Journal of Biotechnology 149 (2010) 115–119

Contents lists available at ScienceDirect

Journal of Biotechnology

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

Composition vector approach to whole-genome-based prokaryotic phylogeny: Success and foundations

Qiang Li[a,1], Zhao Xu[a], Bailin Hao[a,b,c,*]

[a] T-Life Research Center, Fudan University, Shanghai 200433, China
[b] Institute of Theoretical Physics, Academia Sinica, Beijing 100190, China
[c] Santa Fe Institute, Santa Fe, NM 87501, USA

ARTICLE INFO

Article history:
Received 15 August 2009
Received in revised form 8 December 2009
Accepted 21 December 2009

Keywords:
Prokaryote phylogeny
Alignment-free
Composition vector
Whole-genome phylogeny
CVTree

ABSTRACT

Composition vector approach to prokaryotic phylogeny provides an alignment-free and parameter-free method based on whole-genome data. It has also been applied to viruses and fungi. In all studied cases the inferred phylogenetic relationships agree well with taxonomic knowledge in major groupings and fine branchings. In this review article, after demonstrating its successful application to a collection of 892 genomes including 62 Archaea, 822 Bacteria and 8 Eukarya, we will outline some ongoing work towards the foundations of this new approach.

1. Introduction

Being closely related topics, taxonomy and phylogeny are not synonyms. Taxonomists came much earlier, having established their own rules and codes, judicial commissions and validation publications. Phylogeny, especially molecular phylogeny, though appeared later, has also become an established discipline. At present, both the phylogeny and taxonomy of prokaryotes are more and more based on the 16S rRNA analysis introduced by Woese and coworkers in the late-1970s (Woese and Fox, 1977), complemented by multi-alignments of protein-coding or other RNA genes. Many questions have been raised towards the traditional approach. For example, how faithful a phylogenetic tree inferred from a single or a few genes may represent the interrelationships of species? Though lateral transfer of rRNA genes has not been found to exist in nature, ribosomal operons in one genome have been replaced by those of another species in laboratory (Asai et al., 1999). The problem of lateral gene transfer (LGT) becomes even more severe when the phylogeny is based on protein-coding genes. Furthermore, rRNA analysis lacks taxonomic resolution at the level of species (Staley, 2006; Yarza et al., 2008), not to mention its incapability to be applied to viruses.

In order to overcome difficulties caused by using a single or a few genes in phylogenetic reconstructions many whole-genome-based methods have been proposed, see, e.g., reviews (Philippe et al., 2005; Snel et al., 2005). However, most of these methods, if not all, depend on sequence alignment at some stage, which, in turn, invokes many parameters via scoring matrices and gap penalties. Furthermore, the justification of both single or few gene based as well as whole-genome-based phylogenies rely on self-consistency and stability arguments by statistical re-sampling tests such as bootstrap and jackknife. It was not long ago when whole-genome phylogeny could "not resolve the major branchings of the Bacteria" (Hyunen et al., 1999) and "it may not be possible to reconstruct the eubacterial phylogeny reliably by standard methods" (Teichmann and Mitchison, 1999). In view of this situation the parameter-free and whole-genome-based composition vector method to reconstruct phylogeny (Qi et al., 2004b; Hao and Qi, 2004) really made a step forward. The composition vector approach has been applied to the phylogeny of viruses (Gao et al., 2003; Gao and Qi, 2007), prokaryotes (Qi et al., 2004b; Hao and Qi, 2004; Gao et al., 2007; Sun et al., in press), chloroplasts (Chu et al., 2004), and fungi (Wang et al., 2009). In all these cases the phylogenies agree with the corresponding taxonomy in major branchings of the tree. However, in spite of the success, questions may be raised regarding the foundation of the new approach, see, e.g., the comments in the review

* Corresponding author at: T-Life Research Center, Department of Physics, Fudan University, 220 Handan Road, Shanghai 200433, Shanghai, China.
Tel.: +86 21 6565 2305; fax: +86 21 6565 2305.
E-mail address: hao@mail.itp.ac.cn (B. Hao).
[1] Present address: CAS-MPG Partner Institute for Computational Biology, Shanghai 200031, China.

0168-1656/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.jbiotec.2009.12.015

by Snel et al. (2005). In this paper we will touch on some of these questions.

