Systematic and Applied Microbiology 38 (2015) 217–222

Contents lists available at ScienceDirect

Systematic and Applied Microbiology

journal homepage: www.elsevier.de/syapm

Genome sequences as the type material for taxonomic descriptions of prokaryotes

William B. Whitman*

Department of Microbiology, University of Georgia, Athens, GA 30602, USA

ARTICLE INFO

Keywords:
Genomics
Type strain
Bacteriological Code

ABSTRACT

Genome sequencing of type strains promises to revolutionize prokaryotic systematics by greatly improving the identification of species, elucidating the functional properties of taxonomic groups, and resolving many of the ambiguities in the phylogeny of the higher taxa. Genome sequences could also serve as the type material for naming prokaryotic taxa, which will greatly expand the nomenclature governed by the Bacteriological Code to include many fastidious and uncultured organisms and endosymbionts of great biological interest.

The genome sequencing of type strains promises great advances in the systematics of prokaryotes. In addition to improving the general understanding of prokaryotic biology, these advances include improved: (1) identification of prokaryotic species, (2) identification of functional characteristics useful for resolving taxonomic groups, and (3) resolution of the phylogeny of higher taxa. For many prokaryotic species, the genome sequence could also replace live cultures as the type material. This practice would be especially useful for prokaryotes that are difficult to cultivate or maintain in culture collections.

A major goal in prokaryotic systematics is to delineate the relationships of new isolates with the type strains that serve as the basis for taxonomic classification. The focus on type strains follows from the development of the Bacteriological Code [28] and the Approved List [51]. Prior to the Approved List, tens of thousands of bacterial names were present in the literature [52]. However, the descriptions associated with many of these names were so vague that it was impossible to know to what the names referred. Many of the names were also redundant, with the some organisms possessing multiple names. The Approved List discarded all names that were ambiguous, retaining about 2500 names that were either clearly associated with a biological specimen, i.e. a culture, or detailed and unambiguous descriptions. The Bacteriological Code then insured that all future names would possess clear descriptions, usually by deposition of a representative culture in a public culture collection. It also introduced a system for naming the higher taxonomic ranks based upon the genus names of the type strains and a system of priority. For instance, strain ATCC 6051 is the type of the species *Bacillus subtilis*, and *B. subtilis* is the type species of the genus *Bacillus*. Its priority is determined by the date of its original description, in this case by Ehrenberg in 1835. By the rules of priority, any species that is described after Ehrenberg that includes strain ATCC 6051 must be named *B. subtilis*. Similarly, any genus that includes *B. subtilis* must be named *Bacillus*. Because the Bacteriological Code specifies that the name of the higher taxonomic ranks is determined by the name of the genus, the higher taxonomic ranks are similarly constrained. Thus, the family and order that include strain ATCC 6051 must be named *Bacillaceae* and *Bacillales*, respectively, unless they include a species with greater priority, i.e. validly described at an earlier date. This system allows for naming novel species by inserting them into the existing taxonomy. For instance, a new species similar to *B. subtilis* might be named *Bacillus*. A new species less similar to *B. subtilis* might be given a unique genus name but included in the family *Bacillaceae*. An even less similar species might be given unique genus and family names but included in the order *Bacillales*. This clever system insures the stability of names by preventing subsequent authors from overwriting the established nomenclature with their own names.

Two decisions are paramount in this system. One, is the isolate a new species? Two, if an isolate is a new species, what higher taxonomic classifications are appropriate? Genomics will play important roles in addressing both of these questions.

Genomics for species delineation

In the original proposal for the delineation of species based upon genome similarity [59], two measures of genetic relatedness were proposed to set the boundary for prokaryotic species. The first

* Tel.: +1 706 542 4219; fax: +1 706 542 2674.
E-mail address: whitman@uga.edu

http://dx.doi.org/10.1016/j.syapm.2015.02.003
0723-2020/

measure was the change in the melting temperature (or ΔT_m) of heteroduplex DNA formed upon annealing the DNAs from the two strains to be tested. The ΔT_m is directly related to the sequence identity of the DNAs, and a ΔT_m of about 5 °C, the cutoff proposed for prokaryotic species, corresponds to an average sequence identity of about 92% between the hybridizing DNA [7,18]. A second measure was also suggested to be of equal importance, the extent of DNA–DNA hybridization (or DDH). This would be the fraction of DNA capable of forming heteroduplexes under optimal conditions, usually 25 °C below the melting temperature of the homoduplexes. Importantly, it was the DNA sequence itself and not the method used for determining relatedness that was proposed as the ultimate standard for prokaryotic species delineation [59].

These criteria can now be replaced by Overall Genome Relatedness Indices (or OGRI) derived from the genome sequences [9]. ΔT_m and DDH are laborious to determine and prone to errors [21,50]. With the availability of many new genome sequences, it is now possible to calculate surrogates for both ΔT_m and DDH with a very high precision directly from the genome sequence [3,14,46]. This approach will provide the highest possible resolution and much higher reproducibility in delineating species. The average nucleotide identity (or ANI) is a good surrogate for the ΔT_m because it only compares the sequence identity of DNAs that meet a certain threshold of similarity, usually defined by a BLAST score [23]. ANI is readily determined at EzGenome or JSpecies, which calculates the ANI based upon either the BLAST algorithm or the rapid alignment tool MUMmer [9,46]. The Genome Blast Distance Phylogeny tool (GBDP) offers multiple ORGIs to estimate the DDH and genome sequence identity [2,3]. For closely related strains, these genome-based tools yield values highly correlated with DDH and other measures of genome relatedness [3,14,46]. Lastly, specI is a species identification tool developed to form species clusters based on 40, universal, single-copy phylogenetic marker genes [36].

