BRIEFINGS IN BIOINFORMATICS. VOL 7. NO 2. 186–195
Advance Access publication April 24, 2006
doi:10.1093/bib/bbl005

Computational aspects of systematic biology

Timothy G. Lilburn, Scott H. Harrison, James R. Cole and George M. Garrity

Received (in revised form): 25th February 2006

Abstract

We review the resources available to systematic biologists who wish to use computers to build classifications. Algorithm development is in an early stage, and only a few examples of integrated applications for systematic biology are available. The availability of data is crucial if systematic biology is to enter the computer age.

Keywords: *taxonomy; nomenclature; databases; semantics; persistent identifiers*

INTRODUCTION

To most contemporary biologists, the concept of taxonomy conjures up visions of dusty museum cabinets filled with pressed plant parts, insects mounted on pins and bottles of formalin-fixed soft-bodied animals; gathered by naturalists in bygone times and serving the sole purpose of introducing school children to the wonders of science. Yet, taxonomy, or more appropriately systematic biology, is a vibrant field that continues not only to serve as the underpinnings of all modern biology, but also serves to reshape our thinking about the possible relationships of all living and extinct species.

Systematic biology has been defined as 'the scientific study of the kinds and diversity of organisms, and of any and all relationships among them' [1]. A main goal of systematic biology is to produce a classification (see subsequently). Theories of classification can be traced back to Aristotle, whose ideas about 'essentialism' guided efforts at biological classification for centuries. Carl von Linné (better known by his Latinized name, Linnaeus) revolutionized systematics when he introduced his system in the mid-eighteenth century, giving biologists a clear methodology for creating classifications and a set of rules for producing the needed nomenclature. Linnaeus was also an essentialist and although theories of classification now rest implicitly on the theory of evolution, the Linnaean system is still almost universally used by biologists. Ereshefsky [2] gives a lucid account of the history and current state of taxonomic theory. Modern classifications are usually based on a type of relationship. These can be practical, for example, a set of organisms might be classified according to the types of diseases they produce in humans, or they can be based on estimates of relatedness that reflect shared characteristics. The latter case is far more widely seen and there has been a great deal of controversy over how these estimates of relatedness should be produced and used. Estimates of relatedness are derived in two major ways:

(i) according to the shared evolutionary history of the organisms to be classified—evolutionary taxonomists and process cladists explicitly posit evolutionary relationships as the basis for classification, although they differ as to how these evolutionary relationships should be determined.

Corresponding author. George M. Garrity, Department of Microbiology and Molecular Genetics at Michigan State University, East Lansing MI. Tel: +1 517 432 2459; Fax: +1 517 432 2458. E-mail: garrity@msu.edu

T. G. Lilburn, PhD, is a research scientist in the Department of Bacteriology at the American Type Culture Collection in Manassas, VA.

Scott H. Harrison is a doctoral candidate in the Department of Microbiology and Molecular Genetics at Michigan State University.

James R. Cole, PhD, is an Assistant Professor in the Center for Microbial Ecology and manager of the Ribosomal Database Project (RDP-II).

G. M. Garrity, ScD, is a professor in the Department of Microbiology and Molecular Genetics at Michigan State University and Editor-in-Chief of Bergey's Manual Trust.

(ii) according to the number of characteristics shared by the organisms to be classified—pheneticists and pattern cladists reject evolutionary theory and search only for similarity. However, pattern cladists use only characteristics that they consider to be meaningful when comparing organisms while pheneticists consider all characteristics and may either give them all the same weight when comparing organisms or use a weighting scheme based on the input data.

For most life scientists, a classification has a purely functional purpose—it summarizes similarities among groups of organisms so as to facilitate predictions about members of a group. Most often, classifications are derived from phylogenetic analyses. We will restrict this review to resources for systematics, with special emphasis on the needs of end users of the classifications. Because there have been numerous recent reviews of the software and methods used in phylogenetics [3], we will not consider them comprehensively. The driving force in modern systematic biology is, or ought to be, the wealth of new data that floods the databases and serves as the input for modern phylogenetic models. Together, these data and models are radically changing our thinking about the processes and patterns of evolution and are beginning to elucidate the mechanisms that give rise to the molecular structures and functions that form the basis of organismal phenotypes.