2. Materials and methods

2.1. Whole-genome data

We use all the translated protein products in a genome from the National Center for Biotechnological Information FTP site (NCBI, 2009). These are the .faa files with accession numbers preceded by NC_, released by NCBI for the RefSeq database. In fact, our new CVTree Web Server (Xu and Hao, 2009) automatically updates these files from NCBI at the beginning of each month. Full binomina with strain tags, identical to the subdirectory names at NCBI, are used to label the leaves in a tree since abbreviations have become inconvenient when the total number of genomes goes into hundreds and thousands. The analysis performed in this study is based on a collection of 62 Archaea and 822 Bacteria genomes available at NCBI on 31 May 2009, using 8 Eukarya as outgroups. We did not include two bacterial endosymbionts *Carsonella ruddii* (Nakabachi et al., 2006) and *Sulcia Muelleri* (McCutcheon and Moran, 2007) in the analysis, because their genome size and number of genes are much less than all those of known free-living bacteria (see, e.g., Coffeau, 1995). A list of genomes with full organism names and accession numbers is given in the Additional Material for online publication of this paper.

2.2. The CVTree method

CVTree stands for Composition Vector Tree. It is used as the name of the method and the trees obtained by this method. As the CVTree method has been described before (Qi et al., 2004b; Hao and Qi, 2004), we summarize it briefly. For a fixed integer K we count the number of overlapping K-peptides in the collection of all protein products of a genome and form a raw composition vector (CV) of dimension 20^K, allocating the counts in lexicographic order of the amino acid characters. Then a "normalized" CV is obtained by subtracting a predicted count from the real count by making a $(K-2)$-th order Markovian assumption. This subtraction procedure suppresses the effect of random background and highlights the phylogenetic information contained in the normalized CV (Hao and Qi, 2004). The pairwise correlations between CVs are used to generate a distance matrix. Phylogenetic trees are constructed by using the standard Neighbor-Joining (Saitou and Nei, 1987) method implemented in the Phylip (Felsenstein, 2008) package. To facilitate the application of the new method a CVTree web server was published in 2004 (Qi et al., 2004a) and a significant update of this server has just appeared (Xu and Hao, 2009).

2.3. Taxonomic references and the TOBA code

Although there is no unanimously accepted standard prokaryotic taxonomy many bacteriologists take the new edition of the Bergey's Manual of Systematic Bacteriology (Bergey's Manual Trust, 2001) and the closely related online Taxonomic Outline of Bacteria and Archaea (Garrity et al., 2007, referred to as TOBA hereafter) as a good approximation to such a standard (Konstantinidis and Tiedje, 2005). Therefore, we mainly compare the CVTree results with TOBA. When an item is not listed in TOBA we occasionally refer to the NCBI TaxBrowser, which, in fact, is more dynamic and up-to-date though disclaimed to be a taxonomic reference. The Bergey's Manual contains a numerical ordering of taxa together with the Latin names while TOBA uses the latter only. In order to facilitate computer work we have generated a numbering for all the TOBA taxa (available on request to the corresponding author) which basically coincides with the Bergey's ordering. For example, *Escherichia* is the 1st genus in the only family Enterobacteriaceae of the 13th order Enterobacteriales in the 3rd class Gammaproteobacteria of the 12th bacterial phylum Proteobacteria. We introduce a shorthand B12.3.13=1.1 for this lineage where an equal sign "=" is used when an upper taxon contains only one lower taxon. This is called a TOBA code. When a species is not listed in TOBA we assign a tentative code according to the NCBI lineage. These codes are given in the Additional Material for the online publication together with the organism names and accession numbers.

3. Results

In order to compare the CVTree results with prokaryotic taxonomy we proceed from the two extremes: the branchings corresponding to the highest taxonomic ranks and the grouping of species and strains at the lowest rank of genus. When making such analysis we pay special attention to the monophyleticity of the branches. Whenever all leaves in a monophyletic branch come from one and the same taxonomic unit the branch may be "collapsed" and labeled by that taxonomic name.

