Gene Reports 5 (2016) 75–82

Contents lists available at ScienceDirect

Gene Reports

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

Procaryotic species and subspecies delineation using average nucleotide identity and gene order conservation

István Kanyó *, Leonóra Varkula Molnár

Department of Biology, Nikola Durkovic School, Feketic 24323, Serbia

ARTICLE INFO

Article history:
Received 19 July 2016
Accepted 13 September 2016
Available online 15 September 2016

Keywords:
Subspecies delineation
Streptococcaceae
Molecular identification key

ABSTRACT

Two genome-based parameters, gene order conservation (GOC) and average nucleotide identity (ANI) were studied to define how accurately two species, or two subspecies could be distinguished on the basis of these parameters. The first examined species was *Lactococcus lactis* since its two subspecies, the subsp. *lactis* and the subsp. *cremoris* have complete genome sequences available. Comparative analysis of DNA sequences of two subspecies of *Streptococcus equi*, *S. equi* subsp. *equi* and *S. equi* subsp. *zooepidemicus* were performed. In addition, *S. thermophilus* and *S. salivarius*, which formerly were classified into one species, were also compared. The GOC values are 66% or higher if the strains belong to the same species and there is no difference in GOC data of strains belonging to either the same subspecies or different subspecies. However, if two different species are compared, the GOC value decreases below 50% enabling an accurate separation of two species. The GOC value of 66% corresponds to a DNA-DNA hybridisation value of 70%. Unlike GOC, ANI measurements allow distinction of subspecies. If the strains belong to the same subspecies, ANI values are higher than 96%. However, when strains from different subspecies are compared the values are between 88% and 90%. Therefore, the average nucleotide identity can be used to delineate the boundary between subspecies. All genomic data and results of phylogenetic analyses obtained here by comparing *S. salivarius* and *S. thermophilus* support their previous taxonomic classification that they are actually two subspecies and not two separate species.

1. Introduction

Genome comparisons show that gene order conservation (GOC) is high between the genomes belonging to the same species and rapidly decreases (i.e. gene order becomes less conserved) when the species are not closely related (Tamames, 2001). Using this parameter the similarity between two isolates can be defined. A study of the family *Streptococcaceae* showed that the rate of GOC is 70% or more between the strains belonging to the same species, while it was below 20% when the species were from different genera (Kanyó and Nagy, 2014).

A second genome sequence-derived parameter for species identification is the average nucleotide identity (ANI) (Konstantinidis et al., 2006; Konstantinidis and Tiedje, 2005). According to the previous studies the value of 95% ANI is proposed for delineating prokaryotic species (Goris et al., 2007; Richter and Rossello-Móra, 2009). This value approximately corresponds to a 70% DNA-DNA hybridisation (DDH) value. However, these studies did not include comparisons between species and subspecies.

Here we determine the GOC and ANI values for species and subspecies of the family *Streptococcaceae*. One of the studied species is *Lactococcus lactis*, a Gram-positive bacterium used for industrial production of fermented dairy products such as cheese, sour cream, fermented milk, etc. The species currently includes four subspecies *L. lactis* subsp. *lactis*, *L. lactis* subsp. *cremoris*, *L. lactis* subsp. *hordniae*, and *L. lactis* subsp. *tructae* (Siezen et al., 2008; Pérez et al., 2011; Kelly et al., 2013). They could be differentiated on the basis of their arginine and carbohydrate metabolism and growth conditions. Two of the subspecies, *L. lactis* subsp. *lactis* and *L. lactis* subsp. *cremoris*, have known genome sequences (Siezen et al., 2010; Kato et al., 2012). Their analysis has proved the differences between the two subspecies although their 16S rRNA sequences show >99% similarity.

