The Bulletin of BISMiS

Published by Bergey's International Society for Microbial Systematics

Volume 6, part 1 – July 2016

The Bulletin of BISMiS

Published by Bergey's International Society for Microbial Systematics

ISSN 2159-287X

Publisher and Editorial Office

Bergey's International Society for Microbial Systematics
Department of Microbiology and Plant Biology
770 Van Vleet Oval
University of Oklahoma
Norman, OK 73019-0390 USA
Email: paul.lawson@ou.edu

On the cover

Information For BISMiS 2016 Meeting in Pune, India

The Bulletin of BISMiS

Contents of Volume 6, part 1

The Bulletin of BISMiS (2016), Volume 6, part 1, pp 5-11

BISMiS 2016: The Changing Face of Microbial Systematics

Kamlesh Jangid

BISMiS-2016 is just a year away and its planning is already in full swing. The conference will be held from September 12 - 15, 2016 at the Microbial Culture Collection (MCC) in the city of Pune, India. Now is the perfect time to update you about the meeting theme and the local attractions so that you can plan ahead and send us your inputs for sessions and other things that you would like to see at BISMiS.

The Meeting Plan

The meeting will focus on the use of genomic/ enviro-genomic data for the description of novel taxa. The importance of taxonomy in the age of genomics has been highlighted very well in a recent issue of Systematic and Applied Microbiology, Volume 38, Issue 4. With the advancement in sequencing technologies and an increasing proportion of novel taxa descriptions using genomic data, the future of taxonomy holds great potential for DNA and genomic sequences as the primary type material, rather than solely relying on obtaining pure cultures for the description of novel taxa.

The meeting is being held at this crucial juncture when methods and minimum standards are likely to be formulated using genomic data in microbial systematics. It is therefore aimed to be inclusive of the developments in this field and will provide an excellent opportunity to understand the changing face of microbial systematics.

A single advisory-cum-organizing committee has been set-up that will look after the overall planning for the meeting. The meeting will last three and half days and participation is limited to 150 attendees. In summary, 23 talks are planned including, three sunrise talks and six sessions.

Of these, ten talks are reserved for invited speakers from experts in the field. The remaining are 12 speakers will be chosen from the registering participants based upon the submitted abstracts. Each Sun Rise talk is planned for 50 minutes whereas each session talk is planned for 40 minutes including Q & A session.

An open mixer on the first evening will allow extra time for participants to network and exchange ideas. The closing day will include an Indian cultural program performed by students of the Contemporary Art School followed by a closing dinner.

Session Themes

Currently, we are planning to hold sessions on cultures, culturing of as-yet-uncultivated microbes, cyanobacterial taxonomy and on the use of metagenomics for description of novel taxa. Theme ideas for other sessions are open and if you wish to propose any session theme, please contact us.

Important Dates

Abstract Submission: April 1st to June 1st, 2016
Early Bird Registration: Before June 15th, 2016
Late Registration: After June 15th, 2016
On-site Registration (based on availability): September 12th, 2016

Reaching Pune

The meeting venue will be MCC, housed at Sai Trinity, Pashan, Pune. It is strategically located on the outskirts of West-Pune where most academic institutions are located and allow easy commute to the Mumbai International Airport. Pune has its own international airport as well, so one could fly

Contact details
Microbial Culture Collection
National Centre for Cell Science
Maharashtra, India
jangidk@nccs.res.in

directly to Pune. Pick-up will be arranged from both airports for our international participants.
Limited accommodation at discounted rates in select hotels will be arranged for participants. The details will soon be posted on the BISMiS website (www.bismis.org).

Local Attractions and Places of Interest

Pune is the cultural capital of Maharashtra state and there is no better representation of the city than its most revered festival, the 10-day Ganesh Festival (Figures 1A and 1B). Also called Ganeshotsav, the festival ends with a grand immersion of the Ganesh Idols in water bodies. Coincidently, BISMiS 2016 will be hosted during the last four days of Ganesh festival offering a unique opportunity to our participants to witness this grandeur event in action.
Amongst the local attractions, Pune offers a wide range of options from shopping malls to the local joints. If you are a shopping enthusiast, Laxmi Road is the largest market in Pune, offering jewellery, clothes and other wearables.

If you prefer historical monuments, visit Shaniwar Wada, the fort palace of the Peshwa rulers until 1818 or the Aga Khan Palace, which also served as a prison for Mahatma Gandhi. If you would like to explore your inner self, you may visit the OSHO International Meditation Centre at Koregaon, Pune (Figure 2). If you are a food enthusiast, don't forget to try the local delicacies, such as Chitale's Bhakarwadi and Surali Wadi, Kayani's Shrewsbury Biscuits or Budhani's potato wafers (Figure 3).

For nature lovers and photography enthusiast, this is the best time to visit Pune. With monsoon nearing its end, Maharashtra's landscapes around Pune offer the most picturesque sites (Figures 4a-4f). Of the many tourist options, ones that are within a 1-4 hour drive include Sinhgad and Lohagad forts, Lonavala and Khandala Ghats, Bhimashankar wildlife sanctuary, Kaas Plateau, Mahabaleshwar and Panchgani hill stations. As microbiologists, these are of more interest to us since the region is part of the Western Ghat, a biodiversity hostpot and UNESCO World Heritage Site.

If you are planning to stay a little longer, the famous cave paintings of Ajanta and Ellora in Aurangabad along with a short visit to the Lonar Crater, a 50,000-year old meteorite impact crater on Basalt rocks, can be completed in a two-day

Figure 1A. Ganesh Festival in Pune that coincides with BISMiS 2016.
Source: www.storify.com

Figure 1B. Ganesh Festival in Pune that coincides with BISMiS 2016. www.pune.wordcamp.org

trip from Pune.

Get connected with BISMiS

With the program getting constantly updated, frequent updates are not possible through the BISMiS Bulletin and the Microbial Taxonomist Newsletter.

So, make sure you follow us on Facebook (/pages/Bergeys-International-Society-for-Microbial-Systematics-BISMiS/248730865152521) and Twitter (@BISMiS_). Post and tweet your new ideas and what you would like at the meeting.

Figure 2. Tourist attractions in Pune. (Left) Entrance of Shaniwar Wada, the palace of the Peshwa rulers until 1818. (Right) Aga Khan Palace served as a prison for Mahatma Gandhi; Photo Source: Google® images and www.osho.com.

BISMiS 2016 Organizing Committee Members

Brian Austin, University of Stirling, UK

Jongsik Chun, Seoul National University, Korea

Kamlesh Jangid, Microbial Culture Collection, India

Martha E. Trujillo, Universidad de Salamanca, Spain

Paul A. Lawson, University of Oklahoma, USA

Shekhar Mande, National Centre for Cell Science, India

VG Kshirsagar, Association of Microbiologists of India-Pune, India

William B Whitman, University of Georgia, USA

Yogesh Shouche, Microbial Culture Collection, India

Tentative Program Schedule

September 12, 2016 (Day 1)

Registration
Opening Session
President's Speech
Opening Mixer

September 13, 2016 (Day 2)

Sunrise Talk 1
Session 1 (3 talks of 40 min each)
Session 2 (4 talks of 40 min each)

September 14, 2016 (Day 3)

Sunrise Talk 2
Session 3 (3 talks of 40 min each)
Session 4 (4 talks of 40 min each)

September 15, 2016 (Day 4)

Sunrise Talk 3
Session 5 (3 talks of 40 min each)
Session 6 (2 talks of 40 min each)

Discussion by expert panel
Closing Session including valedictory function
Cultural Program and Dinner

Figure 3. (Left) Chitale's Bhakarwadi (Right) Suraliwadi. Photo Source: Google® images

Figure 4a. Lohagad Fort

Figure 4b. Kaas Plateau

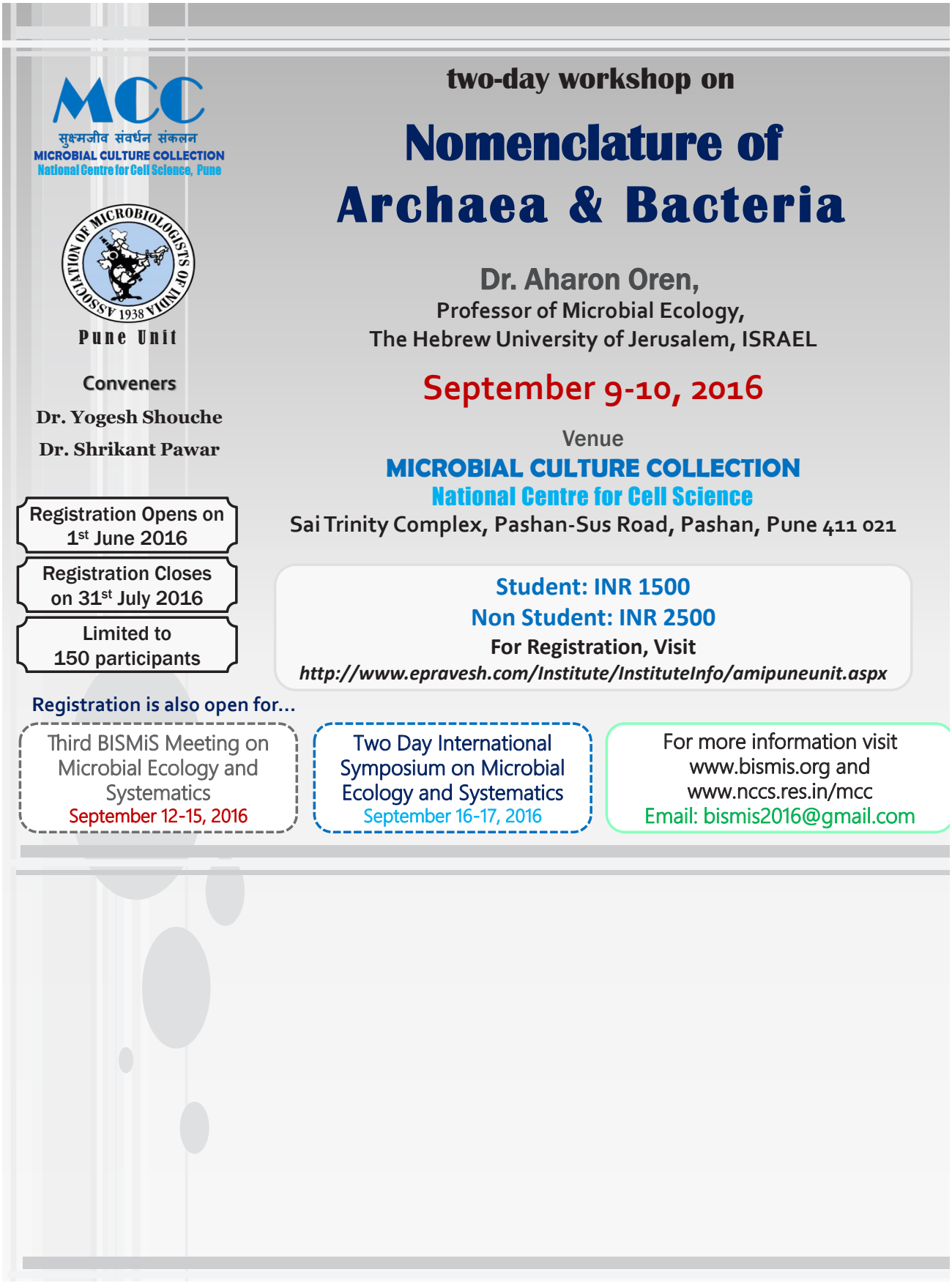
Figure 4c. Ajanta Caves

Figure 4d. Ajanta cave paintings

Figure 4e. Ellora caves

Figure 4f. Lonar Crater.