Fig. 1 is an unrooted CVTree with monophyletic branches corresponding to the highest taxonomic ranks. First of all, the three main domains of life, the Archaea (62), Bacteria, and Eukarya (8), are separated (numerals given in parentheses indicate the number of organisms in a monophyletic branch). Among the 21 bacterial phyla represented by the 822 genomes fifteen phyla do form mono-

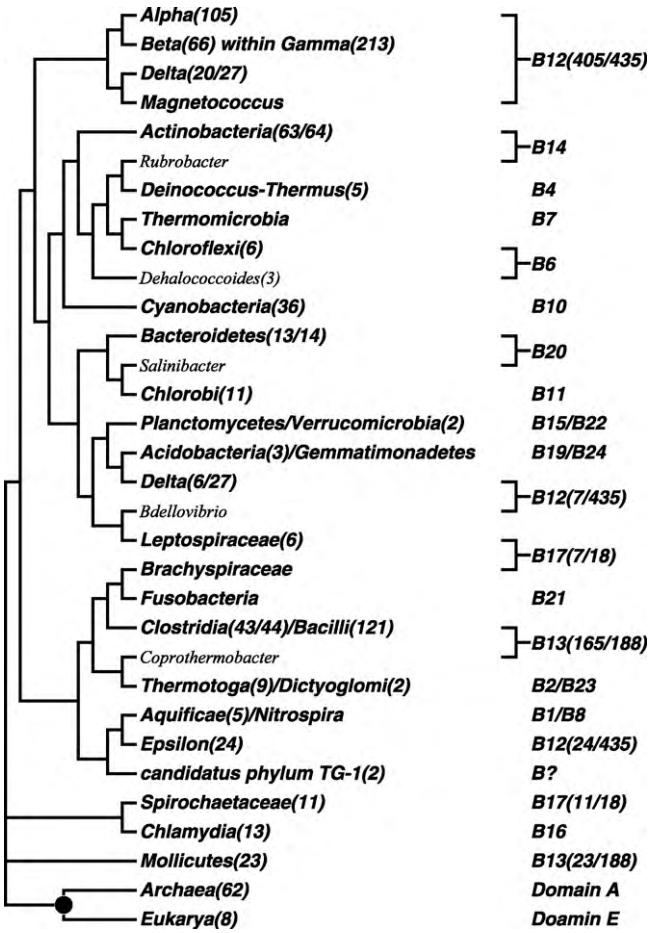

Fig. 1. An unrooted CVTree at $K=6$. Monophyletic branches at the highest taxonomic ranks are shown. Numerals in parentheses give the number or proportion of genomes present in a branch (omitted when it is 1). Outliers are shown in thin font. Listed on the right are the TOBA phylum numbers. Domain A is Archaea. Domian E is Eukarya.

phyletic branches, including five phyla containing only one species. It is a remarkable fact that at the phylum level there are only five outliers, shown in thin font in Fig. 1 (the 3 strains of *Dehalococcoides* are taken as one outlier). Not being appropriate to get into a detailed comparison of phylogeny with taxonomy in this review article, we skip a discussion of the outliers and notice that essentially there are only three phyla that do not form monophyletic clusters.

First, 435 genomes come under a single phylum Proteobacteria which contains 5 classes (groups). Of these 5 classes three do form monophyletic branches in CVTrees: Alphaproteobacteria (106), Betaproteobacteria (66), and Epsilonproteobacteria (24). Furthermore, It was observed in 16S rRNA analysis that the Beta group gets inserted into the Gamma group (Woese et al., 2000). So happens in CVTrees: the combined Beta/Gamma group forms a monophyletic branch comprising 66 + 213 = 279 genomes. The only non-monophyletic class Deltaproteobacteria splits into 3 clusters. The largest cluster labeled as Delta (20/27) in Fig. 1, comprising 20 of the 27 genomes listed under this class, remains in the greater monophyletic branch containing a predominant majority (405/435) of the Proteobacteria. The second cluster Delta (6/27), separated from the main body of Proteobacteria, consists of species from Mycococcales, which has once been classified into a "phylum" together with *Bdellovibrio*, another outlier in Fig. 1, see, e.g., Woese et al. (1985).

Second, two of the three classes of Firmicutes do form a monophyletic group, leaving out the class Mollicutes (23) which makes a separate phylum Terenicutes in NCBI taxonomy. Recently, the class Mollicutes has been removed from the phylum Firmicutes in the new edition of the Bergey's Manual of Systematic Bacteriology (Ludwig et al., 2009).

Third, the phylum Spirochaetes was defined to a large extent by morphological criteria (Woese et al., 1985). It contains a single class which in turn contains a single order. Therefore, only the classification into families makes sense. Fig. 1 shows that the three families of Spirochaetes do not join together to form a monophyletic branch.