The second species studied here is *Streptococcus equi* for which three subspecies are described *S. equi* subsp. *zooepidemicus*, *S. equi* subsp. *equi* and *S. equi* subsp. *ruminatorum* (Fernández et al., 2004; Eyre et al., 2010; Lanka et al., 2010). *Streptococcus equi* subsp. *zooepidemicus* is the most frequently isolated opportunistic pathogen of horses. It is also associated with diseases in a wide range of other animal hosts, including cattle, sheep and pigs, and also humans. *Streptococcus equi* subspecies *equi* is a host-restricted pathogen of horses, the causative agent of equine strangles. The third subspecies is the recently isolated one, *S. equi* subsp. *ruminatorium*. It was isolated from the milk samples of goats and sheep with mastitis (Fernández et al., 2004). Complete genome

Abbreviations: GOC, gene order conservation; ANI, average nucleotide identity.

* Corresponding author at: Department of Biology, Nikola Durkovic School, Bratstva bb, 24323 Feketic, Serbia.
E-mail address: kanyoi@skolafeketic.edu.rs (I. Kanyó).

http://dx.doi.org/10.1016/j.genrep.2016.09.004
2452-0144/

sequences are available for the subsp. *zooepidemicus* and the subsp. *equi* in the NCBI's database (Beres et al., 2008; Holden et al., 2009a; Ma et al., 2011). The comparison of the two subspecies' genomes reveals some differences between them especially in gene content responsible for carbohydrate metabolism and in the number and composition of mobile genetic elements. However, >76% of CDSs found in *S. equi* subsp. *equi* have orthologs in each of the sequenced strains of *S. equi* subsp. *zooepidemicus*.

The third species analyzed here is *S. thermophilus*. This species is closely related to *S. salivarius*, and it was reclassified as *S. salivarius* subsp. *thermophilus* on the basis of DDH experiments under optimal conditions and biochemical analyses (Farrow and Collins, 1984; Hols et al., 2005). However, the DDH studies under stringent conditions together with physiological data resulted in having the two species separated again (Schleifer et al., 1991).

In the present study we use genomic parameters to determine the boundary where two isolates could be classified into the same species even as subspecies and when they should be considered as two separate species.

2. Matherials and methods

2.1. ANI and GOC determination

The genome sequences of the *Lactococcus* and *Streptococcus* species used in this study were obtained from NCBI's FTP site at http://www.ncbi.nlm.nih.gov/Genomes. Gene order conservation was determined using Geneorder4 and Genom. The programs are available at http://binf.gmu.edu:8080/GeneOrder4.0/ and at http://www.skolafeketic.edu.rs/genom/index.php (Mahadevan and Seto, 2010; Kanyó and Nagy, 2014). The sequence alignment and synteny block determination have been done with Geneorder4 while the number of the genes in synteny blocks is estimated by Genom.

The average nucleotide identity was calculated using ANI calculator available at http://enve-omics.ce.gatech.edu/ani (Goris et al., 2007). The program uses best hits and reciprocal best hits for the calculation of ANI values.

2.2. 16S rRNA and DNA-DNA reassociation values determination

The 16S rRNA sequence identity between the species or subspecies was determined as the average identity between all copies of the 16S rRNA gene the species or subspecies possess. The 16S rRNA sequence comparison is done using RDP database (Cole et al., 2014).

DNA–DNA reassociation values between species were obtained from the literature (Farrow and Collins, 1984; Fernández et al., 2004; Jarvis and Jarvis, 1981). We used all available DNA–DNA reassociation values for one species or subspecies. The average DNA-DNA reassociation value is calculated on the basis of these data and it is used for the comparisons. In case of *S. thermophilus* and *S. salivarius* the values determined by Farrow and Collins (1984) are given in Table 3 and the same values were used for the graph in Fig. 3.

