Journal of Industrial Microbiology & Biotechnology (1997) 18, 68–72

Towards microbial data integration

N Larsen[3], R Overbeek[4], S Pramanik[2], TM Schmidt[1,3], EE Selkov[5], O Strunk[6], JM Tiedje[1,3] and JW Urbance[3]

[1]*Department of Microbiology;* [2]*Computer Science;* [3]*Center for Microbial Ecology, Michigan State University, East Lansing, MI 48824, USA;* [4]*Mathematics and Computer Science Division, Argonne National Laboratories, Argonne, IL 60439, USA;* [5]*Institute of Theoretical and Experimental Biophysics, Russian Academy of Sciences, 142192 Pushchino, Moscow region, Russia;* [6]*Lehrstuhl für Mikrobiologie, Technische Universität München, Arcisstraβe 16, D-80290 München, Germany*

There are currently 100–200 microbiology-related databases in existence, although it is impossible to find answers to queries that span even a few of these. The Center for Microbial Ecology (CME) at Michigan State University seeks to change this situation by coordinating the creation of an Integrated Microbial Database (IMD), accessible through the World Wide Web (WWW). Such a system will contain up-to-date phylogeny and taxonomy, gene sequences (including genomes), biochemical data, metabolic models, ecological and phenotypic data. Current main obstacles to creation of an IMD are the lack of a single freely available organismal nomenclature with synonyms and the availability of much critical data. An IMD will have major impacts on microbial biology: currently intractable fundamental questions might be answered, experiments could be refocused, and new commercial possibilities created. An IMD should remain freely available and be created under an open development model.

Keywords: database; integration; microbes; phylogeny

The present situation

Researchers today face a Herculean task when seeking answers to questions that cross the boundaries of existing databases on the characteristics of microorganisms. Currently there are 100–200 databases, each covering some aspect of microbial life: nomenclature, phylogeny and taxonomy, biochemistry, physiology, phenotypic information, ecology, genetics, alignment and sequence data (including the important recent availability of completely sequenced genomes). This growing mass of information is geographically scattered and organized in almost as many ways as there are databases. In the next few years broader databases will undoubtedly emerge, and hundreds may become dozens. Still, these fewer databases may well again become incompatible and not 'traversable'.

A decision to remedy the present situation

The Center for Microbial Ecology (CME) [6] at Michigan State University organized a workshop in August 1995 to discuss integration of microbial data. The workshop participants included representatives of international culture collections: the American Type Culture Collection (ATCC) [4], Deutsche Sammlung von Mikroorganismen und Zellkulturen GmbH (DSM) [11], the Japan Federation of Culture Collections, and the World Data Center for Microorganisms (WDCM) [15], CME [6], the Ribosomal Database Project (RDP) [2], Bergey's Manual Trust [14], Argonne National Laboratories, several prominent US and European database representatives, and representatives from the US National Science Foundation.

It was concluded, that the current major obstacles to data integration are mainly the lack of a single freely available up-to-date organismal nomenclature with synonyms and that a large amount of critical data is unavailable, either because it is organized in ways unsuitable for integration or because it is proprietary. There was also agreement that data availability is a far greater (and costly) problem to solve, than any software or database issue.

It was recommended that the CME should try to bring all microbial databases together in a loose federation, which would then coordinate creation of prototypes, seeking of funds, and further agreements on principle. Each member sub-database would be fully responsible for its data and retain freedom to maintain its data as it prefers, thereby restricting federation members as little as possible. For accessing the IMD, a World Wide Web-based system will be used. A copy of the workshop report, along with progress reports, is available online [7].

Proposed structure of the IMD and current progress

General

The conceptual components of the IMD and proposed linkages are illustrated in Figure 1. Organism nomenclature will serve as the main links between different types of data, thus its central placement. The integration will be organized and queriable with the 16S rRNA-derived phylogeny and/or taxonomy as default starting point. It is however important to offer queries using any data type as starting point. In Figure 1, the most developed databases are those that relate to molecular biology, whereas the database categories in most need of work are currently the phenotypic and ecological data.