The need for algorithmic approaches to systematics

For the most part, systematic biology is still using tools that would be familiar to Linnaeus. Computers are used mainly to create databases of characters relevant to relatively few taxa or, more recently, to try and encompass biodiversity by building and distributing large lists of taxon names. The compilation of these lists has made it clear that, because algorithmic approaches to classification, and even identification, have been frequently rejected by practitioners of systematic biology, the number of taxa to be described far exceeds the existing capacity for describing new taxa. Estimates of described eukaryotic taxa are in the range of 1.7 million species, and the total number of species has been estimated at between 3.5 and 10.5 million [4]. However, these estimates must be viewed with caution as there is no universal concept of either a biological species or higher taxon. Taxon description is rendered more complicated by the need for qualified practitioners to do routine identifications of the already described species and by the need to identify the same taxa under different names (synonymies). Computer applications could reduce this gap considerably. For example, described taxa could be identified using universally applicable methods based on molecular sequences (see barcode methods) or using pattern recognition software. Such algorithms could readily detect and provide a preliminary filter for taxa that conform to previously described taxa, leaving experts free to concentrate on the real issues of circumscribing and describing new taxa [5]. Synonymy can arise through multiple independent descriptions of the same taxon or through reclassification. At present, ~22% of the validly published names applied to prokaryotes are synonyms and the incidence of synonymy in plant and animal names is expected to be considerably higher, but much more difficult to ascertain as the relevant codes of nomenclature do not provide for a centralized control of names. Database technologies would seem well-adapted to dealing with these issues, providing that a good model for such occurrences either exists or can be constructed and that transitive closure and semantic resolution of such events can be readily accomplished in a consistent manner.

Algorithms designed to automate classification and identification require data and these tend to be distributed piecemeal across the scientific community. Furthermore, most of the current databases are generally not interoperable. Tools that integrate data handling and data analysis have become crucial to the successful practice of systematics. In this sense, many of the problems facing systematic biology are informatics problems, such as developing flexible and comprehensive data models, providing methods for dealing with constantly changing nomenclatures and ensuring database interoperability.

This review will survey the resources available to systematic biologists who want to use computers to build classifications. The two most important resources, as already mentioned, are data from the taxa to be classified and algorithms for constructing a classification from the data. Of course, in most cases a new classification is not needed; the biologist simply wishes to place a new taxon into an existing framework, so sources for information about an

existing classification will be mentioned as well. Automated systems for identifying and classifying the higher eukaryotes appear to be following two trajectories: one is based on digital image analysis and the second on molecular sequence comparison. Sequence-based efforts, like Barcodes of Life (http://barcoding.si.edu/index_detail.htm), are still in a nascent state, so it might be instructive to examine the experience of prokaryotic biologists in developing integrated identification and classification systems. After 15 years of development, such systems are now becoming available and, through their successes and shortcomings, illustrate what properties a fully functional classifier might have.

ELECTRONIC RESOURCES FOR SYSTEMATICS

Electronic resources for systematics may be divided into two classes. The first are those that provide information about taxa, including taxonomic information. Resources in this class require that the user already knows the identity of the taxon. The information found here could be used to identify an unknown (say via images or textual descriptions), but no algorithms are supplied. In the second class, we place those resources that enable an investigator to identify an unknown (which would allow it to be placed into a reference classification) or to create a classification that includes the unknown and extant taxa.

Data collections

In order to explore diversity and arrive at a definitive picture of 'what is out there', researchers have recognized that we need an authoritative and non-redundant list of previously described species. This need has led to the establishment of several databases and data portals. The simple establishment of what are often described as lists of species has proven, perhaps, to be a much more daunting task than was initially foreseen.

- Portal sites provide access to information about organisms, including a taxonomy, geographic distribution, where the voucher specimens exist, databases containing information relevant to the organism, etc. [e.g. Global Biodiversity Information Facility (GBIF)]. Some (e.g. Digital Taxonomy) also link to information about data standards, software and other basic information.
- Databases are similar to portals, but are more focused on specific taxa (e.g. AlgaeBase) or on a certain aspect of systematics (e.g. the International Plant Names Index).

These sites do not generally provide real classification and identification services. Rather, they serve up lists of names and provide some rudimentary information about the covered taxa. Neither portals nor databases, in our definition, supply tools to aid in the classification of organisms, that is, they are static. A few of the more prominent services are listed in Table 1.

Identification and classification

Identification places an organism within an established classification. We subscribe to the definition of Sneath and Sokal [6] who regarded classification as

Table 1: A partial list of online resources for systematics

Databases and portals	
SPECIES 2000	http://www.usa.sp2000.org/
Global Biodiversity Information Facility (GBIF)	http://www.gbif.org/
ITIS—Integrated Taxonomic Information System	http://www.itis.usda.gov/
The International Plant Names Index (IPNI)	http://www.ipni.org/
Index Fungorum	http://www.indexfungorum.org/
AlgaeBase	http://www.algaebase.org/
The Universal Biological Indexer and Organizer (uBio)	http://www.ubio.org/
The NCBI Taxonomy database	http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/
BioNET-INTERNATIONAL, The Global Network for Taxonomy	http://www.bionet-intl.org/
Biodiversity and Biological Collections Web Server	http://www.biocollections.org/
Digital Taxonomy	http://digitaltaxonomy.infobio.net/?Home
The Tree of Life	http://www.tolweb.org
Web tools	
The Taxonomicon	http://sn2000.taxonomy.nl/Taxonomicon/Default.aspx
Phylomatic	http://www.phylodiversity.net/phylomatic/

'the ordering of organisms into groups (or sets) on the basis of their relationships'. The term is also used as a noun; a classification is the result of the act of classification. We divide the software applications that follow according to whether they provide identifications or classifications. Computers began to be used in systematics soon after they became more generally available in the 1950s and 60s and were quickly embraced for both purposes. Many of the uses envisioned for computers at that time have only recently become feasible. It seems obvious that, in principle, any software that can identify an organism could be extended to classify organisms as well, but few applications go beyond the identification, based on the previously established criteria.