2.3. Molecular identification key construction

The molecular identification key was created using gene symbols. The appropriate protein names and gene symbols were used as it is suggested in Protein Data Bank (http://www.pdb.org/pdb/home/home.do) and UniProtKB databases (http://www.uniprot.org). However, if different symbols are used for one gene (such as *gla* and *glpF* for gene glycerol uptake facilitator protein) we selected the symbol, which appears in the *E. coli* strains. For cases where *E. coli* does not contain homologue gene, we used the symbol that was given at the strain itself. In the case of the genes, which determine unknown proteins and there is no appropriate gene symbol, the protein amino acid numbers were used. Several homologue protein amino-acid numbers are different between two species or in different strains of the same species (e.g., pore-forming peptide bacteriocin in strains *S. thermophilus LMG 18311*, *S. thermophilus LMD-9* and *S. salivarius CCHSS3* contains 76 amino acids, while in strain *S. thermophilus CNRZ1066* there are 80 amino acids), in the key we used the amino acid number, which is present in most strains. The BLAST program is used to determine the homologue protein sequences (Altschul et al., 1997). Two proteins are homologous if their sequence identity is at least 75%, and the expected value (*E*-value) is $<10^{-5}$. When one or more genes were present only in one species or strain, for that gene we use parenthesis.

2.4. Phylogenetic analyses

For the phylogenetic analysis two sequences were used. The first one was the *GroEL* (hsp60) gene sequence identified in previous studies as an appropriate phylogenetic gene marker for procaryotic species (Goh et al., 1996; Kwok and Chow, 2003; Claesson et al., 2008). The second one was obtained by concatenating the entire sequences of seven housekeeping genes. The nucleotide sequences of the following seven genes were selected: *glyA* (encoding serine hydroxymethyltransferase), *pepX* (encoding x-prolyl-dipeptidyl aminopeptidase), *pepN* (encoding aminopeptidase N), *recN* (encoding DNA repair and genetic recombination protein), *rpoA* (encoding DNA-directed RNA polymerase subunit alpha), *nrdE* (encoding ribonucleotide-diphosphate reductase subunit alpha), and *proS* (encoding prolyl-tRNA synthetase). The selected genes were common in MLST databases of *S. thermophilus*, *S. equi* and *L. lactis* (Delorme et al., 2007; Webb et al., 2008; Delorme et al., 2010; Passerini et al., 2010).

GroEL and concatenated sequences were aligned with ClustalW and a neighbor-joining tree was constructed in MEGA version 6.0 with 500 bootstrap runs (Thompson et al., 1997; Tamura et al., 2013). The Tamura-Nei method was used for estimating pairwise genetic distances. The net average distance between two groups is given by the equation

$$d_A = d_{XY} - ((d_X + d_Y)/2)$$

Where, d_{XY} is the average distance between groups X and Y, and d_X and d_Y are the mean within-group distances. X represents sequences of one subspecies (or sequences of *S. salivarius*) while Y sequences of the other subspecies (or *S. thermophilus*).

3. Results

3.1. Gene order conservation measurements

Numerous complete genome sequences for the two subspecies, *L. lactis* subsp. *lactis* and *L. lactis* subsp. *cremoris*, of *Lactococcus lactis* are available in the database. For the GOC comparison two *L. lactis* subsp. *lactis* (IO and KF147) and two *L. lactis* subsp. *cremoris* strains (NZ9000 and KW2) were used. The GOC values for the other *Lactococcus lactis* strains were previously determined (Kanyó and Nagy, 2014). The ANI and GOC values among the strains are presented in Table 1. The number of genes in the synteny blocks is always higher if two strains belong to the same subspecies. In case of *Lactococcus lactis* the number of protein coding genes (CDSs) is between 2200 and 2500, the number of genes in the syntenic blocks is above 1900 if the strains belong to the same subspecies, and remain below 1900 if they are from different subspecies. However, based on the results of the GOC measurements it is not possible to differentiate between the subspecies. When *L. lactis.* subsp. *cremoris* NZ9000 (number of CDSs 2510) is compared with the other *L. lactis.* subsp. *cremoris* strain (KW2 with number of CDSs 2268) the GOC value is 78% while in the case of *L. lactis.* subsp. *cremoris* KW2 and *L. lactis* subsp. *lactis IO-1* it is 82% (Table 1).

There is only one strain of *S. equi* subsp. *equi*, while there are three strains of subsp. *zooepidemicus* with known complete genome