At the lowest taxonomic rank we note that the 884 genomes studied in this work comprise 290 genera of which 121 contain two or more species/strains. 95 out of these 121 genera agree with the monophyletic branchings in CVtrees. Most of the 26 genera in disagreement have been known and debated by biologists for years and may hint on necessary taxonomic revisions. In the last section of Additional Material for online publication we list the "convergence" of all genera with K changing from 3 to 7.

We note also that a number of previous predictions which first appeared as "disagreement" with taxonomy were confirmed in further releases of the latter. For example, the genus *Oceanobacillus* was assigned to the phylum Proteobacteria (B12) in Bergey's Outline Rel. 3.0 (Garrity et al., 2002), but moved to phylum Firmicutes (B13) in Rel. 4.0 (Garrity et al., 2003), while on all CVTrees it joins class Bacilli in B13 ever since the genome first appeared in Qi et al. (2004b). In Bergey's Outline Rel. 5.0 (Garrity et al., 2004) the species *Thiomicrospira denitrificans* was listed in class Gammaproteobacteria with a footnote saying "The identity of *T. denitrificans* is questionable as it belongs within the Epsilonproteobacteria". All CVTrees confirm this footnote and in TOBA 7.7 (Garrity et al., 2007) the species was renamed *Sulfurimonas denitrificans* within Epsilonproteobacteria. We omit other examples of retrospective verification of CVTree predictions.

In view of the overwhelming agreement of CVtree branchings with taxonomy the few discrepancies should be taken seriously as hints to taxonomic revisions.

4. Discussion

In the previous section we briefly summarized the success of the CVtree approach. However, as a newly proposed method many questions concerning the foundation of this approach remain to be investigated. Not trying to answer all these questions at present time, in what follows we describe a few ongoing work along this line.

4.1. The protein sequence decomposition and reconstruction problem

Given a primary protein sequence of length L and an integer K, decompose the sequence into $L-K+1$ overlapping K-peptides by shifting one letter at a time. Is an amino acid sequence reconstructible from this collection of K-peptides? This problem is solvable because at least the original sequence should be recovered. The uniqueness problem is more interesting and it has a natural relation to the number of Eulerian loops in a graph. It may also be set in the framework of so-called factorizable language, a kind of formal languages (Shi et al., 2007; Li and Xie, 2008; Hao and Xie, 2008). In fact, most of natural proteins do have a unique reconstruction at $K = 5$ or 6 (Xia and Zhou, 2007). This result speaks in favor of using K-peptides instead of primary protein sequences when whole-genome alignment becomes problematic

4.2. The calibration of branch lengths in CVTree

All CVTrees published so far are unrooted trees with "good" topology in the sense of their agreement with taxonomy. However, proper calibration of the branch lengths remains a problem. Quite recently, we succeeded in calibrating the branch lengths and they have been related to the simple p-distance, i.e., the proportion of unmatched sites between two aligned sequences (Li, 2009). The p-distance is a basic genetic distance from which many other more elaborated distances may be obtained or approximated (Nei and Kumar, 2000). A detailed elucidation of these relations will be given in a separate publication.

4.3. The choice of suitable peptide length K

Though K looks like a parameter, there is no need to adjust K in CVTree calculations. Actually, by inspecting trees obtained for various $K \geq 3$, one gains additional knowledge on convergence of the phylogeny. The lower bound $K = 3$ is imposed by the use of a $(K-2)$-th order Markovian assumption in the algorithm (Qi et al., 2004b). Our results showed that the topology of the $K = 4$ tree is better than that of $K = 3$. The $K = 5$ and 6 trees are the "best" ones, and at $K = 7$ the agreement with taxonomy becomes slightly worse. This observation may be explained by the following order-of-magnitude analysis.

Denote the alphabet size of amino acid letters by $|\Sigma| = 20$. A random peptide of length K has a probability of appearance $|\sum|^{-K}$, its expected count in a collection of proteins of total length L is $L|\sum|^{-K}$. In order to ensure that a K-peptide count reflects species-specificity it must be significantly small than that of a random K-tuple, i.e., $L|\sum|^{-K} \ll 1$. For prokaryote genomes we may take $L \approx 10^6$ and thus get $K > 4.6$.

On the other hand, in the subtraction procedure used in Qi et al. (2004b) it is crucial that the number of $(K-2)$-tuples should not be too small, requiring $L|\sum|^{K-2} \gg 1$ which yields $K < 6.6$.

