Systematic and Applied Microbiology 33 (2010) 175–182

Contents lists available at ScienceDirect

Systematic and Applied Microbiology

journal homepage: www.elsevier.de/syapm

Minireview

En route to a genome-based classification of *Archaea* and *Bacteria*?

H.-P. Klenk, M. Göker*

DSMZ – German Collection of Microorganisms and Cell Cultures, Inhoffenstraße 7B, 38124 Braunschweig, Germany

ARTICLE INFO

Article history:
Received 6 October 2009

Keywords:
Gene content
Phylogenetic inference
Supermatrix
Lateral gene transfer
Taxonomy
Whole-genome sequencing

ABSTRACT

Given the considerable promise whole-genome sequencing offers for phylogeny and classification, it is surprising that microbial systematics and genomics have not yet been reconciled. This might be due to the intrinsic difficulties in inferring reasonable phylogenies from genomic sequences, particularly in the light of the significant amount of lateral gene transfer in prokaryotic genomes. However, recent studies indicate that the species tree and the hierarchical classification based on it are still meaningful concepts, and that state-of-the-art phylogenetic inference methods are able to provide reliable estimates of the species tree to the benefit of taxonomy. Conversely, we suspect that the current lack of completely sequenced genomes for many of the major lineages of prokaryotes and for most type strains is a major obstacle in progress towards a genome-based classification of microorganisms. We conclude that phylogeny-driven microbial genome sequencing projects such as the Genomic Encyclopaedia of *Archaea* and *Bacteria* (GEBA) project are likely to rectify this situation.

1. Introduction

Major changes in the way microorganisms are differentiated are often a response to the availability of new technologies (e.g. genetic methods [122,136]). This pattern began as early as the mid 17th century, when Antoni van Leeuwenhook first used the newly invented microscope to observe bacteria, and was followed by the invention of colour staining of bacteria in the 19th century [56], and the classical methods of chemotaxonomy in the mid 20th century [22,89,121]. With the invention of DNA–DNA hybridization [28], 16S rDNA cataloguing [46], and DNA-sequencing [118] in the 1970s, analytical techniques irreversibly moved towards a genetic basis. It is not an exaggeration to state that DNA sequencing has been the key technology for biology over the last 30 years, and will be for the foreseeable future. Enabled by the automation of Sanger sequencing (in the 1990s) and the currently overtaking pyrosequencing methods [36], such technical progress is making genome sequencing a routine analytical method for microbial taxonomists. Techniques based on single-molecule sequencing, as well as nanosequencers, are expected to become available in the near future and even to speed up this development. In particular, the availability of whole-genome sequences has drastically changed the way microbiologists analyse their study objects, from the analysis of single genes or features to a global and integrated analysis of complex pathways and even whole organisms, often termed the '-omics revolution' [96]. However, to date, microbial taxonomy has barely taken the wealth of information contained in completely sequenced genomes into account.

In order to discuss the future impact of genome sequencing on classification, appropriate brief definitions of 'classification' and some other main terms are needed, as well as a clarification of their interrelationships. To quote a modern microbiological textbook: 'Taxonomy can be clearly defined as encompassing characterization, classification, and nomenclature. [...] Classification is the arrangement of prokaryotes into groups' [138]. It is now widely agreed upon that classification must aim at a natural system in accordance with the branching order imposed by the course of evolution [59,60,75]. This view can be traced back to no less a person than Charles Darwin [52], who famously stated that 'our classifications will come to be, as far as they can be so made, genealogies' [25]. For the majority of researchers (apparently including Darwin [53]) this also implied that only monophyletic taxa should be accepted (at least above species rank; see below) [38–41], whereas the school of 'evolutionary taxonomy' preferred paraphyletic groups in some situations [66,127,100]. In contrast, the 'phenetic' approach was based on the assertion that it is unlikely that evolution can be reconstructed with certainty and, hence, organisms should be classified using overall similarity and clustering analyses [124,126–128]. However, this view has fallen into disregard [40,66,73], even though many methods introduced by pheneticists are still in use in other areas [87]. As a consequence, trees obtained using appropriate phylogenetic inference methods [43] (as opposed to clustering techniques yielding ultrametric dendrograms [124,126,128]) have become essential for the classi-

Abbreviations: DDH, DNA–DNA hybridization; GEBA, Genomic Encyclopaedia of *Archaea* and *Bacteria*; LGT, Lateral Gene Transfer.

* Corresponding author. Tel.: +49 0 531 2616 272; fax: +49 0 531 2616 237.
E-mail address: markus.goeker@dsmz.de (M. Göker).

0723-2020/$ – see front matter © 2010 Elsevier GmbH. All rights reserved.
doi:10.1016/j.syapm.2010.03.003

fication of organisms because they are needed to assess monophyly [38].

However, the relationship between classification and phylogenetic reconstruction is not as obvious as it might seem at first sight. In a section entitled 'The Irrelevance of Classification' on p. 45 of his textbook [43], the prominent phylogeneticist Joseph Felsenstein expresses the view that the 'delimitation of higher taxa is no longer a major task of systematics, as the availability of estimates of the phylogeny removes the need to use these classifications'. While it is apparent that phylogeny can now replace classification (of higher ranks) in areas such as comparative biology [43,58], biodiversity conservation [37] and target selection in genome sequencing projects (see below), assuming competition seems less plausible in other respects. Taxa of higher rank can be regarded as named assertions of monophyly, which are needed to efficiently summarize and communicate a phylogenetic hypothesis, and often also to point to 'interesting' features of the allegedly monophyletic group. A conversion step from a phylogenetic tree to a hierarchical classification, such as the Linnaean system, is necessary because the number of ranks is limited [39,66]. While the recent taxonomy mostly employs Linnaeus-style taxon definitions based on unique character combinations (diagnoses), but respects the monophyly criterion, it may be preferable to use only apomorphic characters [59,75]. However, character-based definitions might not always be possible [85], and alternative systems such as the PhyloCode have been proposed [29,30]. Apparently, the fact that the phylogeny-classification conversion is not straightforward has more to do with the underlying taxonomic philosophy [65,76,140] and less to do with the characters available to infer the trees or to 'define' the taxa. In other words, this topic is not specific to genome-scale data and therefore will not be considered below.

