent splits between two trees.

We programmed methods CV, FCGR, RTD, ACS, 2DNV, 2DSV, 2DMV, IC-PIC, and BBC according to the algorithms published with the methods. All methods were implemented using the Python language. Phylogenetic tree visualization and comparison functions were implemented based on a Python environment for tree exploration (ETE) and DendroPy Python Packages (Huerta-Cepas et al. 2010; Sukumaran and Holder 2010). The web server was implemented based on the web2py framework (http://www.web2py.com/, last accessed February 26, 2013).

Acknowledgments

This work was supported by National Natural Science Foundation of China (81102746, 81100077, 31201768), Beijing Natural Science Foundation (5113033, 6133033), Scientific Research Foundation of the State Human Resource Ministry and the Education Ministry for Returned Chinese Scholars, PUMC Youth Fund, the Fundamental Research Funds for the Central Universities, Basic Scientific Research Operation Cost of State-Leveled Public Welfare Scientific Research Courtyard (YZ-12-15), Special Foundation of China Postdoctoral Science (2012T50067), China Postdoctoral Science Foundation (2011M500014, 2012M510011), the Research Fund for the Doctoral Program of Higher Education (20111106120028), "Major Drug Discovery" major science and technology research "12nd Five-Year Plan" (2012ZX09301-002-001-025), China Medical Board of New York (A2009001), and Program for New Century Excellent Talents in University (NCET-74) to Z.L.

References

Almeida JS, Carrico JA, Maretzek A, Noble PA, Fletcher M. 2001. Analysis of genomic sequences by chaos game representation. *Bioinformatics* 17:429–437.

Cohen E, Chor B. 2012. Detecting phylogenetic signals in eukaryotic whole genome sequences. *J Comput Biol.* 19:945–956.

Deng M, Yu C, Liang Q, He RL, Yau SS. 2011. A novel method of characterizing genetic sequences: genome space with biological distance and applications. *PLoS One* 6:e17293.

Deschavanne PJ, Giron A, Vilain J, Fagot G, Fertil B. 1999. Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. *Mol Biol Evol.* 16:1391–1399.

Domazet-Loso M, Haubold B. 2009. Efficient estimation of pairwise distances between genomes. *Bioinformatics* 25:3221–3227.

Felsenstein J. 1989. PHYLIP—phylogeny inference package (version 3.2). *Cladistics* 5:164–166.

Felsenstein J. 2004. Inferring phylogenies. Sunderland (MA): Sinauer Associates.

Gao L, Qi J. 2007. Whole genome molecular phylogeny of large dsDNA viruses using composition vector method. *BMC Evol Biol.* 7:41.

Gao Y, Luo L. 2011. Genome-based phylogeny of dsDNA viruses by a novel alignment-free method. *Gene* 492:309–314.

Gates MA. 1985. Simpler DNA sequence representations. *Nature* 316:219.

Hatje K, Kollmar M. 2012. A phylogenetic analysis of the brassicales clade based on an alignment-free sequence comparison method. *Front Plant Sci.* 3:192.

Haubold B, Pfaffelhuber P, Domazet-Loso M, Wiehe T. 2009. Estimating mutation distances from unaligned genomes. *J Comput Biol.* 16:1487–1500.

Huang G, Zhou H, Li Y, Xu L. 2011. Alignment-free comparison of genome sequences by a new numerical characterization. *J Theor Biol.* 281:107–112.

Huerta-Cepas J, Dopazo J, Gabaldon T. 2010. ETE: a python environment for Tree exploration. *BMC Bioinformatics* 11:24.

Jeffrey HJ. 1990. Chaos game representation of gene structure. *Nucleic Acids Res.* 18:2163–2170.

Jun SR, Sims GE, Wu GA, Kim SH. 2009. Whole-proteome phylogeny of prokaryotes by feature frequency profiles: an alignment-free method with optimal feature resolution. *Proc Natl Acad Sci U S A.* 107:133–138.