Furthermore, we may consider the variation of the total number of existing K-peptides with K in a given genome. For K small this number is limited by 20^K, independent of the proteome size. When K gets greater, it is limited by a linearly decreasing function $L - M(K-1)$, where L, as before, is the total length of the protein sequences and M is the number of such sequences. Fig. 2 shows how the total number of K-peptides varies with K for several genomes.

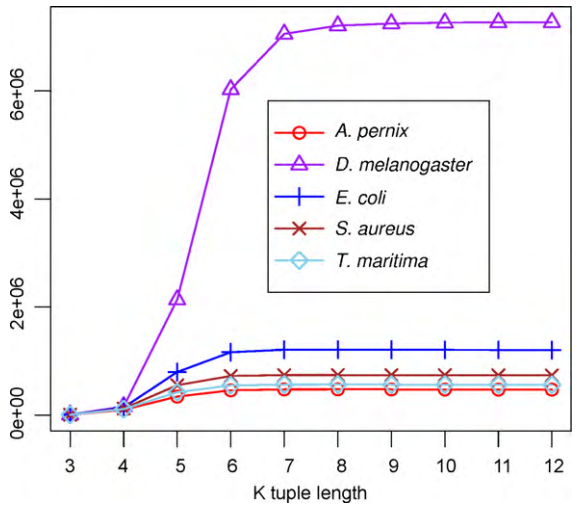

Fig. 2. The total number of nonzero *K*-peptides versus *K* for the fruitfly and four prokaryotic genomes.

Indeed, it gets saturated when $K > 6$ and $K = 6$ is enough for our purpose.

4.4. Distance versus dissimilarity

The distance between organisms is defined via correlations between CVs (Qi et al., 2004b). However, this definition does not guarantee that all the triangular inequalities hold as required by the distance axioms. For example, for the 892 genomes used to produce the results analyzed in this study, there are $892 \times 891 \times 890/6 = 117{,}891{,}180$ triples. The number of triples that violate the triangular inequality is 4550, 166, 0, 0, and 4, for $K = 3$, 4, 5, 6, and 7, respectively. We see that there is no violation at all for $K = 5$ and 6 that produce the "best" tree. Even at $K = 3$, the 4550 violations only make a tiny share in the total number of triples. We note that most violations occur among closely related genomes and there is no evident correlation between violated triangles and species "misplacement" in the CVtree. This point is being further investigated.

4.5. Statistical tests: bootstrap and jackknife

As described in the previous sections, in order to evaluate the inferred trees we have adopted a viewpoint different from that in the traditional molecular phylogeny. Being a kind of theoretical construction based on a unified data type (genomes) the CVTree results are evaluated by direct comparison with taxonomy as collection of experimental facts. This is made possible by the achievement of taxonomy as well as by the high resolution power of CVTrees. Nevertheless, we have also performed statistical tests by using bootstrapping (Qi et al., 2004b; Wang et al., 2009) and jackknifing (ongoing work). On the other hand, if one increases the number of genomes and calculates the larger and larger CVTrees, this process can be interpreted as an "anti-jackknife" test. Starting from the109-genome tree (Qi et al., 2004b), our analyses of the 440-genome tree (Gao et al., 2007) and the present study based on 892 genomes provide a successful example of passing the anti-jackknife test. An elaborated exploration of these statistical tests for CVTrees will make the subject of another publication.

4.6. Evolutionary model underlying CVTree

In principle, one can envision a "*K*-peptide picture of protein evolution" (Hao and Qi, 2004) to justify the CVTree approach. Suppose one could observe proteins present in the primordial "soup" of microbial organisms without knowing the coding, transcription and translation machinery involving nucleotide sequences. The collection of proteins must be less diversified than one sees nowadays. If one could collect all *K*-peptides, say, for $K = 5$, from these ancestral microbes, they must have taken a small part of the 20^5 points of the "5-string space". These polypeptides then evolved by growth, fission, fusion, mutation, and rearrangement. All the new pentapeptides generated by these mechanisms must be somehow related to those existed before, but the peptide composition diverges with time. Therefore, it makes sense to directly model the evolution of CVs, though the realization of this model may be extremely difficult. For example, to model mutations at the amino acid level one should invoke something similar to the 20 by 20 scoring matrices used in protein sequence alignment. We are working on this evolution model in order to test CVTree by computer simulation.