Figure 4. Tourist attractions near Pune. Photo Source: Google® images


MCC
सूक्ष्मजीव संवर्धन संकलन
MICROBIAL CULTURE COLLECTION
National Centre for Cell Science, Pune
ASSOCIATION OF MICROBIOLOGISTS OF INDIA 1938
Pune Unit
Conveners
Dr. Yogesh Shouche
Dr. Shrikant Pawar
two-day workshop on
Nomenclature of Archaea & Bacteria
Dr. Aharon Oren,
Professor of Microbial Ecology,
The Hebrew University of Jerusalem, ISRAEL
September 9-10, 2016
Venue
MICROBIAL CULTURE COLLECTION
National Centre for Cell Science
Sai Trinity Complex, Pashan-Sus Road, Pashan, Pune 411 021
Registration Opens on 1st June 2016
Registration Closes on 31st July 2016
Limited to 150 participants
Student: INR 1500
Non Student: INR 2500
For Registration, Visit
http://www.epravesh.com/Institute/InstituteInfo/amipuneunit.aspx
Registration is also open for...
Third BISMiS Meeting on Microbial Ecology and Systematics
September 12-15, 2016
Two Day International Symposium on Microbial Ecology and Systematics
September 16-17, 2016
For more information visit
www.bismis.org and
www.nccs.res.in/mcc
Email: bismis2016@gmail.com

Two-Day International Symposium on Microbial Ecology and Systematics

September 16-17, 2016

Venue

CSIR- National Chemical Laboratory

Dr. Homi Bhabha Road, Pashan, Pune, 411008

Speakers

Aharon Oren, The Hebrew University of Jerusalem, ISRAEL

Boris A Vinatzer, Virginia Tech, USA

Brian Austin, University of Stirling, UK

Brian P. Hedlund, University of Nevada, USA

Ch. Sasikala, JNT University, Hyderabad, INDIA

Jörg Overmann, DSMZ, GERMANY

Martha E. Trujillo, Universidad de Salamanca, SPAIN

Paul Lawson, University of Oklahoma, USA

Pelin Yilmaz, Max Planck Inst. for Marine Microbiol., GERMANY

S. Shivaji, LV Prasad Eye Institute, Hyderabad, INDIA

William B. Whitman, University of Georgia, USA

Yogesh Shouche, MCC-NCCS, INDIA

Student Member (AMI or BISMiS): INR 2500
Student Non-Member: INR 3000
AMI or BISMiS Member: INR 5000
Non-Member: INR 8000

Fees to be paid on:
http://www.epravesh.com/Institute/InstituteInfo/amipuneunit.aspx

For More Information visit
www.bismis.org or www.nccs.res.in/mcc
Email: bismis2016@gmail.com

MCC
सूक्ष्मजीव संवर्धन संकलन
MICROBIAL CULTURE COLLECTION
National Centre for Cell Science, Pune

Association of Microbiologists of India 1938
Pune Unit

ncl

Conveners
Dr. Yogesh Shouche
Dr. Shrikant Pawar

Registration Opens on 1st June 2016

Registration Closes on 31st July 2016

Limited to 500 participants

Registration is also open for

Third BISMiS Meeting on Microbial Systematics and Metagenomics
September 12- 15, 2016

Two-Day Workshop on Nomenclature of Archaea and Bacteria by Prof. Aharon Oren
September 9-10, 2016

The Bulletin of BISMiS (2016), Volume 6, part 1, pp 13-14

I Culture, I Characterise, I Archive

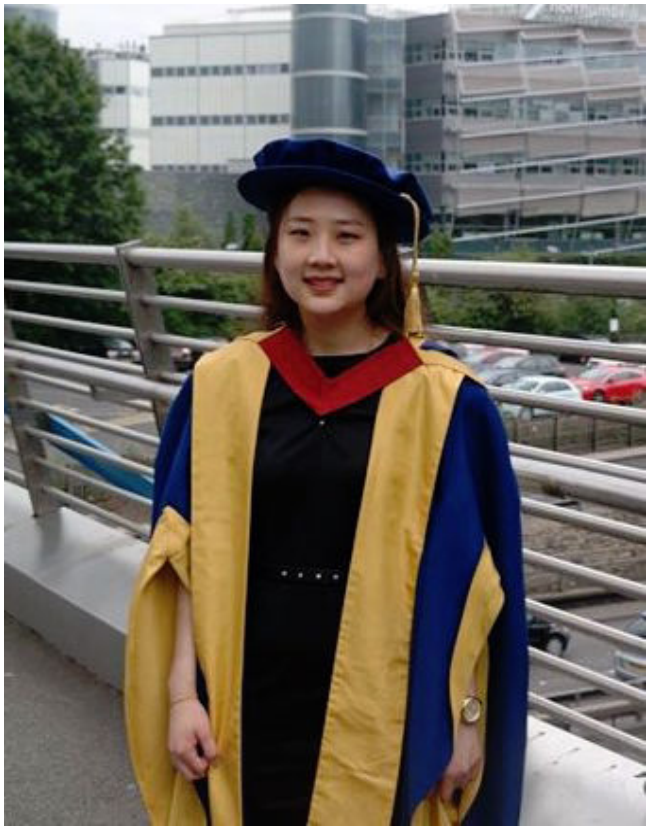
Jesmine Lim

I have always been fascinated about microorganisms. I still remember during my high school biology class when we first learned about these tiny bacteria by peeking through a microscope to observe an active colony. My curiosity was raised and I was keen to learn more about these beautiful microorganisms. And so I begin my journey, immersing myself in the realm of biology by studying related courses during pre-university and later pursuing a BSc and MRes in Biotechnology.

Upon finishing my masters degree, I was lucky to be offered a PhD studentship by Northumbria University under the supervision of Dr. Amanda Jones, Prof. Stephen Cummings and Prof. John Woodward. My research project was about characterising the prokaryotic community in one of the most well preserved lake sediment in Japan that has an excellent chronological storage of intrinsic lake sediment dating back to 150,000 years ago. In the broadest sense, this research project intended to utilise the diversity of the lake sediment bacteria to construct a novel palaeoenvironment biomarker. Two major characterisation techniques were employed in this study, notably the culture-dependent and molecular techniques.

By using the former technique, I was exposed to the line of work in microbial taxonomy. Typically, the investigation practices following the culture-dependent path, involve the initial separation of bacteria through differentiation and differential centrifugation, followed by the cultivation of isolated samples using uniquely engineered media, where its chemical compositions were designed to maximise the bacteria recovery rate based on the findings from molecular techniques.

The results from 16S RNA gene sequencing and phylogenetic analyses revealed a strong dependence between the bacterial diversity and the lake water chemistry. We discovered that the freshwater sediments were comprised of 19 distinct Actinobacteria taxa while Gram-positive bacterial taxa such as *Bacillus* and *Paenibacillus* were predominant in brackish sediment. Further phenotypic and chemotaxonomy analyses have also enabled us to identify novel strains from the genera of *Dermacoccus*, *Dietzia*, *Leifsonia* and *Rhodoccocus*.

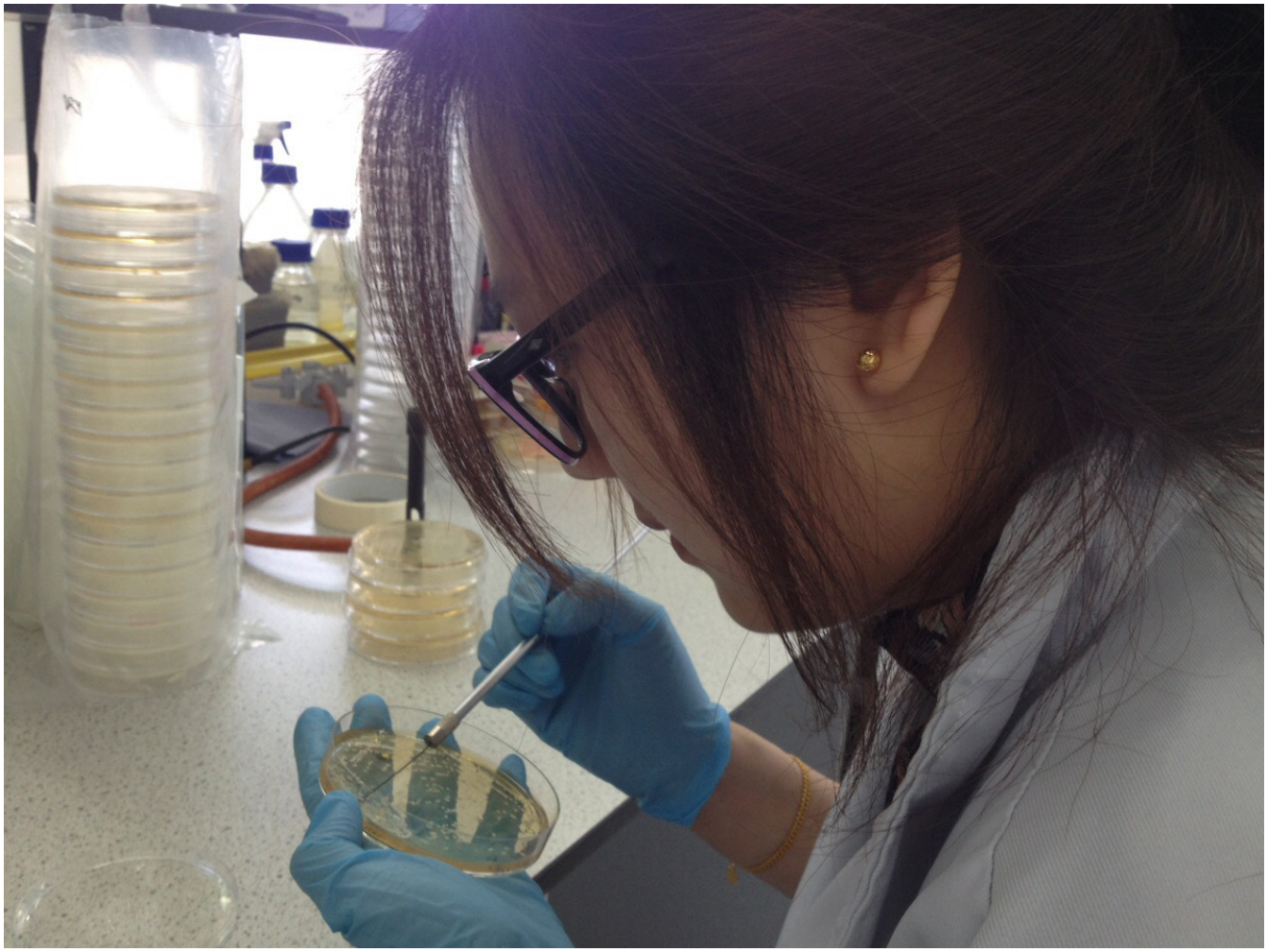


Jesmine Lim

Apart from doing research, I also had a great opportunity involving academic teaching. I started out by demonstrating for the microbial taxonomy practical class. Then I was able to deliver lectures and IT practical classes for masters students during my final year of PhD.