This review, which is based on a lecture given at the Leopoldina Symposium on Recent Advances in Microbial Taxonomy (Zürich, March 2009; see also [122,140]), describes the promises that genome sequencing offers for phylogeny and the classification of species and higher ranks. The potential reasons why microbial systematics and genomics have not yet been reconciled are described, along with the measures needed to rectify this situation. As an example of a phylogeny-driven large-scale microbial genome sequencing project, the Genomic Encyclopaedia of *Archaea* and *Bacteria* (GEBA) project is introduced. Recognizing the essential role of the inference of the species tree, both an optimistic and a pessimistic view of the future of genome-based classification are presented, and the potential impact on microbial classification is discussed.

2. The promise of genome-scale data for phylogeny and taxonomy

Although the 16S rDNA gene has been tremendously valuable for establishing the molecular phylogeny of prokaryotes over the last three decades [49], it suffers from the same limits as any other single-gene phylogenetic approach. The resolving power of the gene and its encoded molecule is often rather limited in extensively sampled, diverse clades, particularly close to the species level, and the 16S rDNA gene(s) actually represent only about 0.1% of the coding part of microbial genomes. In addition to the limited sequence space covered, in some cases 16S rDNA inter-operon differences of up to 9% are observed [137]. Similarly, inter-strain differences of up to 16% have been observed. Moreover, 16S rDNA sequence identity is only roughly correlated with DNA–DNA hybridization (DDH) values, which are used as the ultimate decision criterion to establish a new species [14,72]. The availability of numerous genome sequences and rapidly growing metagenomic sequence data raised many expectations in phylogenomic analyses, not only in (i) functional genomics to support the improved annotation of genes, as well as the recognition of taxon-specific gene families as the basis for specific physiological features, but also for (ii) taxonomic and evolutionary purposes, that is, for 'ending incongruence' [51] between single-gene phylogenies by inferring better resolved and more reliable phylogenies from multiple loci. This of course would have a considerable impact on the natural classification system, based only on the inclusion of monophyletic groups [38–41].

The promise of genome-scale data for phylogenetic reconstruction is twofold. Firstly, sampling more characters is a general means of improving the signal–noise ratio. In the context of phylogenomics, an improvement in precision and statistical support has been observed in many 'supermatrix' datasets obtained by concatenating alignments of the large number of genes found in complete genomes [107]. The analysis of supermatrices, which use information from each character (nucleotide or amino acid) directly, is based on the same principles as the analysis of single loci, in which considerable experience regarding the relative performance of phylogenetic methods (e.g. the advantages of explicit evolutionary models) has been gained. Following the identification of orthologous loci, very large sequence alignments, sometimes equivalent to more than 100,000 bp or 33,000 amino acids, were compiled, resulting in impressive resolution of the trees of, for example, yeasts [117], higher plants [70], birds [57], Metazoa [34] and Eukaryotes [151]. Supermatrix analyses of the two prokaryotic domains [149] and all three domains of life [19] based on 31 genes have also been conducted. In comparative empirical studies on phylogenomic methods, supermatrices often performed best, under the proviso that the 'true' reference taxonomy used really corresponds to natural relationships [35,145].

However, a number of issues are far from being settled, such as the choice of method in orthology detection [1,18], inference of sequence alignments [95] and alignment filtering [33]. The effect of missing data [142,143] (usually caused by missing genes in phylogenomic datasets) and of the incongruence between single loci [48] are not yet fully understood. However, these phenomena are particularly prominent in prokaryotic genomes, which differ in size from 416,000 bp (*Buchnera aphidicola*) [106] to almost 13 Mbp (*Sorangium cellulosum*) [123], and only about 80 genes are universally conserved [78]. Moreover, lateral gene transfer (LGT; see below) is supposed to be particularly frequent in prokaryotic genomes. Thus, supertrees [13], inferred by analysing loci separately and using algorithms to combine the trees afterwards, are sometimes regarded as more appropriate than supermatrices [26,47]. A third but not yet frequently used approach is the 'superdistance method' [35], where information from single genes is integrated by combining distance matrices [21].

Moreover, the ratio of phylogenetic and non-phylogenetic signals remains constant even if the size of a dataset is increased, and mis-specified models may lead to maximum support for wrong groupings, as in the case of heterogeneity in base composition [71,108]. Likewise, long-branch attraction artefacts [42] may easily occur in phylogenomic analyses if key taxa (which would subdivide the long branches) are missing [11,86]. Technical problems include the need for high-performance implementation of established phylogenetic methods [54,129,152] and for extended analysis pipelines, particularly in the light of the current lack of standardized file formats for phylogenomics, which increases the effort needed to plug existing components together. The degree of conservation varies dramatically between genes (and codon positions [42]), which may be beneficial in combined analyses because distinct loci provide information for resolving distinct parts of the phylogeny (this may even hold true when combining morphological and molecular characters [8,11]), although selecting an appropriate evolutionary model [88,131] and partitioning scheme [101] for maximum likelihood or Bayesian analysis is more diffi-