Kolekar P, Kale M, Kulkarni-Kale U. 2012. Alignment-free distance measure based on return time distribution for sequence analysis: applications to clustering, molecular phylogeny and subtyping. *Mol Phylogenet Evol.* 65:510–522.

Kolekar PS, Kale M, Kulkarni-Kale U. 2011. Genotyping of Mumps viruses based on SH gene: development of a server using alignment-free and alignment-based methods. *Immunome Res.* 7:1–7.

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. 2004. Versatile and open software for comparing large genomes. *Genome Biol.* 5:R12.

Li M, Badger JH, Chen X, Kwong S, Kearney P, Zhang H. 2001. An information-based sequence distance and its application to whole mitochondrial genome phylogeny. *Bioinformatics* 17:149–154.

Liu ZH, Jiao D, Sun X. 2005. Classifying genomic sequences by sequence feature analysis. *Genomics Proteomics Bioinform.* 3(4):201–205.

Liu ZH, Meng JH, Sun X. 2008. A novel feature-based method for whole genome phylogenetic analysis without alignment: application to HEV genotyping and subtyping. *Biochem Biophys Res Commun.* 368:223–230.

Liu ZH, Sun X. 2008. Coronavirus phylogeny based on base-base correlation. *Int J Bioinform Res Appl.* 4:211–220.

Liu ZH, Zeng X, Yang D, Ren GM, Chu GY, Yuan ZR, Luo K, Xiao PG, Chen SL. 2012. Identification of medicinal vines by ITS2 using complementary discrimination methods. *J Ethnopharmacol.* 141:242–249.

Liu ZH, Zeng X, Yang D, Chu GY, Yuan ZR, Chen SL. 2012. Applying DNA barcodes for identification of plant species in the family Araliaceae. *Gene* 499:76–80.

Nandy A, Harle MS, Basak C. 2006. Mathematical descriptors of DNA sequences: development and applications. *Arch Org Chem.* 9:211–238.

Otu HH, Sayood K. 2003. A new sequence distance measure for phylogenetic tree construction. *Bioinformatics* 19:2122–2130.

Page RD. 1996. TreeView: an application to display phylogenetic trees on personal computers. *Comput Appl Biosci.* 12:357–358.

Qi J, Luo H, Hao B. 2004. CVTree: a phylogenetic tree reconstruction tool based on whole genomes. *Nucleic Acids Res.* 32:W45–W47.

Robinson DF, Foulds LR. 1981. Comparison of phylogenetic trees. *Math Biosci.* 53:131–147.

Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol.* 4:406–425.

Sims GE, Jun SR, Wu GA, Kim SH. 2009. Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. *Proc Natl Acad Sci U S A.* 106:2677–2682.

Sims GE, Kim SH. 2011. Whole-genome phylogeny of *Escherichia coli/Shigella* group by feature frequency profiles (FFPs). *Proc Natl Acad Sci U S A.* 108:8329–8334.

Sukumaran J, Holder MT. 2010. DendroPy: a Python library for phylogenetic computing. *Bioinformatics* 26:1569–1571.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol.* 28:2731–2739.

Wang Y, Hill K, Singh S, Kari L. 2005. The spectrum of genomic signatures: from dinucleotides to chaos game representation. *Gene* 346:173–185.

Xu Z, Hao B. 2009. CVTree update: a newly designed phylogenetic study platform using composition vectors and whole genomes. *Nucleic Acids Res.* 37:W174–W178.

Yu C, Liang Q, Yin C, He RL, Yau SS. 2010. A novel construction of genome space with biological geometry. *DNA Res.* 17:155–168.

Yu ZG, Chu KH, Li CP, Anh V, Zhou LQ, Wang RW. 2010. Whole-proteome phylogeny of large dsDNA viruses and parvoviruses through a composition vector method related to dynamical language model. *BMC Evol Biol.* 10:192.

Zeng X, Yuan ZR, Tong X, Li QS, Gao WW, Qin MJ, Liu ZH. 2012. Phylogenetic study of Oryzoideae species and related taxa of the Poaceae based on atpB-rbcL and ndhF DNA sequences. *Mol Biol Rep.* 39(5):5737–5744.