In addition, I gained plentiful of new insights

Contact details
Faculty of Health and Life Sciences
Northumbria University
Newcastle upon Tyne, NE1 8ST, UK
jesmine86@hotmail.com



Jesmine Lim working in her laboratory

attending regional and international conferences. By engaging with other microbiologists working in a similar area of interest, new research ideas were stimulated, which I found extremely helpful during the planning stage of my experiments. Not only have I broadened my horizon on the frontier of microbiology, the people that I met at conferences have inspired and motivated me to be a better scientist as I learnt the impact and importance of our work in solving real life problems. These experiences shaped me to actively seek new skills and techniques as well as being appreciative to the work of microbial taxonomy.

As soon as I completed my PhD, I received a job offer as a senior research scientist in a biotechnology company. My main role in the company is working towards the discovery of novel enzymes for industrial applications via metagenomics sequencing and bacterial cultivations. I am in charge of two EU funded research projects that involve 17 collaborators from within the EU. From time to time, I am also involved in other projects, working with other colleagues towards the completion of their projects. Without doubt, the training and experiences I acquired during my PhD studies have become extremely valuable for this position.

The Bulletin of BISMiS (2016), Volume 6, part 1, pp 15-15

BISMiS Business : Introduction of Bulletin Associate Editor

Paul A. Lawson

It is my pleasure to introduce Jang-Cheon Cho as an associate editor of the Bulletin. Jang-Cheon is a full professor in the Department of Biological Science at Inha University, Republic of Korea, where he has served as a faculty member since 2005. He also has been holding an Inha Fellow Professorship since 2009. After his undergraduate studies at Seoul National University, Korea, he completed his Ph.D in the Department of Microbiology at the same university in 2000. The major theme of his Ph.D dissertation was microbial ecology of groundwater bacteria dwelling in contaminated aquifers. Jang-Cheon joined Steve Giovannoni's lab at Oregon State University in 2001 and worked as a postdoctoral researcher in the field of marine microbiology and microbial systematics for 4 years.

Jang-Cheon's research interests lie in the area of aquatic microbiology, ranging from marine environments to freshwater ecosystem, with special focuses on cultivation of majority of the uncultured bacteria, microbial taxonomy of oligotrophic marine bacteria and their genomes.

He has been isolating and reporting some of abundant marine Proteobacteria, including "*Candidatus* Puniceispirillum" (SAR116 clade) and "*Candidatus* Pelagibacter" (SAR11 clade). He has also taxonomically created several higher taxa related to the PVC superphylum, such as the phylum *Lentisphaerae*, the class *Opitutae* and the order *Puniceicoccales*, contributing to the 2nd edition of Bergey's Manual of Systematic Bacteriology by writing with Brian Hedlund.

In recent years, he has studied on the isolation of bacteriophages infecting representatives of major bacterial assemblages and showed that "*Candidatus* Puniceispirillum"-infecting SAR116 phages contributes significant proportions to marine viromes.

Jang-Cheon Cho

Jang-Cheon was an associate editor of Journal of Microbiology from 2009 to 2011 and Journal of Microbiology and Biotechnology during 2012 to 2014. He has been serving as an associate editor for Microbes and Environment since 2012. Jang-Cheon has been a member of the organizing committee for K-J-T (Korea-Japan-Taiwan) symposium on microbial ecology since 2009. He has been serving as an associate editor of International Journal of Systematic and Evolutionary Microbiology (IJSEM) since 2010. As an associate editor of IJSEM, he has been handling manuscripts mostly on aquatic isolates (mainly from saline environments) belonging to the *Proteobacteria*, *Flavobacteriaceae*, *Verrucomicrobia* and *Lentispharae*.

Jang-Cheon Cho at a meeting

Contact details for Professor Cho
Professor in Department of Biological Sciences
Inha University, Inharo 100
Republic of Korea
chojc@inha.ac.kr

The Bulletin of BISMiS (2016), Volume 6, part 1, pp 17-34

Genome Sequence-Based Microbial Taxonomy: I Will Know It When I See It

Micah Krichevsky

Taxonomy provides the context for the single isolates grouped in the Bacteria and Archaea domains. All taxonomies result from analysis of observational data. The first step in establishing a taxonomy or phylogeny is classifying a cohort of organisms to establish coherent groups. As John Locke (1690) posited, and here paraphrased: People make species, nature does not. People make boundaries of species and nature does not. Thus, uncertainty and subjectivity affect the process of establishing species (Krichevsky, 2011) and determining strain identity (Krichevsky at al, 2013).

A whole genome sequence denotes the status of the genetics of the organism at the time the determination is made. It is a snapshot, but much more complete than any partial sequence or a survey of select genes. Use of whole genome sequences mitigate the informational noise potentially present due to genes excluded from the classifying analysis. The availability of whole genome sequences for constructing a taxonomy or phylogeny facilitates development of even more robust products. However, the interpretation of the underlying data to produce a taxonomy or phylogeny has constraints in common with precedent methods.

In all cases, the construction process consists of interpretation of the original observational data. At each stage of the process, methodological, interpretive, and philosophical choices are available. Thus, there are components of subjectivity in the process. This paper discusses the procedural uncertainties and subjective decisions inherent in the process of construction of a whole genome-based taxonomy or phylogeny. The microbial taxonomist, who must be pragmatic, must build, adapt, or choose a taxonomy or phylogeny to fit the need. All of the above alternatives for each of the construction processes force final decisions to be judgmental.Thus: I will know it when I see it.

Contact details
Bionomics International
3023 Kramer Street
Wheaton, Maryland 20902, USA
michahk@bioint.org

Background

All efforts at classification require some form of direct comparison of attributes of the included cohort. The historical basis for classification of bacteria utilized various specific components and activities of the organism as surrogates for the complete cell in that comparison. Size and shape as visually perceived were the first, and still are, used. Physiological and biochemical characteristics, immunological tests, and macromolecular components and their sequences are in widespread current use. Each of these, either singly or in combination, acts as surrogates for the complete cell. For most of these, data acquired for each of these attributes forms the basis for comparisons. However, for DNA-DNA hybridization, multiple physical samples or, if done electronically, genome sequences are directly compared to generate taxonomically useful results, rather than just the data for specific attributes. This is also true for immunologic tests. DNA-DNA hybridization utilizes the whole genome but gives no specific differentiation information.

Like hybridization, the whole or partial genome sequence of an organism can provide information without the need for expression of features. But, unlike hybridizations, comparisons of genome sequences do not need to match physical samples of organisms in question. Some sequence comparisons, as with hybridizations, can be attributed independently, or they may involve attribute predictions. The result is comparisons

like those of phenotypic expression testing.
A common set of such partial genome sequences are housekeeping genes and the most common of those, used alone, is 16S rDNA. Taxonomists commonly use one or more housekeeping gene sequences in classification. The 16S rDNA sequence data and housekeeping gene sequences support proposals for a new species and for constructing or enhancing classifications. Non-housekeeping genes may be added to the classifier panel, as well. Similarly, partial DNA sequences find application as classifier tools in other microorganisms. As useful and insightful as such partial sequences are as classifiers, they are but a cropped snapshot of their part of the whole genome.

Having access to the whole genome sequence facilitates detection of the locus of single point changes in genes of specific interest. Such changes may be missed simply by virtue of being in some part of the genome which is not included in the partial DNA sequences selected for study.

Short history of use of DNA sequences in prokaryotic classification

Developing coherent and stable taxonomies for prokaryotes is a continuing challenge due to the diversity of these organisms, the frequent lack of useful morphological features like those used for most eukaryotes, and potential horizontal transfer of genes. All efforts at classification require some form of direct comparison of attributes of the included cohort. The historical basis for classification of bacteria utilized various specific components and activities of the organism as surrogates for the complete cell in that comparison. Size and shape as visually perceived were the first, and still are, used. Physiological and biochemical characteristics, immunological tests, and macromolecular components and their sequences are in widespread current use. However, for the past few decades, these laborious laboratory experiments have been supplanted by molecular approaches and sequencing technologies, which now dominate taxonomic procedures.

The first genotypic approach bacteriologists adopted to classify prokaryotes on the basis of their phylogenetic relatedness was DNA-DNA hybridization (DDH) (Wayne, et al. 1987), which became the nominal “gold standard” for classifying species of prokaryotes. But, like the phenotypic and biochemical methods, this is still a primarily wet-lab technique, encumbered by all the problems such procedures entail. DNA-DNA hybridization requires multiple partial or whole genome physical samples for direct comparison to generate taxonomically useful results. Compared to the earlier phenotypic analyses, DNA-DNA hybridization utilizes the whole genome. Nevertheless, it gives no specific differentiation information.

However, new electronic procedures for whole genomes are now available ranging from the construction of the sequence, through methods for comparing and aligning sequences to the development of taxonomies, phylogenies and identification resources. Examples follow that illustrate this active field.

Sequencing of single “housekeeping” genes eventually became commonplace as a means of classifying bacteria.

The most accepted criteria for selection of the genes for these essential proteins are such that they should not be subjected to horizontal gene transfer (HGT), should be present in all bacteria, preferably in single copies, and at least two highly conserved regions for the design of PCR primers (Yamamoto and Harayama, 1995). The first, and currently most extensively used, single gene employs the small-subunit ribosomal RNA gene (16S rDNA), but others have been utilized, such as the gyrB and rpoD genes. “In addition to the single gene based methods, MultiLocus Sequence Typing (MLST) has been widely used for genotypic characterization and classification of prokaryotes by comparing multiple housekeeping gene sequences (Maiden, et al. 1998). However, usually a different set of genes is useful for different sets of organisms, and some difficulties occur in primer design for amplification of genes in all strains if the analysis is not conducted all in silico” (Ozen, et al. 2013).

Whole genome sequencing is now inexpensive and commonplace. Directly comparing apparently

related microorganisms using measures such as the Average Nucleotide Identity (ANI) provide an *in silico* analog to the re-annealing process of DDH. An advantage of this method is that it avoids the wet-lab complications of older methods. The study of Goris, et al. (2007) on pair wise comparison of completely sequenced genomes showed the ANI of the core genes yielded results similar to analysis of 16S rDNA sequence identity and DDH similarity values. They concluded that a 70% DDH value corresponds to 95% ANI. Hence, ANI provides an alternative to the tedious DDH method (Goris, et al. 2007).

Others choose to limit comparisons to a set of conserved essential genes common to all genomes compared: the "core-genome". Core-genomes of phylogenetically coherent groups are considered to be more stable and contain genes that are less prone to horizontal gene transfer (Ozen et al., 2013). These comparisons are more comprehensive than analyses that use a select few genes in MLST. It is hoped that the core-genome contains a more consistent set of genes common across larger groups. MLST gene sets may not translate very well across less closely related taxonomic groupings, requiring different gene sets for different group cores.

The Purpose of classification

The act of classification allows thinking about biological organisms in an orderly construct. This process facilitates assigning names to the resulting classes (groups). Conventionally, taxonomies result in species names which, in turn, are grouped into multilevel taxa, also given names. Phylogenies result in described clades which group as lineages. Frequently, lineages are given names. Phylogenies aim to describe the historical evolutionary relationships of the cohort under study.

This paper raises the question, whether the use of whole genome data can provide a way to arrive at a universal taxonomy, as recommended by Staley (2009), not dependent on subjective decisions in the process of developing such a taxonomic structure. Many applications of microbiology depend on assigning names to groups of organisms and subsequently placing an individual organism in a named group, e.g., human, animal, and plant pathology; patents;
biodiversity and ecological studies; and so forth. We strive for a "correct identification."