With rapid advance of new sequencing technology and reliable automatic annotation, many bacterial genomes are becoming available, including those which are not cultivable at present. The cost of sequencing a genome may decrease drastically. Therefore, in not-too-distant future a microbiologist may first submit a genome to CVTree to see where an organism goes before designing and performing more specific phenotyping experiments. In this sense CVTree may add a helpful and definitive tool in prokaryotic phylogeny and taxonomy.

Acknowledgements

This work was supported by the National Basic Research Program of China (973 Program 2007CB814800) and the Shanghai Leading Academic Discipline Project No. B111.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jbiotec.2009.12.015.

References

Asai, T., Zaporojets, D., Squires, C., Squires, C.L., 1999. An *Escherichia coli* strain with all chromosomal rRNA operons inactivated: complete exchange of rRNA genes between bacteria. Proc. Natl. Acad. Sci. U.S.A. 96, 1971–1976.

Bergey's Manual Trust, 2001–2009. Bergey's Manual of Systematic Bacteriology, vol. 1–5, 2nd ed. Springer-Verlag, New York.

Chu, K.H., Qi, J., Yu, Z.G., Anh, V., 2004. Origin and phylogeny of chloroplasts revealed by a simple correlation analysis of complete genomes. Mol. Biol. Evol. 28, 70–76.

Coffeau, A., 1995. Life with 482 genes. Science 276, 445–446.

Felsenstein, J., 2008. PHYLIP (Phylogeny Inference Package) ver. 3.68. Available from http://evolution.genetics.washington.edu/phylip.html.

Gao, L., Qi, J., Sun, J.D., Hao, B.L., 2007. Prokaryote phylogeny meets taxonomy: an exhaustive comparison of composition vector trees with systematic bacteriology. Sci. Chin. Ser. C-Life Sci. 50, 587–599.

Gao, L., Qi, J., Wei, H.B., Sun, Y.G., Hao, B.L., 2003. Molecular phylogeny of coronaviruses including human SARS-CoV. Chin. Sci. Bull. 48, 1170–1174.

Gao, L., Qi, J., 2007. Whole genome molecular phylogeny of large dsDNA viruses using composition vector method. BMC Evol. Biol. 7, 41, doi:10.1086/1471-2148/7/41.

Garrity, G.M., Johnson, K.L., Bell, J.A., Searles, D.B., 2002. Taxonomic Outline of the Procaryotes, Bergey's Manual of Systematic Bacteriology, Rel. 3.0, 2nd Ed. Springer, New York, doi:10.1007/bergeysoutline200210.

Garrity, G.M., Bell, J.A., Lilburn, T.G., 2003. Taxonomic Outline of the Procaryotes, Bergey's Manual of Systematic Bacteriology, Rel. 4.0, 2nd Ed. Springer, New York, doi:10.1007/bergeysoutline200310.

Garrity, G.M., Bell, J.A., Lilburn, T.G., 2004. Taxonomic Outline of the Procaryotes, Bergey's Manual of Systematic Bacteriology, Rel. 5.0, 2nd Ed. Springer, New York, doi:10.1007/bergeysoutline200405.

Garrity, G.M., Lilburn, T.G., Cole, J.R., Harrison, S.H., Euzéby, J., Tindall, B.J., 2007. Taxonomic Outline of Bacteria and Archaea (TOBA), Rel.7.7, Michigan State University, available from www.taxonomicoutline.org.

Hao, B.L., Qi, J., 2004. Prokaryote phylogeny without sequence alignment: from avoidance signature to composition distance. J. Bioinform. Comput. Biol. 2, 1–19.

Hao, B.L., Xie, H.M., 2008. Factorizable language: from dynamics to biology. In: Schuster, H.G. (Ed.), Chapter 5 in Reviews of Nonlinear Science and Complexity. Wiley-VCH, Weinheim.