Identification

Phenotype-based DELTA—DEscription Language for TAxonomy (http://delta-intkey.com/). These tools for creating classifications and identification keys have been very popular [7], but development of DELTA was discontinued in 2000. Designed as a single-user tool, it was felt that DELTA would not be able to meet the demands of the collaborative, multi-user efforts that are seen as the future of systematics. An open source version, Free DELTA is actively under development and a proposed replacement, BioLink (http://www.biolink.csiro.au/), is available.

Linnaeus II (http://www.eti.uva.nl/products/linnaeus.php) was developed by ETI Bioinformatics to support the development of taxonomic databases and identification keys by researchers.

There are automated identification applications for certain taxa. Two examples of such systems are the Digital Automated Identification SYstem (DAISY) (http://chasseur.usc.edu/pups/projects/daisy.html), which was initially developed for insect identification [8, 9], and the Automated Bee Identification System (ABIS) [10, 11]. DAISY is now being developed as a universal identification system, but ABIS is not available. Both use machine learning approaches to pattern recognition, training the software to recognize taxa based on digital images. There are drawbacks to the pattern recognition approach. Although the approach approximates the morphometric methods used by humans to identify taxa, computers are less flexible in the way images are interpreted and consequently are more likely to fail to arrive at the same identification. The way the software is trained is critical because a poorly chosen training set can lead to incorrect identifications.

There are many manual and automated approaches to the identification of the prokaryotes. A recent review [12] lists 30 commercially available kits and instruments for the *Enterobacteriaceae* alone. Typically, the tests are inoculated by a technician and after a fixed length of time for growth the instrument compares the growth/no-growth patterns in a diagnostic table (Data Matrix) with a database to produce an identification. The approach is probabilistic, using a Bayesian approach to estimating the likelihood that the isolate belongs to the suggested taxon and determining whether or not the identification might be a definitive one. These approaches were developed by Lapage *et al.* [13] in the 1970s. Recently the technique of phenotypic microarrays has been developed for testing prokaryotes for up to 2000 properties [14]. Biolog, Inc., the developers of the approach, have also sought to find a set of tests that would be useful in comparing all *Bacteria*, thus enabling the calculation of phylogenetic relationships based on phenotypic properties [15]. Historically, such biochemical and nutritional tests have been targeted at identifying members of specific groups of *Bacteria* and have not, therefore, been useful for establishing evolutionary relationships among diverse species of bacteria.

Molecular sequence-based Recently, the Consortium for the Barcodes of Life has proposed an alternative to classical means of identifying eukaryotes dubbed 'DNA barcoding' [16, 17]. Relying on the sequencing of one or a few universally distributed genes to identify the organism, DNA barcoding has been proposed as a rapid, inexpensive and generally applicable way of identifying a broad range of organisms. A concerted effort is now underway to establish the necessary databases, tools and information management systems needed to make this approach successful (http://barcoding.si.edu/index_detail.htm). The barcoding approach mimics the approach taken by microbiologists for the last 15 years in identifying prokaryotes using the sequences of small subunit ribosomal RNA (SSU rRNA) genes. The Ribosomal Database Project (RDP; http://rdp.cme.msu.edu/index.jsp) has been collecting ribosomal RNA sequence data for about 12 years and supplies tools for sequence comparison, phylogenetic analysis and probe matching, among others. Researchers often use this database to obtain

a preliminary identification of a newly isolated prokaryote. Another project, Ribosomal Differentiation of Microorganisms (RIDOM) (http://www.ridom-rdna.de/index.html), is designed to help researchers identify bacterial and fungal pathogens [18]. Researchers can upload sequences or chromatograms. In the latter case, an interface with the phred/phrap sequence assembly tools is provided. This allows users to edit their sequences in order to raise the confidence level of their identifications. RIDOM also provides information on the appearance of the organism, the diseases it is associated with and the nomenclature. In cases where the ribosomal RNA sequence does not contain sufficient information to provide an identification, RIDOM also supplies information on other means of identifying the bacterium.

A more recent trend in the identification of prokaryotes and single-cell eukaryotes is the use of multi-locus sequence typing (MLST) [19]. This technique allows discrimination among organisms below the species level, which makes it very useful for epidemiological studies. The technique detects allelic variation among strains by amplifying and then sequencing seven or more well-conserved genes. The technique allows strain characterizations to be made over the internet through DNA sequence comparison with data that is available from 34 different MLST databases (http://pubmlst.org/) [20]. Twenty-one of these databases are affiliated with the MLST website (http://www.mlst.net) at the Imperial College, London [21]. The 'mlst.net' also supplies software for analysis of the relationships among strains.