Classification versus identification: Both taxonomy and phylogeny are the results of classification. Thus, the product presents the presumed orderly natural relationships among the classified objects. The data submitted for classification are all of equal weight.

The process of identification attempts to allocate a new specimen to the "correct" and previously described taxon or lineage. The characters are weighted by ability to distinguish among taxa or lineages. In the case of sequences, the closest match to an annotated sequence is often considered identification. Thus, the identification of the object sequenced depends on the reliability of the target annotation of the closest matching sequence.

Factors that may affect building a whole genome-based universal taxonomy or phylogeny

At all steps in the classification and subsequent interpretation, there are inherent error rates and choices that must be made. The first group of factors involves machine and operator errors. These are directly measurable by well-known methods. The remaining factors are the direct result of decisions of procedure and interpretation by the investigators.

Effect of reproducibility of sequence determination

-Mechanics
- *Machine differences
- *Error rates
  - ~Machine errors
  - ~Operator errors

Effect of differences in analysis methods for creating a taxonomy or phylogeny

Levels of annotation

-Annotation of functions of specific sequences
*Differences in labeling of gene sequences
*Relevance of choice of housekeeping genes
*Quality of Genome Sequence (Chain, 2009)
~Standard Draft
~High-Quality Draft
~Improved High-Quality Draft
~Annotation-Directed Improvement
~Noncontiguous Finished
~Finished
*Use of presumed taxa names
*Robustness of the set of examples
*Investigator currency and bias with name changes
*Relevance of use of type strains and species

Stability of genomes within taxa

-Taxa with robust cores
-Taxa with limited and variable genome cores
-Horizontal gene transfer
-Variations in mutation rates

Agreement on conventions for biological meaning of names

-Absence of a judicial entity for "correct usage of names"

Alternatives from which to choose are available at all process steps. Thus, professional judgment is the final determinant of the process. Of course, each step of the process varies as to its overall contribution to the final result. For example, the extent of machine error contributes less than which clustering algorithm is chosen. Further, the choice of decision of which machine to use largely depends on financial and administrative constraints such as existing institutional resources, budget, prior experience, and available personnel support.

Assembling the final sequence

Comparison of machine error rates and final editing of assembled sequences are active subjects for publication. Examples follow which illustrate the range of alternative tools available to the classifier.

McElroy, et al., (2014) tabulate the properties of the common sequencing machines by manufacturer, platform, throughput (per machine run), reported errors, and depth (for viruses and bacteria). The error rates listed are machine and chemistry dependent on the individual platform. In addition to naturally occurring indels, the process of constructing the sequence confounds interpretation of indels. Process steps such as library preparation, sequencing biases, and algorithm artifacts can contribute to sequence interpretation. While these variable error rates are low in the context of most biological measures, the errors become significant in detection of variants. A bacterial genome would have ~50 indels and ~5,000 substitutions in the best reported machine rates. They also describe various sources of the errors and methods for overcoming errors in deep sequencing. The categories include sample collection and PCR errors, sequencing errors, alignment errors, and haplotype reconstruction errors. Various methods for overcoming errors are available as choices to the classifier.

Kisand and Lettieri (2013) describe the availability of open source tools for assembly and annotation of Next Generation Sequencing (NGS) data. They use pyrosequencing technology results as their example. They conclude: "Free and open source software (FOSS) tools for assembly and annotation of NGS data are being developed rapidly to provide accurate results with less computational effort. Usability is not high priority and these tools currently do not allow the data to be processed without manual intervention (emphasis added). Despite this, genome assemblers now readily assemble medium short reads into long contigs (>97-98% genome coverage). A notable gap in pyrosequencing technology is the quality of base pair calling and conflicting base pairs between singlereads at the same nucleotide position. Regardless, using draft whole genomes that are not finished and remain fragmented into tens of contigs allows one to characterize unknown bacteria with modest effort."

Lou, et al., (2013) describe a process starting with library preparation and computer processing of the reads produced by a high throughput sequencing machine. They report significant improvement in error rates and yield in comparison to various

barcoding methods. They predict that this process "efficiently reduces sequencing error, [and] this method will be broadly enabling in projects where high-throughput sequencing is applied to detect variation in complex samples such as tumors, microbialpopulations, and environmental communities."

Considerations regarding computational subjectivity

The increasing availability of computers led to concomitant programs to use them. This software functionality fostered development of complex and diverse classification methods. The Classification Society (www.classification-society.org) was established on 17 April 1964 "to promote the scientific study of classification and clustering (including systematic methods of creating classifications from data)". Biological classification was an early beneficiary of computational classification alternatives. These alternatives present classifiers with choices to make as to which computational methods best fit their particular needs.

For example, Willcox, et al., (1973) described the theoretical and practical issues of deciding whether to use Bayesian probabilities or maximum likelihood in constructing identification matrices. Use of computers for taxonomy and identification already had a rich history by the time of their paper.

Two of the major steps in the computational process involving considerable judgment are searching for and comparing sequences. Multiple techniques for searching and comparing sequences are available. Lorenza Bordoli (2006) summarized the overall topic, including theory and practice. Nevertheless, "The biologist's expertise is still essential to the analysis of the results!" Thus, the biologist makes subjective decisions in choosing parameters for the operation and interpretation of a Blast search.

One can choose among a number of variations of the original BLAST program, as well as alternative heuristic search programs. Non-heuristic programs are more sensitive and optimized for completeness of comparisons. Two resources listing programs in both categories are: http://en.wikipedia.org/wiki/BLAST#Alternative_versions and http://www.ebi.ac.uk/Tools/sss/.

Alternate choices for methods commonly used for computing a phylogeny

The construction of a sequence-based phylogeny involves the choice of a model that a particular computational method is designed to measure. Fredj Tekaia of the Institut Pasteur uploaded a PowerPoint discussion of "Molecular Phylogeny." The parts of the presentation that illustrate common alternative philosophical choices of computational methodology follow as text below. (For the full discussion, see: www.pasteur.fr/~tekaia/BGDA/Phylogeny.ppt)

"Phylogenetic tree construction methods:

"There are three main families of methods:
• Parsimony

• Distance Methods

• Maximum likelihood Methods"

"Methods directly based on sequences"
• Maximum Parsimony: find a phylogenetic tree that explains the data, with as few evolutionary changes as possible.

• Maximum likelihood: find a tree that maximizes the probability of the genetic data given the tree.
Methods indirectly based on sequences:

• Distance based methods:
(Neighbour Joining (NJ)): find a tree such that branch lengths of paths between sequences (species) fit a matrix of pairwise distances between sequences."

"Evaluation of different methods"
• None of the previous methods of phylogenetic reconstruction makes any guarantee that they yield the one true tree that describes the evolutionary history of a set of aligned sequences

• There is at present no statistical method allowing comparisons of trees obtained from different phylogenetic methods; nevertheless many attempts have been made to compare the relative consistency of the existing methods.

• The consistency depends on many factors, including the topology and branch lengths of the real tree, the transition/transversion rate and the variability of the substitution rates.

• In practice, one infers phylogeny between sequences which do not generally meet the specified hypothesis.

• One expects that if sequences have strong phylogenetic relationships, different methods will result in the same phylogenetic tree."

The classification process requires choosing one or more of the available computational methods. In practice, alternate methods often yield trees which are similar but not identical. Identical trees rarely result. A frequent comparison of alternate trees in the literature is describing the occurrence of joining nodes in common. Presentation of alternate complete tree diagrams together is less common.

Comparison of phylogenetic trees obtained by different methods:

Examples

The examples below are trees constructed from16S rDNA sequences. The examples demonstrate that the results of different phylogeny methods yield similar but not identical results. (Since the differences observed are properties of the algorithms, whole genome sequences will yield similar differences).

Neighbor-joining method and maximum-parsimony method: Zhang, J.-Y., et al., (2011) published an example of a tree constructed from two computations of the same data (Figure 1). Many trees supporting the naming of new species present the justification for the new name in the same way.

"Phylogenetic tree reconstructed by the neighbour-joining method showing the clustering of strain V2M29^T with other species of the genus Phycicoccus. Other related genera within the family Intrasporangiaceae are also included in the tree. Arthrobacter globiformis DSM 20124^T was used as an outgroup. GenBank accession numbers are given in parentheses. Numbers represent confidence levels (percentages higher than 50% are shown) from 1000 replicate bootstrap resamplings. Closed circles indicate those branches also found in the

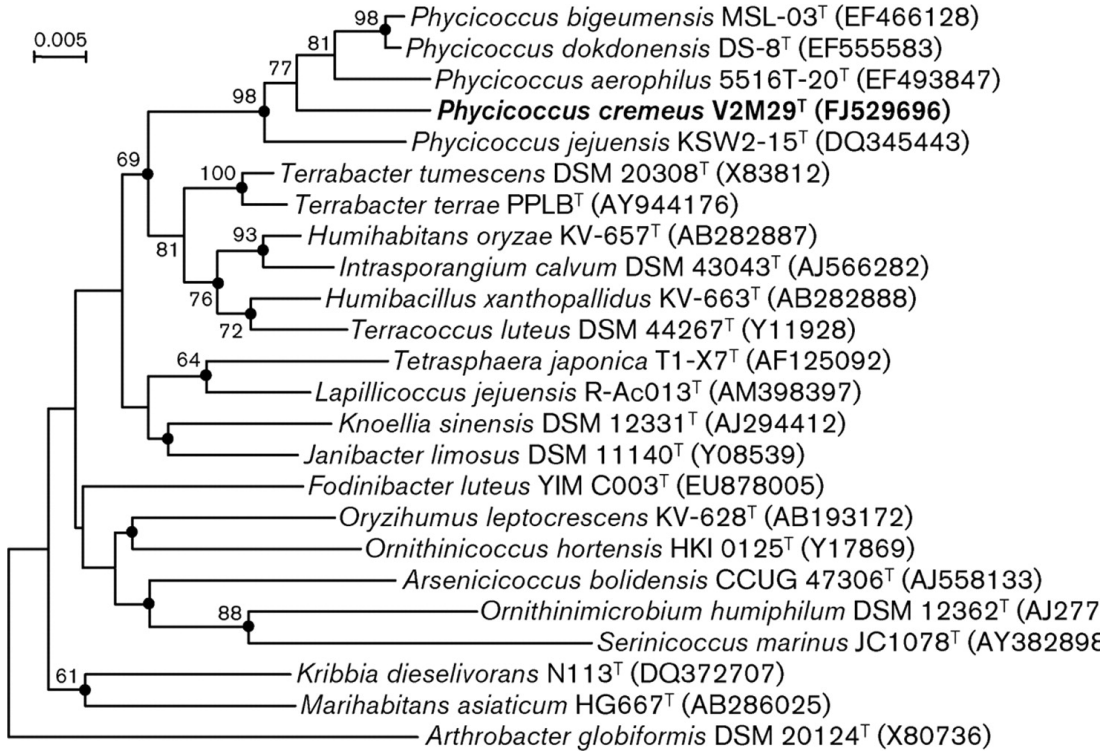


Figure 1. Phylogenetic tree reconstructed by the neighbour-joining method showing the clustering of strain V2M29T with other species of the genus *Phycicoccus*. (Reprinted with permission from Microbiology Society)

tree reconstructed by the maximum-parsimony method."