Hyunen, M., Snel, B., Bork, P., 1999. Lateral gene transfer, genome surveys, and the phylogeny of prokaryotes. Science 286, 1443a.
Konstantinidis, K.T., Tiedje, K.V., 2005. Towards a genome-based taxonomy for prokaryotes. J. Bacteriol. 187, 6258–6264.
Li, Q., 2009. A heuristic evolutionary model for K-string composition and the problem of unique reconstruction of sequences. Ph.D. Thesis. Fudan University, Shanghai (in Chinese).
Li, Q., Xie, H.M., 2008. Finite automaton for testing composition-based reconstructibility of sequences. J. Comput. Syst. Sci. 74, 870–874.
Ludwig, W., Schleifer, K.-H., Whitman, W.B., 2009. Revised road map to the phylum Firmicutes. In: Bergey's Manual of Systematic Bacteriology, vol. 3 (The Firmicutes), 2nd Ed. Springer Verlag, New York.
McCutcheon, J.P., Moran, N.A., 2007. Parallel genomic evolution and metabolic interdependence in an encient symbiosis. Proc. Natl. Acad. Sci. U.S.A. 104, 19392–19397.
Nakabachi, A., Yamashita, A., Toh, H., Ishikawa, H., Dunbar, H.E., Moran, N.A., Hattori, M., 2006. The 160-kilobase genome of the bacterial endosymbiont *Carsonella*. Science 31, 267.
NCBI, 2009. The NCBI FTP site: ftp://ftp.ncbi.nih.gov/genomes/Bacteria/.
Nei, M., Kumar, S., 2000. Molecular Evolution and Phylogenetics. Oxford University Press, New York.
Philippe, H., Delsuc, F., Brinkmann, H., Lartillot, N., 2005. Phylogenomics. Annu. Rev. Ecol. Evol. Syst. 76, 541–562.
Qi, J., Luo, H., Hao, B.L., 2004a. CVTree: a phylogenetic tree reconstruction tool based on whole genomes. Nucl. Acids Res. 32, Web Server Issue, W45–W47.
Qi, J., Wang, B., Hao, B.L., 2004b. Whole-proteome prokaryote phylogeny without sequence alignment: a K-string composition approach. J. Mol. Evol. 58, 1–11.
Saitou, N., Nei, M., 1987. The neighbor-joining method: a new method for reconstructing phylogenetic tress. Mol. Biol. Evol. 4, 406–425.
Shi, X.L., Xie, H.M., Zhang, S.Y., Hao, B.L., 2007. Decomposition and reconstruction of protein sequences: the problem of uniqueness and factorizable language. J. Korean Phys. Soc. 50, 118–123.
Snel, B., Huynen, M.A., Dutilh, B.E., 2005. Genome trees and the nature of genome evolution. Annu. Rev. Microbiol. 59, 191–209.
Staley, J.T., 2006. The bacterial species dilemma and the genomic-phylogenetic species concept. Phil. Trans. R. Soc. B 361, 1899–1909.
Sun, J.D., Xu, Z., Hao, B.L., in press. Whole-genome based archaea phylogeny and taxonomy: a composition vector approach. Chin. Sci. Bull., doi:10.1007/s11434-010-0008-7.
Teichmann, A.A., Mitchison, G., 1999. Is there a phylogenetic signal in prokaryote proteins? J. Mol. Evol. 49, 98–107.
Wang, H., Xu, Z., Hao, B.L., 2009. A fungal phylogeny based on 82 complete genomes using the composition vector method. BMC Evol. Biol. 9, 195.
Woese, C.R., Fox, G.E., 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc. Natl. Acad. Sci. U.S.A. 74, 5088–5090.
Woese, C.R., Olsen, G.J., Ibba, M., Söll, D., 2000. Aminoacyl-tRNA synthetases, the genetic code, and the evolutionary process. Microbiol. Mol. Biol. Rev. 64, 202–236.
Woese, C.R., Stackebrandt, E., Macke, T.J., Fox, G.E., 1985. A phylogenetic definition of the major eubacterial taxa. System. Appl. Microbiol. 6, 143–151.
Xia, L., Zhou, C., 2007. Phase transition in sequence unique reconstruction. J. Syst. Sci. Compl. 20, 18–29.
Xu, Z., Hao, B.L., 2009. CVTree update: a phylogenetic tree reconstruction tool based on whole genomes. Nucl. Acids Res. 37, Web Server Issue, W174–178, published online April 26, doi:10.1093/nar/gkp278.
Yarza, P., Richter, M., Peplies, J., Euzeby, J., Amann, R., Schleifer, K.-H., Ludwig, W., Glöckner, F.O., Rosselló-Móra, R., 2008. The All-Species Living Tree project: a 16S rRNA-based phylogenetic tree of all sequenced type strains. Syst. Appl. Microbiol. 31, 241–250.