Classification

Phenotype-based Although molecular sequence-based analysis has come to dominate most discussions of classification, even to the point that it has been suggested that morphology is not to be used to produce classifications [22], a strong case can be made for the continuing importance of some morphological features in classification [23, 24]. There are a few widely used software packages that can be employed to analyse phenotypic data using a cladistic/phylogenetic approach. These include PAUP* [25], PHYLIP [26], MacClade [27] and Mesquite [28]. Although the phenetic approach has fallen from favour, there are at least two software packages that allow researchers to take this approach to morphological data, even though general statistics packages like S-Plus (Insightful, Inc., Seattle, WA) offer similar functionality. These packages are NTSYSpc (Numerical Taxonomy System 2.2) [29] which takes a phenetic approach, and can be used to discover pattern and structure in multivariate data, and SYN-TAX [30], another suite of software for multivariate data analysis.

Molecular sequence-based Today, most classifications are produced using cladistic/phylogenetic analyses of molecular sequence data. A great deal of software is available for this purpose. Since this software is regularly reviewed, we will mention only a few services and software packages that are explicitly intended to integrate phylogenetic analysis and classification. The data used in these analyses are readily available from the major sequence databases (GenBank, EMBL and DDBJ). All three of these databases use the National Institute for Biotechnology Informatics (NCBI) Taxonomy for annotating their sequence data. This taxonomy is regularly updated by NCBI, with up to 100 new species added each day. Corrections based on internal research and outside information are also made, but NCBI makes no claims for the authority or accuracy of the taxonomy and classification. Many sequences are associated with incompletely classified organisms. Nevertheless, this classification and taxonomy provides a convenient way to start exploring various taxonomic groups and to assemble data sets for phylogenetic analysis. All three databases provide tools for browsing the classification and taxonomy [Taxonomy Browser (http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/index.cgi), TAXY (http://www.ebi.ac.uk/taxy/index.html) and TXSearch (http://sakura.ddbj.nig.ac.jp/uniTax.html)]. The NCBI taxonomy is being widely adopted by databases such as PIR (http://pir.georgetown.edu/home.shtml) and BIND (http://www.bind.ca/Action) that use the taxon ID as a permanent identifier. This could, however, be problematic as NCBI does not regard these identifiers as permanent and indicates that taxon IDs can disappear as a result of various taxonomic acts.

The self-organizing self-correcting classifier (SOSCC) of Garrity and Lilburn [31] is an algorithm that builds classifications from molecular sequence-based distance matrices and uses dynamically reordered heatmaps [32] to visualize the results (http://taxoweb.mmg.msu.edu/). The algorithm allows one to review and modify the classification, identify possible classification errors and facilitates *ad hoc* testing of alternative classifications/hypotheses.

It also reveals higher taxonomic structure in the underlying data. Partial functionality of this algorithm is available on the 'Taxomatic' website, along with taxonomic visualizations created with this algorithm and by statistical analysis of evolutionary distance data.

Web Accessible Sequence Analysis for Biological Inference (WASABI) (http://www.lutzonilab.net/aftol/) is a tool for participants in the Assembling the Fungal Tree of Life (AFToL) project [33]. WASABI automates the tasks of data assembly and quality assurance for sequences submitted by participants in the AFToL project. It also carries out phylogenetic analyses based on these sequences. Partial functionality, in the form of a BLAST search of user-submitted data sets, is available through this website.

The *mor* (http://mor.clarku.edu/) [34] is also part of the AFToL project. It automates the process of finding, screening and analysing fungal large subunit rRNA sequences in the public databases. *mor* targets one group of fungi, the homobasidiomycetes, and generates classifications of this group by interpreting the generated trees to produce node-based definitions of the taxa that are consistent with the PhyloCode [35]. The PhyloCode (http://www.ohiou.edu/phylocode/) is a set of rules or procedures for producing taxon names that are consistent with the phylogeny.

emerencia (http://emerencia.math.chalmers.se/) [36] is a third program that is being used to identify, if not classify, fungi based on molecular sequences. It scans GenBank accessions for sequences that are from unidentified or insufficiently identified taxa and periodically screens those sequences against a set of regularly updated sequences from known organisms.

RDP-II (http://rdp.cme.msu.edu/index.jsp) [37] is a set of web-based services that comprise perhaps the most widely-used software for the identification of organisms. This project started as a database of SSU rRNA sequences from all taxa in 1992 and over the years has added services that allow users to find the best match for their own sequence. The RDP-II also provides a classification service, using a naïve Bayesian classifier and provides multiple taxonomic views, including the nomenclatural taxonomy of prokaryotes distributed by Bergey's Manual Trust [38]. The collection, alignment and classification of database sequences are fully automated, with the result that this is probably the largest sequence database devoted to identification and classification. Work is underway to extend the existing pipeline to include the SOSCC algorithm and other tools for exploring prokaryotic taxonomy that are currently found on the 'Taxomatic' (http://taxoweb.mmg.msu.edu/) website.