The tree constructed by the neighbor-joining method has 23 nodes. The maximum-parsimony method tree had 13 nodes in common. Thus, there are two trees representing the same data. It rests with the biologist to interpret the results.

Maximum Parsimony and Maximum Likelihood method: When both of the phylogeny trees, resulting from two methods, are available the trees can be compared in detail. The methods in this example are maximum parsimony and maximum likelihood (Coêlho, et al., (2011)).

Most clusters in the two results are similar in content and differ mainly in apparent branch extent (Figure 2). Consider the structural differences in the six member cluster beginning with Bacillus thuringiensis. The internal relationships within this cluster are structurally different. In the overall topologies, the placement of Bacillus polyfermenticus differs substantially. By Maximum Parsimony, it is an outlier, while the Maximum Likelihood calculation places it in a cluster of three. Which cladogram is the "better" is in the philosophical eye of the beholder.

The boundary problem

Use of whole genome sequences for determining similarity provides a powerful tool to determine relationships. A high degree of similarity provides confidence that two strains are members of the same taxonomic or phylogenetic group. As observed similarities diverge, the decision as to group membership is less clear. Members of closely related groups will have divided group loyalties and overlapping edges. Therefore, the boundaries between groups are not clear cut. Developing pragmatic criteria for deciding group membership is historically classification method dependent. Such criteria are critical to the use of taxonomies and phylogenies.

Paraphrasing John Locke (1690), People make species, nature does not, and people make boundaries of species, nature does not. The descriptive making of species and giving each a name is regulated by various bodies of the scientific unions (Krichevsky, M. 2005). There is no agreed organization that defines the criteria for establishing the boundary of a species. By itself, the whole genome sequence, or any part thereof, is a characteristic in isolation. To give the characteristic substantial meaning it must be compared to other sequences. Thus, a collection of whole genomes is subjected to some form of classification. Depending on the method of classification chosen by the classifier, the groups so formed will be given some group identifier. The class name may be a species name, clade designation, or simply a group number, etc. Construction of identifiers is subject to the preferences of the person(s) performing the classification. This overall process is basically the same as the historical practice of developing taxonomies based on phenetic or partial sequence information. Sneath and Sokal (1973) presented dendrograms that varied markedly depending on the algorithm used. Different algorithms will yield results that resemble each other but often will not be identical as exemplified by 16S rDNA sequences (Krichevsky, 2011). Borriss, et al. (2011), present the common alternatives for assembling a sequence and comparing it to others. Obviously, identical sequences will cluster together regardless of the algorithm selected. As the sequences diverge, so too will the resulting topology. Meinel and Krause (2012) studied 37 reported topologies constructed by a variety of methods. They describe determinants of the divergence in the topologies depending on the analytic data models used.

For various characteristics assessed (e.g., DNA hybridization, 16S rDNA, various phenetic commercial and private identification methods) microbiologists developed various decision planes for the various metrics of similarity. The wish to develop discrete boundary breakpoints is powerful, as we all would like certainty to facilitate placement decisions. Various methodology-dependent boundary numbers exist for the bacteria such as DNA hybridization data, about 70% (Wayne, et al., 1987), and for 16S rDNA sequence data, about 97% (Stackebrandt and Goebel,

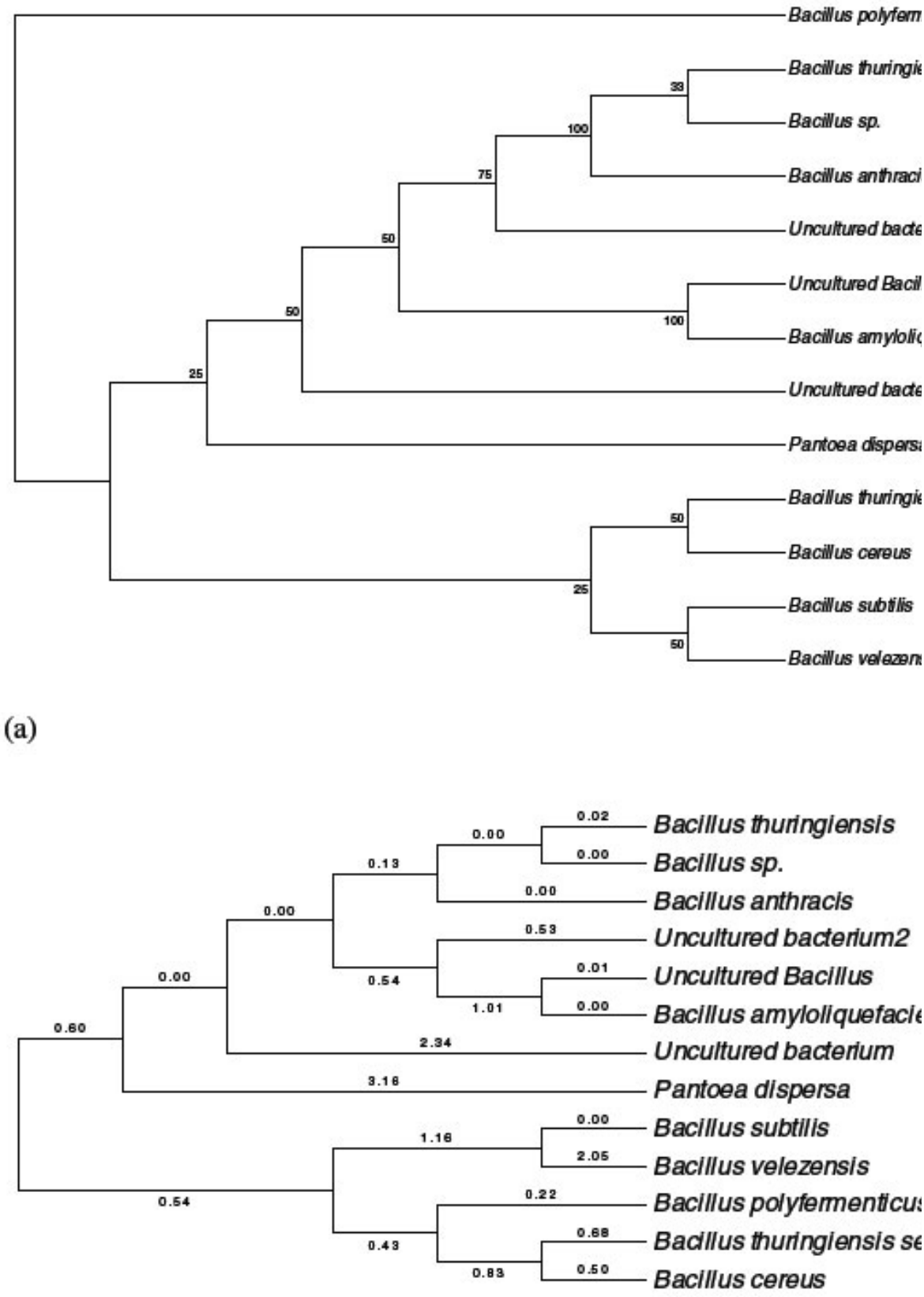


Figure 2. Phylogenetic analysis showing the relationship of the 16S rRNA gene sequences of isolates. The trees were generated using (a) Molecular Phylogenic analysis by Maximum Parsimony (MP) analysis of taxa (b) Maximum Likelihood (ML) method. (Reprinted with permission from the Microbiology Society)

1994). For commercial phenotypic identification kits the vendors use internal proprietary statistical criteria for defining membership in a species. These criteria have great practical utility. However, an objective discrete boundary escapes discovery. "Speciation is the evolutionary process by which new biological species arise." (From Wikipedia, the free encyclopedia) Accepting this definition, the evolutionary process, as the phrase implies, occurs due to genetic changes that do not simply jump over a discrete boundary fence into a new species domain.

Goris, J., et al., (2007) stated evidence for the lack of a discrete boundary as follows: "Our results…... together with data from other studies……., suggest that DDH values are continuous, i.e. theoretically, every value between 0 and 100% could be obtained in DDH experiments. These data are supportive of a continuous gradient of genetic relatedness rather than discrete species boundaries."

The boundary problem as a subset of the overlapping peak problem

Many examples of the need to resolve the areas and limits of overlapping peaks in variable data exist. The advent of computers allowed development of algorithms for separation of the areas into constituent components. Fields such as column chromatography, gel electrophoresis, geology, genomics, taxonomy, etc. have this need in common. This paper focuses mainly on prokaryotes. The issue is common to all biological speciation. For example, Kurtzman (2014) discusses in detail issues in circumscription of both genera and species of yeast.

A cursory Google search for "how calculate overlapping peak areas" yielded well over a million hits. The distribution of similarities in building a taxonomy or phylogeny is an example of the overlapping peak problem. As evolutionary events take place, the similarities form overlapping peaks which drift apart until the peaks are far enough apart to be considered distinct.

Figure 3 is an idealized example of two normally distributed overlapping variables. The frequency of occurrence in each variable, models the similarity values to the respective type strain of each of the peaks. (In the real world, the type strain may not be the most common genotype or phenotype. Further, the distribution of similarity values may not be normal.) The distribution shown represents an early stage of speciation. The underlying shape of the second variable is distorted by the numerical contribution of the first variable.

The late Stephen Lapage (Personal communication) stated that the updating of the probabilistic identification matrices used to define species at NCTC (National Collection of Type Cultures (NCTC), a Culture Collection of Public Health England occurred once a year. Any more often ran the risk of distortion by localized outbreaks causing submission of large numbers of the same strain. In such a case, the distribution would be skewed as a result.

Figure 4 illustrates the effect on perceived overlapping similarity distributions of isolates within two clusters as the relative frequencies of isolation change. Cluster 1 is a theoretical normal distribution of similarities to an ideally representative type strain. Cluster 2 is the same distribution displaced by similarity to a second type strain. That is, the frequency of isolation of members of the second cluster is identical in shape to that of the first. Most often, the frequency of isolation would be different for members of two related clusters. Multiplying the frequency values of Cluster 2 by 0.25 and by 0.1 illustrates the effect of disparate frequency of isolation of the second cluster members.

Using the observed frequency at the crossover points of the two lines, i.e., the minimum is a simple, and common, way to allocate the areas between the two clusters. However, the crossover point shifts as the frequencies of Cluster 2 accumulate. That is, as a corollary to Lapage's observation, the putative "discrete" breakpoint will vary as the relative frequency of isolation of the two "species" changes.