Recent work suggests that criteria based upon surrogates either of the ΔT_m or DDH may yield substantively different results depending upon the taxon [32]. Because they measure very different properties of DNA, each of the cutoffs have very different implications for genetic relatedness [17,47]. ANI, formula d_4 of GBDP and specI measure sequence identity. They are similar to the standard measures of phylogenetic relatedness and measure the diversity acquired during the accumulations of substitutions and deletions by neutral and other evolutionary processes. In contrast, surrogates of DDH, such as formula d_6 of GBDP, measure the fraction of DNA that is homologous between two strains or the shared gene content and should reflect the prevalence of horizontal gene transfer (HGT) and other processes that insert or remove genes. Given that these are very different evolutionary processes, it is possible for the DNA of two stains to exceed the species cut off by one criterion but not the other. In fact, the ratio of the ANI and ORGIs based upon formula d_6 of GBDP, a surrogate of DDH, varies about two-fold among different prokaryotic lineages [32].

There are several advantages for using measures of sequence identity, whether or not they are based upon the entire genome, such as ANI or formula d_4 of GBDP, or small groups of genes, such as specI or multilocus sequence analyses [12,34,36]. Sequence identity is widely used in phylogenetic studies and is supported by a solid theoretical understanding of the evolutionary processes and a wealth of experimental evidence. Second, sequence identity has clock-like properties and provides the promise of correlation with times of divergence [25]. Discovery of divergence times of prokaryotic groups will enable correlation of the formation of species with the fossil record, the established evolutionary record of eukaryotes and the geological record [4,6]. Third, sequence identity has been used to proposed thresholds for higher taxa in addition to species [29,62]. Thus, classification can proceed by a uniform set of criteria from ancient to modern groups.

In contrast, surrogates of DDH possess many disadvantages. One of the major arguments for surrogates of DDH is that it extends a tradition of DDH as the major genetic criterion for prokaryotic species delineation [34]. This argument neglects the practical and theoretical difficulties of DDH. Although the DDH has been widely used, the accuracy of values reported in the literature is generally quite low [17]. DDH is not symmetrical. Thus, the DDH of strain A to strain B may be different from that of strain B to strain A. When this occurs, there is no theoretical basis for choosing the lessor or greater value or an average of the values. The DDH is also sensitive to the genome size, which is known to vary within species. Thus, even though DDH has been widely used to delineate prokaryotic species, it lacks a precise physical and chemical interpretation. While the 'average' DDH may provide a good sense of prokaryotic diversity, the particular values for any lineage are suspect.

Genomics for identification of functional properties of taxonomic groups

In addition to setting the criteria for delineation of species, genomics can play an important role in how thresholds are applied. While thresholds are necessary to maintain uniformity in taxonomic ranks among phylogenetic lineages, there are many reasons why they should be applied flexibly [12,35]. First, no matter what threshold is chosen, there will be certain groups that fall just below or above the threshold and would be inappropriately subdivided or grouped, respectively (Fig. 1). A related problem is the difficulty in applying thresholds to all strains in a species. For instance, if the threshold is 95% similarity, strains A and B may both possess 95% similarity to the type strain but <95% similarity to each other. In these cases, there may be little value in grouping these strains as separate species. Lastly, there will likely be some lineages where the evolutionary processes are so complex that comparisons of sequence similarity are of little value (Fig. 1D). Genomics will help recognize these lineages and avoid creation of superfluous species. Thus, thresholds are necessary but not sufficient for classification, and other factors such as the physiology and ecology of the groups being classified will have to be considered [49].

Because genome sequences provide enormous insights into the biology of organisms, they are an excellent tool for identifying features that will assist in the final classification [45]. For complex processes, such as development, stress response, quorum sensing, more will be inferred from the genome sequence than ever directly measured for most species. Likewise, for many fastidious organisms, more will probably be known about their physiology and metabolism from their genome sequence than it will ever be possible to observe directly. The genome sequence also provides enormous insights into the evolutionary processes within a group. By revealing deep insights into the biology of the organisms, genome sequencing will reveal the functional criteria most appropriate for creating biologically relevant classifications.

Historically, polyphasic taxonomy has served this role. Polyphasic taxonomy analyzes the relationships among prokaryotes by combining many types of evidence, from ecological to molecular, and often includes sequence as well as growth and chemotaxonomic data [10,19,56,58]. However, many of the growth tests and chemotaxonomic analyses are time-consuming and expensive to perform [57]. Because of their low reliability, the type strains must often be reanalyzed each time a new isolate is added to a group [56]. Importantly, these methods often provide very limited information about the biologically relevant properties of an organism or those properties likely to play a significant role in an organism's persistence in the environment or evolution. For instance, growth experiments are typically conducted in laboratory media with enormous quantities of single substrates. By their very nature,