An emerging trend in classification that is worth-mentioning is the use of whole genome phylogenies. Genome phylogenies are constructed in two basic ways. The first approach uses whole genome features, that is, gene order, gene content or a statistic of the entire genome. The second can be thought of as sequence-based; a set of orthologous genes is used to build a phylogeny, either by concatenating the genes, aligning them and building a tree or by calculating trees for each gene and then combining the trees into a single 'super-tree'. Recent reviews cover this emerging field quite thoroughly [39–41]. Variations and innovations continue to appear, as well [42–46]. As the cost of sequencing decreases, it is anticipated that this approach will become more useful and more widely applied.

DISCUSSION

Current state of computational approaches to systematic biology

On the whole, most available applications must be seen as prototypes in which some necessary components work, and others do not. Some of the most ambitious projects [47, 48] fall into this category. Navigation to the software resources mentioned here will quickly confirm this view. We should also extend our definition of prototype to include projects of limited taxonomic scope. No matter how well-implemented such projects may be, their tight focus effectively makes them prototypical. However, if they are readily adaptable to other taxa they might serve as models for the development of more inclusive projects.

The RDP, for example, is a relatively mature service that began as a collection of sequence data. During the first years, it supplied only these sequence data to users. Later, the service supplied tools that used these data to allow end-users to retrieve data more conveniently and to identify user sequences by comparing them with those in the database. Taxonomic information about the organisms from which the sequences were obtained was added in 2001 and 2 years later the Bayesian classifier was added. Thus, its systematic biology services have been developed gradually, even though the impetus for building the database in the first

place was to facilitate the phylogenetic classification of the prokaryotes. With the addition of the SOSCC service, the RDP-II will become a truly integrated systematic biology service. Like the RDP-II, most current efforts to address problems in systematic biology using computational tools recognize the need for integrated solutions, built on databases. At present, the larger initiatives are focused on gathering nomenclatural information for all the known species. The developers are also planning to accommodate new taxa; GBIF hopes to capture the authoritative names for 40% of all the known species by the end of 2005 [48]. Smaller, more focused, initiatives are able to go beyond nomenclature in the capture of data relevant to the classification of the organisms they are focused on. The AFTOL project, like the RDP-II, is not only using sequence data to guide the classification of the fungi, but also to database morphological data. The Taxonomicon makes an effort to link various metabolic, ecological and phenotypic hierarchies [47].

Issues in designing and implementing systems biology applications

If all the data relevant to a classification are to be integrated, interoperability among databases and services is essential. GBIF has taken great care to develop data models and protocols that facilitate interoperability [49]. How linkage may reliably proceed between databases, and how there may be an efficient, ongoing process of correction and resolution of missing entries and revisions or extensions to underlying data models are the questions they are facing and that need to be dealt with by all systematists that provide databases or analysis services. The establishment of meaningful collaborations among database owners, application developers and organizations is another issue that affects the users of systematics applications and databases. As software applications and resources for systematic biology mature, questions as to the reliability of data transactions will persist, simply because it is so difficult to keep track of the names of things. Looking at the Taxonomicon, we see that even the name 'Bacteria' can represent six different taxonomic ranks and has 13 synonyms, each supposedly a valid name for that group of organisms. Uncertainties as to the correct names of taxa are introduced by the complex relationships among names (which are often treated as unique identifiers) and taxa (which are dynamic and subject to constant revision), which can exist in one-to-one, one-to-many and many-to-many relationships. Those relationships are further confounded by the non-persistent nature of biological names which can introduce serious hidden errors of interpretation by end-users who may be unaware of these peculiarities.

The nomenclature is not persistent because our understanding of the relationships among the taxa is constantly changing; indeed, even our ideas about what constitutes a species is under constant scrutiny [50, 51]. Isaac *et al.* [51] attribute the burgeoning number of species to changes in our ideas about species, rather than to an increase in the number of newly described species. Nomenclature is frequently debated in cases where some species show heterogeneous association with various other species (such as those which has been seen with *Clostridia* species [52]) and this leads to the establishment of competing, parallel classifications. In order to clear the hurdles of accurately identifying, cataloguing, updating and retrieving information pertinent to species, a system of persistent identifiers needs to be established. There are at least two proposals under consideration: Life Science Identifiers (LSIDs) [53] and Digital Object Identifiers (DOIs) (http://www.doi.org/). LSIDs are used in Page's Taxonomic Search Engine (http://darwin.zoology.gla.ac.uk/~rpage/portal/) [54], while DOIs are being developed as unique identifiers of names, taxa, exemplars and associated data by the NamesforLife initiative [55]. The aim of both is to allow researchers to reach data and metadata relevant to an organism, no matter what label the organism may bear. With a proper transactional architecture, automated efforts at exploratory data analysis are achievable and could deploy, for example, insights that come from taxonomists in the field or biological data from specialist laboratories. A major problem facing plans for persistent identifiers is sociological; the information providers must be provided with a meaningful incentive to make accommodations and cooperate with the proposed initiatives. There are trends in informational technology that may assist general efforts in taxonomy. For instance, the semantic web seeks to increase interoperability and aid in service discovery and invocation [56]. Labelling and characterizing taxonomic information in a way that is cumulative, reproducible and computable for the practically significant meaning may be a significant milestone for biologists.