This example presents the simplest case wherein the two frequency distributions are symmetrically within each cluster. The actual events in speciation likely will distort the shapes making the relationships even more complicated. In general, as more examples of overlapping clusters are found, changes in their relative abundance will modify the shape of the

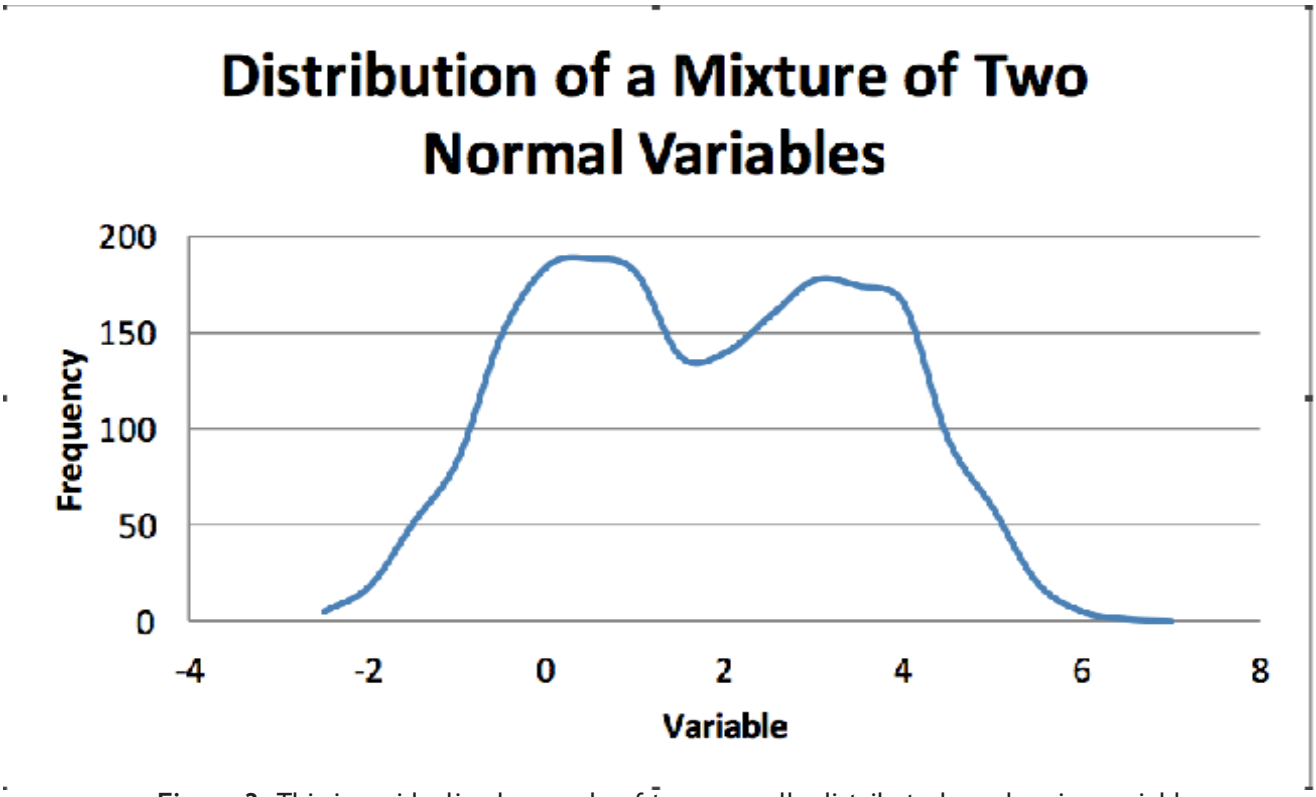


Figure 3. This is an idealized example of two normally distributed overlapping variables.

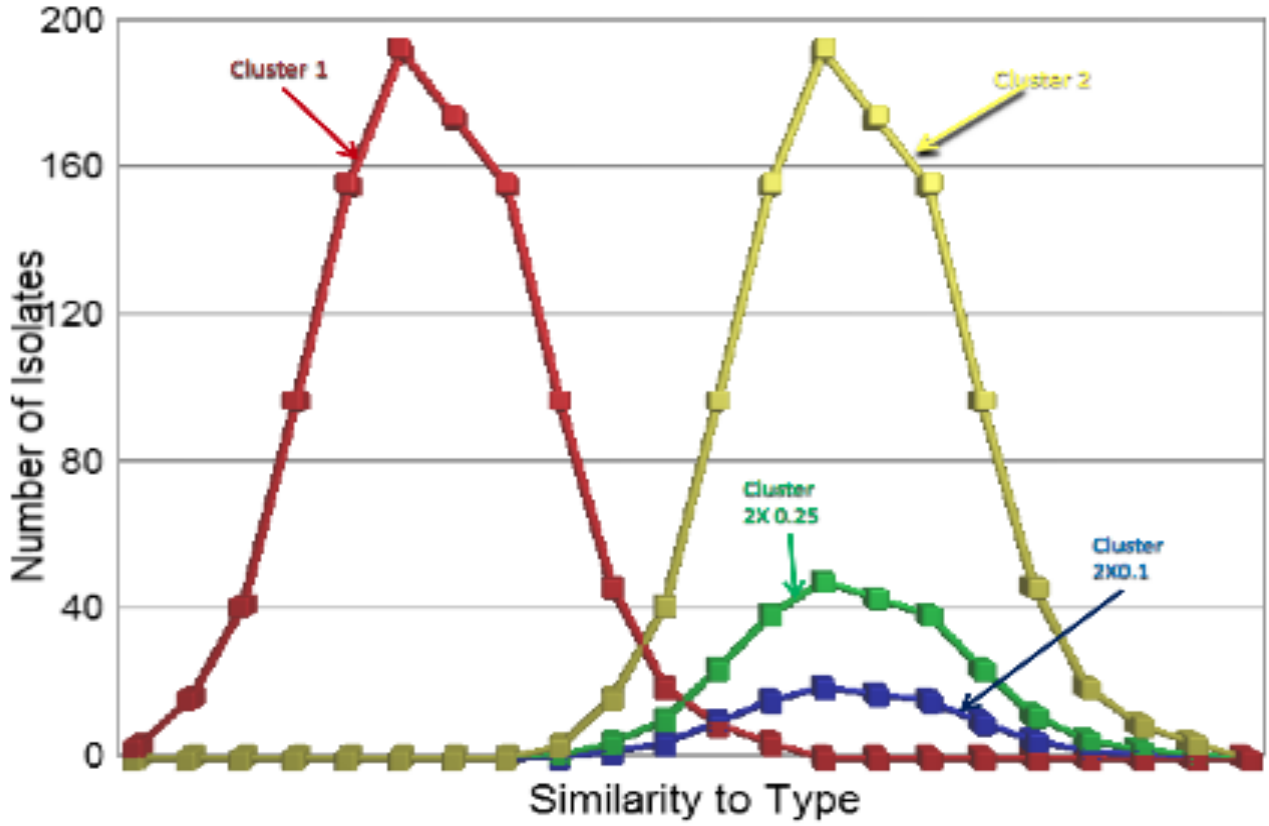


Figure 4. Effect of changing relative abundance of two "species."

frequency curve and shift the similarity minima in unpredictable ways.

This effect of distorting the perceived shape by changes in relative abundance of isolates in overlapping is independent of metric used to calculate similarity. The distortion is a mathematical property of the numbers of objects in each cluster measured.

A Geological Metaphor for the Boundary Problem
Geological examples illustrate the boundary problem for classification of related objects. They serve as a metaphor for degrees of boundary edges.

Edge clearly defined (Figure 5): At the top of the Australian monolith (named variously, Uluru / Ayers Rock) one is clearly on the Rock.(Ernieski. 1995 Uluru Sunset. Flickr photo library, see: https://www.flickr.com/photos/ernieski/). The edge clearly is defined as the junction of Rock and surrounding plain. The boundaries are apparent. Species boundaries are seldom as distinct.

Edge ambiguity (Figure 6): As depicted by this elegant old woodblock print (Hiroshige, U., ca. 1855-1858), the height of Mt. Fuji approaches zero roughly asymptotically. The boundary edges of the volcano are approximate. Species boundaries are similarly indistinct as more and more examples are discovered.

Overlapping peaks (Figure 7): (Original uploader was Lupinelawyer at en.wikipedia)
Both of the above geological formations are isolated geographically from related formations. The Presidential Range in New Hampshire is a good analogy for the edge problem in microbial classification. Each peak has a name of a USA president or historical person. There are hiking trails from one mountain to the next. As one walks the trails, there is no clear point defining where one mountain ends and another begins. Species overlaps result from a variety of contributing mechanisms; important are speciation and genetic exchange.

Opinionated Conclusions

As Edward Frenkel (2013) discussed in elegant detail, mathematical knowledge is independent of all human activity. Thus, mathematics alone is not subject to interpretation.

The components of a taxonomic classification can be represented by a cross section of a pyramid (Figure 8).

Observations: The base data are the actual observations or measurements made in the laboratory. These may be the results of individual tests, the actual macromolecular sequences found, the peaks on a chromatographic record, etc. All other information arising from these observations are interpretations or conclusions derived from the base data.

Annotation: The second layer has two very different components, context and communication.

Context: The database should contain, or point to, information as to the tests performed, the history of the object studied (where and how it was obtained), the conditions of the observations (when and by whom the observations were made, etc.) This information gives the context for the observations.

Communication: The motivation for obtaining data is to interpret the data in some meaningful way and communicate the interpretations to others. (“If you cannot communicate [your results] effectively, you are a hobbyist. No one pays the salary of hobbyists.” (W.A. Wood, personal communication to his graduate student)

Metadata: The concept of metadata refers to annotation of the database as a whole. This annotation describes the structure and overall contents of the database. Proper metadata summarizes the elements to the degree sufficient for a person (or computer) to access and manipulate the database contents. The metadata may be a part of the database itself or contained in a separate database. Thus, the metadata form a database describing one or more databases.

The dividing line between data and their interpretation is critical to reasoning about

Figure 5. Uluru / Ayers Rock

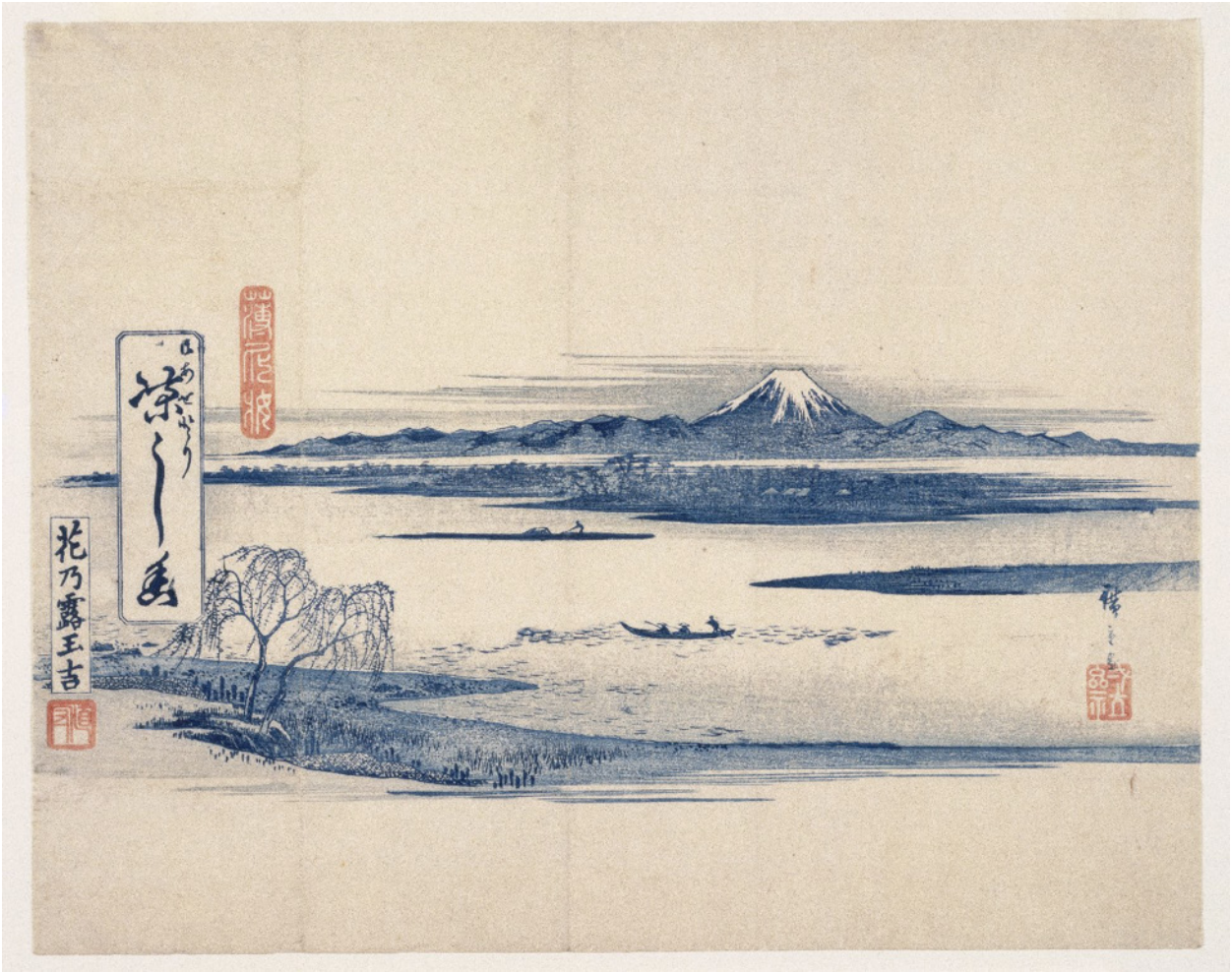


Figure 6. Woodblock print of Mt. Fuji, Japan

microbial information. All taxonomies, phylogenies and identifications are subjective interpretations of observations. Thus, there cannot be an objectively correct identification. Paraphrasing Hans Yu of Health Canada (personal communication), the community can agree on a consensus, thus lending credence to the identification.