For non-sequence data such as nomenclatural references in the literature, as well as phenotypic,

ecological and metabolic data, the underlying data models are neither simple nor static. The Nomencurator [57] has a dynamic three-level model that flexibly tracks adjustments to various changes and updates to the nomenclature. The NamesforLife model of Garrity and Lyons [55] also uses a multi-level XML model, coupled with unique, persistent identifiers to provide a transparent metadata layer to disambiguate the constantly changing relationships among biological names, taxa and exemplars (the actual biological entities that are subject of the classification). This model also supports multiple taxonomic views and 'time travel' to provide end-users with a view of how taxonomy has changed over time (http://dx.doi.org/10.1601/tx.0). Taxonomicon presents an evolving, iteratively adjusted data model that is refined to fit meaningful categorizations of data corresponding to various species.

CONCLUSION

Methods for automatically identifying and/or classifying organisms exist or are in development. Methods for identification are much more mature than methods for classifying and are generally based on the measurement of some phenotypic property, whether it is a description of morphology or the assessment of metabolic capabilities. Even though identification methods based on molecular sequence analysis are now coming into wider use for prokaryotes and eukaryotic microorganisms, clinical laboratories continue to rely on phenotypic tests. The automation of identification based on phenotypic properties is fairly straightforward in the case of microorganisms. Automating the identification of eukaryotic macrobiota with this type of data is a much more complex proposition; if molecular methods can be brought to bear on these organisms, it promises to simplify identifications immensely. The foregoing comments apply equally well to the automated creation of classifications. Here, though, molecular methods are much better established, primarily due to the use of sequence data in phylogenetic analysis.

Another advantage of molecular approaches to identification and classification is the availability of data in a standard format. Although many electronic sources of information for a taxon of interest may exist, at least one type of data is, usually, conspicuously absent. Phenotypic data is not available in significant amounts. Although it has been very important in identifying and classifying prokaryotes in the past and continues to underlie most of the identification systems used in clinical laboratories, no significant, electronic sources of phenotypic data exist [58]. Interoperable databases of classification data for eukaryotes are also lacking.

In order to move ahead, the barriers between databases, and between databases and applications need to be reduced. One giant step towards such interoperability will be the institution of methods to tame the nomenclature issues so that biologists can ensure that the names they use are correct or, if not, that they can find the correct name along with the history of labels associated with the organism they are interested in. The automation of identification will also free researchers to apply their intellectual energy to the exploration of new areas in systematics and biodiversity. The discovery of new species and novel, deep-branching lineages equivalent to phyla [59] and the need to discriminate among organisms below the species level [60] are certain to be drivers of future developments in computational systematic biology. The ability of computational approaches to adapt to new discoveries, present clear depictions of alternative classifications and integrate disparate data types relevant to the classifications, will play a key role in the surveys of the natural world.

Key Points

- Interoperability: is the ability of systems to work together via data exchange and is facilitated by the use of software that recognizes the same protocols and can read from and write to shared file formats. Semantic interoperability ensures that the same items, concepts and so on are identified using the same labels across systems. Interoperability ensures that users can seamlessly query information from multiple repositories at the same time.
- Exemplar: a representation of an organism by reference to curated material and data.
- Taxon: a group of organisms defined by a name. The plural is taxa. Synonymy: The existence of more than one valid name for a taxon.
- Machine learning: an approach to programming in which a program is designed to modify some aspect of itself so that in repeated runs with the same input data its performance improves.

Acknowledgements

Portions of this work were funded and supported by the Office of Science (BER), US Department of Energy, Grants No. DE=FG02-04ER63933 and DE=FG02-04ER63933.

References

1. Simpson GG. *Principles of Animal Taxonomy*. New York: Columbia University Press, 1961.
2. Ereshefsky M. *The Poverty of the Linnaean Hierarchy*. Cambridge, UK: Cambridge University Press, 2001.
3. Felsenstein J. *Inferring Phylogenies*. Sunderland, MA: Sinauer Associates, 2004.
4. Alroy J. How many named species are valid? *Proc Natl Acad Sci USA* 2002;**99**:3706–11.
5. Blaxter ML. The promise of a DNA taxonomy *PhilosTrans R Soc Lond B Biol Sci* 2004;**359**:669–79.
6. Sneath PHA, Sokal RR. Numerical taxonomy *The Principles and Practice of Numerical Classification*. San Francisco, CA: W. H. Freeman, 1973.
7. Askevold IS, O'Brien CW. DELTA, an invaluable computer program for generation of taxonomic monographs *Ann Entomol Soc Am* 1994;**87**:1–16.
8. Gauld ID, O'Neill MA, Gaston KJ. Driving Miss Daisy: the performance of an automated insect identification system. In: Austin AD, Dowton M, (eds). *Hymenoptera: Evolution, Biodiversity and Biological Control*. Collingwood, VIC: CSIRO, 2000;303–12.
9. Watson AT, O'Neill MA, Kitching IJ. Automated identification of live moths (Macrolepidoptera) using Digital Automated Identification SYstem (DAISY) *Systematics and Biodiversity* 2003;**1**:287–300.
10. Arbuckle T, Schroder S, Steinhage V, Wittmann D. Biodiversity informatics in action: identification and monitoring of bee species using ABIS. In: Hilty LM, Gilgen PW, (eds). *15th International Symposium on Informatics for Environmental Protection*. Metropolis: Zurich, 2001; 425–30.
11. Arbuckle T. Automatic identification of bees' species from images of their wings. In *9th International Workshop on Systems, Signals and Image Processing*. UK: Manchester: UMIST, 2002.
12. O'Hara CM. Manual and automated instrumentation for identification of *Enterobacteriaceae* and other aerobic gram-negative bacilli *Clin Microbiol Rev* 2005;**18**:147–62.
13. Lapage SP, Bascomb S, Willcox WR, Curtis MA. Identification of bacteria by computer: general aspects and perspectives *J Gen Microbiol* 1973;**77**:273–90.
14. Bochner BR, Gadzinski P, Panomitros E. Phenotype microarrays for high-throughput phenotypic testing and assay of gene function *Genome Res* 2001;**11**:1246–55.
15. Bochner BR, Gomez V, Franco-Buff A. Modern Phenotypic Taxonomy Using Phenotype MicroArrays *In International Congress of Bacteriology and Applied Microbiology*. CA: San Francisco, 2005.
16. Hebert PDN, Cywinska A, Ball SL, deWaard JR. Biological identifications through DNA barcodes *R Soc Lond* 2003;**270**: 313–21.
17. Hebert PD, Gregory TR. The promise of DNA barcoding for taxonomy *Syst Biol* 2005;**54**:852–59.
18. Harmsen D, Rothganger J, Frosch M, Albert J. RIDOM: Ribosomal Differentiation of Medical Micro-organisms Database *Nucleic Acids Res* 2002;**30**:416–17.
19. Enright MC, Spratt BG. Multilocus sequence typing *Trends Microbiol* 1999;**7**:482–87.
20. Jorgensen HJ, Mork T, Caugant DA, *et al*. Genetic Variation among *Staphylococcus aureus* strains from norwegian bulk milk *Appl Environ Microbiol* 2005;**71**: 8352–61.
21. Aanensen DM, Spratt BG. The multilocus sequence typing network: mlst.net *Nucleic Acids Res* 2005;**33**: W728–33.
22. Scotland RW, Olmstead RG, Bennett JR. Phylogeny reconstruction: the role of morphology *Syst Biol* 2003;**52**: 539–48.
23. Jenner RA. Accepting partnership by submission? Morphological phylogenetics in a molecular millennium *Syst Biol* 2004;**53**:333–42.
24. Wiens J. The role of morphological data in phylogeny reconstruction *Syst Biol* 2004;**53**:653–61.
25. Swofford DL. *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods)*. 4th edn. MA: Sunderland, Sinauer Associates, 2003.
26. Felsenstein J. *PHYLIP (Phylogeny Inference Package)*. In 3.6th edn. Department of Genome Sciences, University of Washington, WA: Seattle, Distributed by the author 2005.
27. Maddison DR, Maddison WP. MacClade version 4 *Analysis of phylogeny and character evolution*. MA: Sunderland, Sinauer Associates, 2000.
28. Maddison WP, Maddison DR. Mesquite: a modular system for evolutionary analysis. In 1.06th edn. http://mesquiteproject.org 2005.
29. Rohlf FJ. NTSYSpc *Numerical Taxonomy System*. Setauket, NY: Exeter Publishing, 2005.
30. Podani J. SYN-TAX 2000. Budapest, Hungary: Scientia 2001.
31. Garrity GM, Lilburn TG. Self-organizing and self-correcting classifications of biological data *Bioinformatics* 2005;**21**:2309–314.
32. Lilburn TG, Garrity GM. Exploring prokaryotic taxonomy *Int J Syst Evol Microbiol* 2004;**54**:7–13.
33. Lutzoni F, Kauff F, Cox JC, *et al*. Assembling the fungal tree of life: Progress, classification, and evolution of subcellular traits *Am J Bot* 2004;**91**:1446–80.
34. Hibbett DS, Nilsson RH, Snyder M, *et al*. Automated phylogenetic taxonomy: an example in the *Homobasidiomycetes* (Mushroom-forming fungi) *Syst Biol* 2005;**54**:660–8.
35. de Queiroz K, Gauthier J. Phylogenetic Taxonomy *Ann Rev Ecol Syst* 1992;**23**:449–80.
36. Nilsson RH, Kristiansson E, Ryberg M, Larsson K-H. Approaching the taxonomic affiliation of unidentified sequences in public databases - an example from the mycorrhizal fungi *BMC* 2005;**6**:178.
37. Cole JR, Chai B, Farris RJ, *et al*. The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis *Nucleic Acids Res* 2005;**33**: D294–96.
38. Garrity GM, Bell J, Lilburn TG. Taxonomic Outline of the Prokaryotes *Bergey's Manual of Systematic Bacteriology*. 2nd edn, Release 5.1, 2004.
39. Delsuc F, Brinkmann H, Philippe H. Phylogenomics and the reconstruction of the tree of life *Nat Rev Genet* 2005;**6**: 361–75.