The choice of a classification method protocol for taxonomy or phylogeny is subjective. For every step of building a taxonomy or phylogeny the developer makes professional judgments and interpretations. The first judgment is what to investigate, or as Monod (1971) termed the starting concept, an ethical choice. ".....the principle of objectivity as the condition of true knowledge constitutes an ethical choice and not a judgment arrived at from knowledge....."

The very initial study stages involve choices of the cohort of isolates, strains, and putative species. The developer should consider the quality level of the genomes required (and available) to support the aim of the investigation (Chain, 2009), methodologies of preparation of the DNA samples, choice of sequencing machines and sequence construction algorithms all are decisions of the developer.

The building of a taxonomy or phylogeny usually is a personal, an institutional, or at most, a multi-institutional effort. Within that context, consistency of methodological choices persists. Major taxonomic compendia such as Bergey's Manual, the Prokaryotes, that of the US National Library of Medicine, Ribosomal Data Project, and similar efforts for eukaryotes, are the work of either a small institutional group or sectional contributions by experts. Thus, they have strong components of the opinions of the various experts.

At each step in the process of generating a taxonomy or phylogeny alternative methodologies are available. Perhaps the most important philosophical decisions are choosing which machines and algorithm(s) to use in constructing the final classification. Chun and Rainey (2014) reviewed the properties of sequencing machines, the variety of software available for analyzing the data and relationships among the genomes, and the overall status of the available information on whole genome sequences that are publically available. Thus, the taxonomist has multiple tools from which to choose. In addition, they point out the disparity of the number of sequenced type strains of prokaryotes versus the number of species named, emphasizing the need for expanding the coverage of the prokaryote universe.

Each combination of method and metric yields a different display of relationships having varying degrees of resemblance. There is no truly objective means available for doing comparisons among methods (Tekaia). The resultant taxonomies or phylogenies are similar, but not identical, interpretations of the relationships among the sequences considered.

Taxonomies and phylogenies based on whole genome sequencing may overcome problems found with expression-based approaches and those of using single or a limited few sequenced genes. Still, questions remain as to whether this approach will significantly improve classification of bacteria, especially at the species level. Species of many genera of important bacteria have been difficult to differentiate because of a key feature of prokaryotic genomes: the propensity of isolates to exchange genes horizontally. Many important species survive in rapidly changing environments. They possess survival mechanisms that rely on shedding or acquiring genes under stress. Their pan-genomes, the total number of genes found in the gene pool of a set of genomes (Snipen, et al., 2009), may predominantly be comprised of lifestyle genes, needed for adaptation, as compared to the set of essential, mostly "housekeeping" components of their core-genomes. These non-core gene segments can readily be exchanged with closely related, and some not-so-closely related, isolates with the consequence that taxonomic boundaries could be blurred if entire pan-genomes are used for classifications. All taxonomic methodologies must accommodate this phenomenon, but whole genome sequencing and comparison of pan-genomes does little to obviate the problem. Comparisons of core-genomes may only limit the problem for sets of related taxa. Core-genome sizes appear to vary too much among the wide range of prokaryotes, such that a successful single method of core-genome

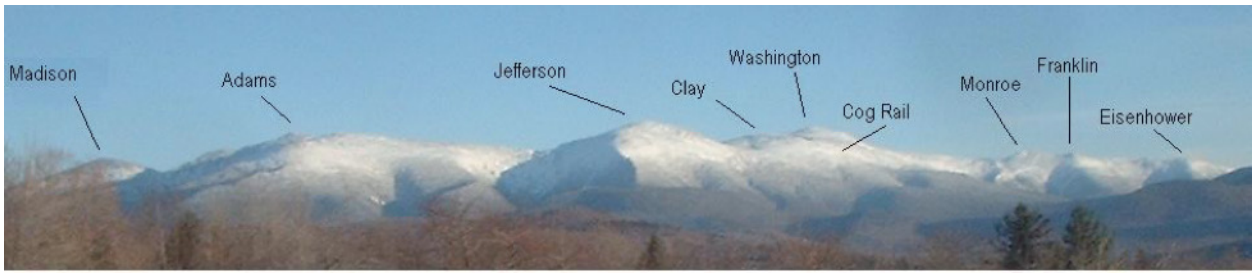


Figure 7. Presidential Range in New Hampshire

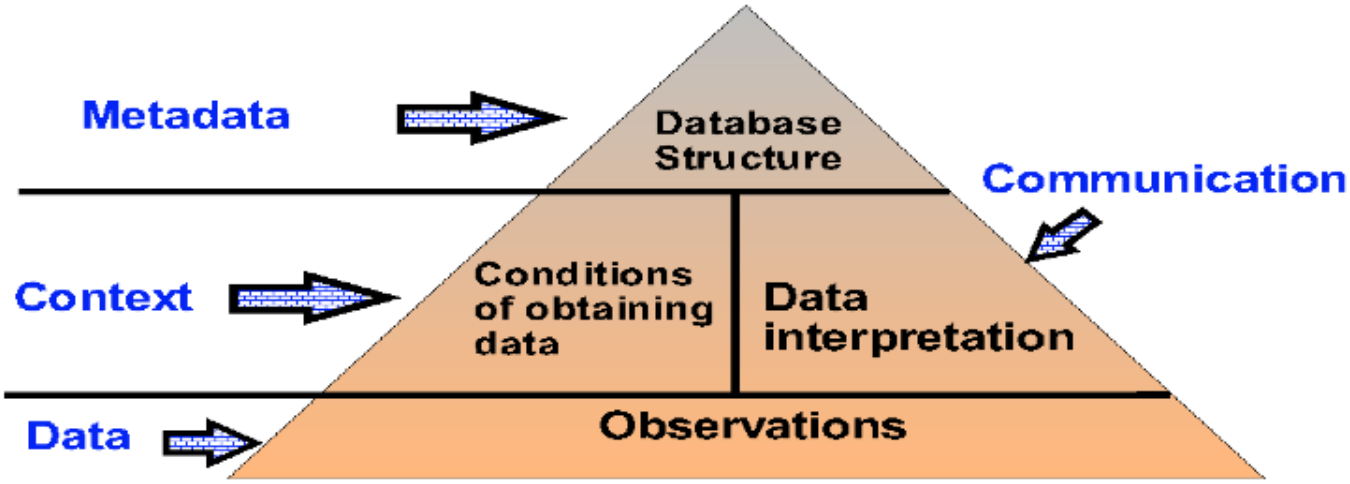


Figure 8. Representation of components of a taxonomic or phylogenetic database

comparison is unlikely to be devised. One taxon's core gene may be part of another taxon's lifestyle genes or "shell."

Most developers of taxonomies and phylogenies for cellular microorganisms utilize existing taxon names at some stage in the development. However, no body exists for the adjudication of the correct usage of a name (Krichevsky, 2004, 2005). Thus, competing taxonomies and phylogenies are permissible. Any taxonomy or phylogeny developed has no "official" standing. Paraphrasing Hans Yu of Health Canada (personal communication), the community can agree on a consensus, thus lending credence to the result.

The International Committee on the Taxonomy of Viruses (ICTV) (http://www.ictvonline.org) performs such a function through a structure of committees of experts that oversees the description and naming of new viruses. These committees have judicial roles in the detailed structure description of each virus, its name as well as its placement in the overall taxonomy of viruses. The final taxonomic decisions are the responsibility of the ICTV as a whole. The virologist community accepts the ICTV opinions as decisive. This system is unique in biology.

Use of nomenclatural annotation rather than de novo calculation to select cohort members

Various uses of annotation nomenclature substitute for wet laboratory work in constructing strain data cohorts. The usual reason is practicality.

For example, the justification for proposing a new species name involves demonstrating the difference of the new strain from a selection of similar (but not too similar) type strains. Similarly, in clustering, type strains that are dissimilar are

chosen as anchoring outliers. In both these cases, the original annotation of the type strains delineates which type strains are chosen for inclusion in the study. If the type strains are accompanied by a whole genome sequence, much laboratory effort is mitigated.

Identification systems utilizing annotation of reported nomenclature. The accumulation of large numbers of publically available whole genome sequences facilitates in silico studies. They often rely on the reported nomenclature to allow construction of a study cohort for establishing an identification utility. The strength of such systems is the richness of examples of sequences annotated with the same name. The weakness is the reliance on the judgment of name assignment by the annotators. Two examples follow:

Using annotated names from such rich sources as the National Center for Biotechnology Information (NCBI) [http://www.ncbi.nlm.nih.gov/genome] provides a method of approaching Yu's consensus. When enough examples of whole genome sequences with a common annotated name are available, the cohort with that name represents an operational definition of the named species. However, similar judgmental choices exist in this process as described for constructing taxonomies or phylogenies. The constructor of the resultant identification system must decide on, and demonstrate, the system's practicality for the purpose of the system. For example, the decision criteria for deciding boundaries between taxa and the rational for them should be clearly stated by the constructor since closely related clusters will overlap in an unpredictable manner as discussed below.

Two examples of whole genome-based identification systems illustrate the practical utility and differences of their approaches. Both utilize cohorts of whole genome sequences gleaned from NCBI. Both made use of the annotated taxon names to define the groups. Thus, the resultant groups represented the collective taxonomic opinions of original submitters of the sequences. Both author groups make their software available to perform identifications based on their databases. The methods described for determining identity are different.

Larsen, et al, (2014) constructed a data set from NCBI of 1,647 complete genomes (1,535 Bacteria and 112 Archaea). They curated the original taxonomic annotation utilizing various resources. Discrepancies were checked and corrected manually. Using a set of complete genomes, they developed and evaluated five identification methods. Four of the methods involved chromosomal core genes. The fifth, KmerFinder, examines the number of co-occurring k-mers (substrings of k nucleotides in DNA sequence data). The KmerFinder method exhibited the overall highest accuracy and correctly identified from 93% to 97% of the isolates in the evaluations sets. They stated that the results indicated that methods that sample only chromosomal, core genes have difficulties in distinguishing closely related species which only recently diverged.