40. Snel B, Huynen MA, Dutilh BE. Genome trees and the nature of genome evolution *Annu Rev Microbiol* 2005; **59**:191–209.
41. Savva G, Dicks J, Roberts IN. Current approaches to whole genome phylogenetic analysis *Briefings in Bioinformatics* 2003; **4**:63–74.
42. Battistuzzi FU, Feijao A, Hedges SB. A genomic timescale of prokaryote evolution: insights into the origin of methanogenesis, phototrophy, and the colonization of land *BMC Evol Biol* 2004;**4**:44.
43. Henz SR, Huson DH, Auch AF, *et al*. Whole-genome prokaryotic phylogeny *Bioinformatics* 2005;**21**:2329–35.
44. Konstantinidis KT, Tiedje JM. Genomic insights that advance the species definition for prokaryotes *Proc Natl Acad Sci USA* 2005;**102**:2567–72.
45. Kunin V, Ahren D, Goldovsky L, *et al*. Measuring genome conservation across taxa: divided strains and United Kingdoms *Nucleic Acids Res* 2005;**33**:616–21.
46. Tekaia F, Yeramian E. Genome Trees from Conservation Profiles. PLoS *Comput Biol* 2005;**1**:e75.
47. Brands SJ. Systema Naturae 2000. Amsterdam: http://www.taxonomicon.net/ 1989–2005.
48. Wake MH, Prance G, Hoshi M, *et al*. GBIF 3rd Year Review. http://circa.gbif.net/Public/irc/gbif/pr/library?l=/review_documents/report_ph_pdf/ 2005.
49. Edwards JL, Lane MA, Nielsen ES. Interoperability of biodiversity databases: biodiversity information on every desktop *Science* 2000;**289**:2312–14.
50. Silvertown J, Servaes C, Biss P, Macleod D. Reinforcement of reproductive isolation between adjacent populations in the Park Grass Experiment *Heredity* 2005;**95**: 198–205.
51. Isaac NJB, Mallet J, Mace GM. Taxonomic inflation: its influence on macroecology and conservation *Trends Ecol Evol* 2004;**19**:464–69.
52. Collins MD, Lawson PA, Willems A, *et al*. The phylogeny of the genus *Clostridium*: proposal of five new genera and eleven new species combinations *Int J Syst Bacteriol* 1994;**44**: 812–26.
53. Clark T, Martin S, Liefeld T. Globally distributed object identification for biological knowledgebases *Briefings in Bioinformatics* 2004;**5**:59–71.
54. Page RD. A Taxonomic Search Engine: federating taxonomic databases using web services *BMC Bioinformatics* 2005;**6**:48.
55. Garrity GM, Lyons C. Future-proofing biological nomenclature *Omics* 2003;**7**:31–3.
56. Lord P, Bechhofer S, Wilkinson MD, *et al*. Applying semantic web services to Bioinformatics: Experiences gained, lessons learnt. In: McIlraith SA, Plexousakis D, van Harmelen F, (eds). *3rd International Semantic Web Conference*. Japan: Springer, 2004;841.
57. Ytow N, Morse DR, Roberts DM. Nomencurator: a nomenclatural history model to handle multiple taxonomic views *Biol J Linn Soc* 2001;**73**:81–98.
58. Field D, Garrity GM, Morrison N, *et al*. eGenomics: Cataloguing our Complete Genome Collection *Comparative and Functional Genomics* 2006;**6**:363–68.
59. Dawson SC, Pace NR. Novel kingdom-level eukaryotic diversity in anoxic environments *Proc Natl Acad Sci USA* 2002;**99**:8324–29.
60. Nikkari S, Lopez FA, Lepp PW, *et al*. Broad-range bacterial detection and the analysis of unexplained death and critical illness *Emerg Infect Dis* 2002;**8**:188–194.