Zhang, W., et al., (2014) compared pairs of the average nucleotide identity (ANI) of whole genome sequences of pairs of 1,226 bacterial strains. The goal was to determine whether such whole genome comparison can be used directly for bacterial species definition. Two bacterial strains from the same named species had a significantly higher ANI than those of two strains from different named species. They concluded that the ANI between the query and the reference genomes serve to determine whether two genomes come from the same named species. Bacterial species definition based on ANI with a cut-off value of 0.92 matched well (81.5%) with the current bacterial species definition. The average ANIs within the same named species among the 124 bacterial species ranged from 0.988 to 0.839. The ANI value was deemed to be consistent with the standard for traditional bacterial species definition, and it could be used in bacterial taxonomy for species definition. The value of 81.5% is the maximum accuracy rate observed for named species in the database. The metric of accuracy used is the ratio of correspondence to the annotated names in the source database. Considering the average ANI instead of the maximum within a taxon, the average ANI for two strains in the same species is 0.936, which is larger than that for two strains in the same genus (0.836) and also larger than that for two strains in the same family (0.789). The values have pragmatic utility for the particular data set. The values will change as new, or different, sequences

join the cohort.

The two examples above use published annotated whole genome sequences, albeit in different ways, to build an identification resource. Both rely on the name given the annotation as the reference for testing the efficacy of their algorithms. The Larsen, et al (2014) paper described editing procedures to validate the final nomenclature in the database. Zhang, W., et al., (2014) accepted the nomenclature as given in the NCBI database. While not explicitly stated, their threshold results for higher taxa also followed NCBI designations. Thus, both papers utilized clustering by names to form groups rather than a de novo clustering calculation. Such harvesting of groups by names becomes a practical necessity as the number of available whole genomes grows. These methods rely on the assumption that the taxon names are a useful method for establishing the species definitions. The use of names as a first step selection method introduces unknown level of informational noise. Again, such selection criteria will be a practical first step. For specific purposes, the genome sequences, per se, will still find use without regard to the bias introduced by using the names as the database selection criterion as well as for the challenge set.

Epilogue

The entire microbiological community depends on the professional opinions of experts (peer reviewers) to provide Yu's consensus. The aim is to fulfill the proverb of Confucius (~475 B.C.) (Loosely translated, see Ambrose (2013) from the original Chinese as: "The beginning of wisdom is to call things by their right name". Adamopoulos (1999) modified the proverb to give it biological context: "Old Chinese proverb: "the beginning of wisdom is to call things by their right names" – at least until the biology changes".

With respect to the role of the whole genome in the taxonomy or phylogeny of cellular microorganisms, the microbiological community could agree on a conventional protocol for "approved methodology" to allow more consistent comparisons of relationships among sequences produced by different investigators. The agreed protocol would generate the sequence used in constructing the "approved" classification. Alterative classifications from other protocols could be used as supplementary material as well. Peirce (1877) described four levels of belief. The "approved" classification approximates his third (the authoritarian) category wherein "Following the method of authority is the path of peace. Certain non-conformities are permitted; certain others (considered unsafe) are forbidden." Publishing a taxonomy or phylogeny establishes a descriptive authority. Indeed, non-conforming taxonomies and phylogenies are rampant in the literature. The methodological rigor used to construct the authority falls within the Peirce category of the scientific method.

To achieve a universal consensus definition of a "Species" and a classification, whole genome sequences may become a necessary component, but sequencing alone is not sufficient. Why expect that a discrete species boundary exists in nature? Assuming a consensus methodology evolves, other conditions will need to evolve as well. Some considerations follow.

Whole genome information, housekeeping gene sequencing, polyphasic analysis, phenotypic arrays, and so forth, provide complementary views of the relationships among microbes. The choice of the preferred view rests with the individual. Some want a pleasing, coherent classification. Others are interested in having a classification which helps them discern specific traits such as pathogenicity, production of products, etc. The distinction between these aspects blurs quickly. The microbiologist, as always, must build, adapt, or choose a taxonomy or phylogeny to fit the need. Thus, "I will know it when I see it."

Acknowledgement: This paper relied heavily on the critical reviewing and helpful suggestions by Mark C. Segal, US Environmental Agency. However, the author takes sole responsibility for opinions, suggestions and specific content.

References

Adamopoulos DA (1999) Old Chinese proverb: "the beginning of wisdom is to call things by their right names"--at least until the biology changes. Fertility and Sterility 71(5): 967-8.

Borriss R, Rueckert C, Blom J, Bezuidt O, Reva O & Hans-Peter Klenk (2011) Whole Genome Sequence Comparisons in Taxonomy. In Taxonomy of Prokaryotes, Methods in Microbiology , vol. 38, Rainey and Oren (eds.). Academic Press, Waltham; pp 409-436

Brenner D, Staley J, & Krieg N (2000) Classification of prokaryotic organisms and the concept of bacterial speciation. In Bergey's Manual of Systematic Bacteriology 2nd ed., vol. 1, Boone, Castenholz, and Garrity (eds.). Springer-Verlag, New York; pp 27-31.

Chain PSG, Grafham DV, Fulton RS, FitzGerald MG, Hostetler J, Muzny D, et al (2009) Consortium Genomic Standards Consortium Human Microbiome Project Jumpstart. and J. C. Detter. Science. Published in final edited form as: Oct 9, 2009; 326(5950): 10.1126/science.1180614.NIHMSID: NIHMS506003

Chun J and Rainey FA (2014) Integrating genomics into the taxonomy and systematics of the Bacteria and Archaea. Int. J. Syst. Evol. Microbiol 64: 316–324.

Coêlho MM, Ferreira-Nozawa MS, X Nozawa SR & Santos ALW (2011) Isolation of endophytic bacteria from arboreal species of the Amazon and identification by sequencing of the 16S rRNA encoding gene. Genet. Mol. Biol. 34(4): 676–680.

Ernieski (1995) Uluru Sunset - Flickr photo library. https://www.flickr.com/photos/ernieski/

Frenkel E (2013) Love and Math: The Heart of Hidden Reality. Basic Books, A Member of the Perseus Books Group, New York.

Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme P & J. M. (2007) DNA–DNA hybridization (DDH) values and their relationship to whole-genome sequence similarities. Int. J. Syst. Evol. Microbiol 57: 81–91.

Hiroshige U, ca. 1855-1858 Woodblock print of Mt Fugi File:Brooklyn Museum – "Cool Picture View of Mt. Fuji"

Kisand V and Lettieri T (2013) Genome sequencing of bacteria: sequencing, de novo assembly and rapid analysis using open source tools. BMC Genomics 14: 211.

Krichevsky MI (2004) Taxonomy: a Moving Target for Sequence Data. In Database Annotation in Molecular Biology: Principles and Practice, A. M. Lesk, (ed.). John Wiley & Sons, Ltd, Chichester.

Krichevsky MI (2005) Codes of Nomenclature. Nature Encyclopedia of Life Sciences. John Wiley & Sons, Ltd. Chichester.

Krichevsky MI (2011) What is a bacterial species? I will know it when I see it. The Bulletin of BISMiS 2: 17-23.

Krichevsky MI et al. (2013) What Is A Bacterial Strain? I Will Know It When I See It. The Bulletin of BISMiS, 4: 7-14.

Kurtzman CP (2014) Use of gene sequence analyses and genome comparisons for yeast systematics. Int. J. Syst. Evol. Microbiol. 64: 325–332.

Larsen MV, Cosentino S, Lukjancenko O, Saputra D, Rasmussen S, Hasman H, et al. (2014) Benchmarking of Methods for Genomic Taxonomy. J. Clin. Microbiol. 52: 1529–1539.

Locke J (1690) An Essay Concerning Human Understanding, Book 3, Chapter V, Of the Names of Mixed Modes and Relations (Published multiple times)

Lou DI, Hussmann JA, McBee RM, Acevedo A, Andinoc R, W.H. Press & Sawyer. SL (2013) High-throughput DNA sequencing errors are reduced by orders of magnitude using circle sequencing. Proc. Natl. Acad. Sci. U. S. A.: 110: (49) 19872-19877.

Lupinelawyer at en.wikipedia Presidential Range (Original uploader)

Maiden MC, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, et al. (1998) Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proc. Natl. Acad. Sci. U. S. A. 95: 3140–3145.

McElroy K, Thomas T, and Luciani F (2014) Deep sequencing of evolving pathogen populations: applications, errors, and bioinformatic solutions. Microbial Informatics and Experimentation 4:1

Meinel T and Krause A (2012) Meta-Analysis of General Bacterial Subclades in Whole-Genome Phylogenies Using Tree Topology Profiling. Evolutionary Bioinformatics 2012. 8: 489–525

Monod J (1971) Chance and Necessity: An Essay on the Natural Philosophy of Modern Biology, Austryn Wainhouse (translator). New York Vintage, New York.

Ozen AI, Vesth T, & Ussery DW. Comparative Genomics (2013) In The Prokaryotes: Prokaryotic Biology and Symbiotic Associations, Rosenberg, DeLong, Lory, Stackebrandt. and Thompson (eds.), Springer, Heidelberg; pp. 209-227.

Peirce CS (1877) The Fixation of Belief. Popular Science Monthly 12:1-15.

Sneath PHA. and Sokal RR (1973) Numerical Taxonomy: The Principles and Practice of Numerical Classification. Freeman, San Francisco.

Snipen L, Almøy T & Ussery DW (2009) Microbial comparative pan-genomics using binomial mixture models. BMC Genomics 10: 385.

Stackebrandt E and Goebel BM (1994) Taxonomic Note: A Place for DNA-DNA Reassociation and 16s rRNA Sequence Analysis in the Present Species Definition in Bacteriology. Int. J. Syst. Evol. Microbiol 44: 846-849.

Staley JT (2009) Universal species concept: Pipe dream or step toward unifying biology? J. Ind. Microbiol. Biotechnol. 36: 1331-1336.

Tekaia F see: www.pasteur.fr/~tekaia/BGDA/Phylogeny.ppt

Wayne LG, Brenner, DJ, Colwell RR, Grimont PAD, Kandler O, Krichevsky MI, et. al. (1987) Report of the Ad Hoc committee on reconciliation of approaches to bacterial systematics. Int. J. Syst. Bacteriol. 37:463-464.

Willcox WR, Lapage SP, Bascomb S, & Curtis MA (1973) Identification of bacteria by computer: theory and programming. J. Gen. Microbiol. 77, 317-330.

Yamamoto S and Harayama S (1995) PCR amplification and direct sequencing of gyrB genes with universal primers and their application to the detection and taxonomic analysis of Pseudomonas putida strains. Appl. Environ. Microbiol. Mar; 61(3):1104-9. Erratum in Appl. Environ. Microbiol 1995 Oct; 61(10):3768.

Zhang JY, Liu XY, & Liu SJ (2011) Phycicoccus cremeus sp. nov., isolated from forest soil, and emended description of the genus Phycicoccus. Int. J. Syst. Evol. Microbiol. 61: 71-75.

Zhang W, Du P, Zheng H, Yu W, Wan L, & Chen C. (2014) Whole-genome sequence comparison as a method for improving bacterial species definition. J. Gen. Appl. Microbiol. 60:75-78.