Phylogenies and taxonomies

Microbial phylogeny, based upon 16S rRNA aligned sequences, will serve as a natural framework for organizing

Correspondence: Dr N Larsen, Department of Microbiology, Giltner Hall, Michigan State University, E Lansing, MI 48824-1101, USA
Received 20 June 1996; accepted 2 November 1996

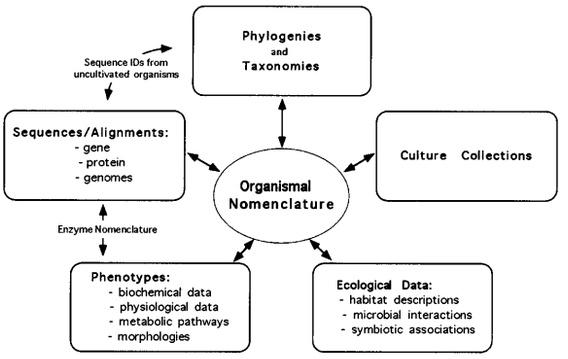

Figure 1 Proposed linkages of categories of data about microorganisms through organismal and enzyme nomenclature.

and accessing the data and displaying the results of many queries, but an IMD will also support taxonomies derived from other characters. The IMD effort will be completely dependent on the availability of a good up-to-date 16S rRNA-based phylogeny, such as the one now maintained by the Ribosomal Database Project (RDP) [2]. Currently there are many known organisms for which the 16S rRNA sequence has not been determined. Until the 16S rRNAs are known for all cultivated microbes, an IMD will have to use other criteria in order to provide comprehensive phylogeny-based queries. To accelerate this phylogenetic 'gap-filling', CME has initiated an effort (in coordination with DSM) to determine systematically the 16S rRNA sequence from all type strains for which such determination has not previously been done.

A consistent nomenclature

The workshop acknowledged the prime importance of a single freely available, up-to-date list of prokaryotic organism names, with a history of their synonyms, former names and orthographic variants. A sole source of organism names would probably be welcomed and adopted by all current databases, thus making it possible to connect these databases without losing many links. A large proportion of bacterial names has changed over the last few years in response to the recent 16S rRNA-derived classification, and when connecting multiple databases by names this becomes a serious problem. CME is currently coordinating the development and curation of a comprehensive and up-to-date prokaryotic organismal nomenclature via a WWW-dedicated server machine.

Phenotypic data

As another major element of the IMD, we propose to organize the gathering of phenotypic data from bacteria (including images) via the World Wide Web. The CME has initiated, with Bergey's Manual Trust, creation of such a WWW interface [8] by which geographically separated curators can build a database of phenotypic microbial characteristics. We are also exploring the feasibility of including existing phenotypic data.

Metabolic reconstruction

We believe that 100–200 phylogenetically diverse bacterial genomes will be completely sequenced (or near-complete) before year 2000. It will then be possible to assemble complete functional models for many common bacteria, and to characterize the groups they belong to in far more detail. By functional model we mean a hierarchical layout of the microbe's parts and its metabolism that describes functional units at both low and high levels, not unlike a blueprint of an engine, its parts and how they interact. The assignment of function to genes will be relevant not only to a given organism, but to groups that possess functionally equivalent genes. When linked together, different types of data can be compared, predictions evaluated, and key experiments devised. With enzyme properties included, these models will eventually become dynamic (like cranking the engine, to stay with the analogy). Work in this direction has already begun. Two of the authors of the present paper, RO and EES, are developing, in collaboration with CME, a WWW-based, metabolic reconstruction environment [9]. See also the PUMA system, a system for presenting functional overviews and metabolic reconstructions [13].

A simple computer model

We believe microbial data should be brought together in an open WWW-based system in which it is easy to participate. We envision a loose federation of databases, where each member curates its own data as preferred. Members would however commit to the following: 1) upload ASCII data (or some common format for non-ASCII data); 2) guarantee consistency of format and content of uploads; 3) describe the data fields and how they relate (either in concise English or in a formal notation). In addition, members will be urged to implement strong error control mechanisms and WWW annotation interfaces. We believe an integration will only be successful if contributors do not have to change their ways significantly, and if they are able to view and query their data in a much larger context. We propose a WWW-based database model, which does not provide for maintenance and curation, but which can accommodate any kind of data and software helper applications (viewers), and comes with computer source code. A single speed-optimized compilation would contain all data to be queried, disallowing changes and (at least initially) real-time additions to the data. Maintenance, on the other hand, must necessarily be curated by domain experts using hardware and software with which they are familiar. Each mirror site of an IMD would install local versions of all the data that the given site considers important to query, with access allowed for any client.

There are systems in various states of completion that could be built upon [3,12]. As our underlying query engine we chose the SRS project [10] developed by Thure Etzold and coworkers at European Molecular Biology Laboratory (EMBL). It is a fast engine already in production use with a simple query language, and over 120 molecular biology databases have already been defined under this system, ie added data can be queried in that whole context. Version 5 of SRS also fits the model above.

Since August 1995 CME has sponsored the development of a prototype that includes a Java-based phylogenetic interface, which provides convenient (and fast) viewing of phylogeny or taxonomy (Figure 2). This prototype can also

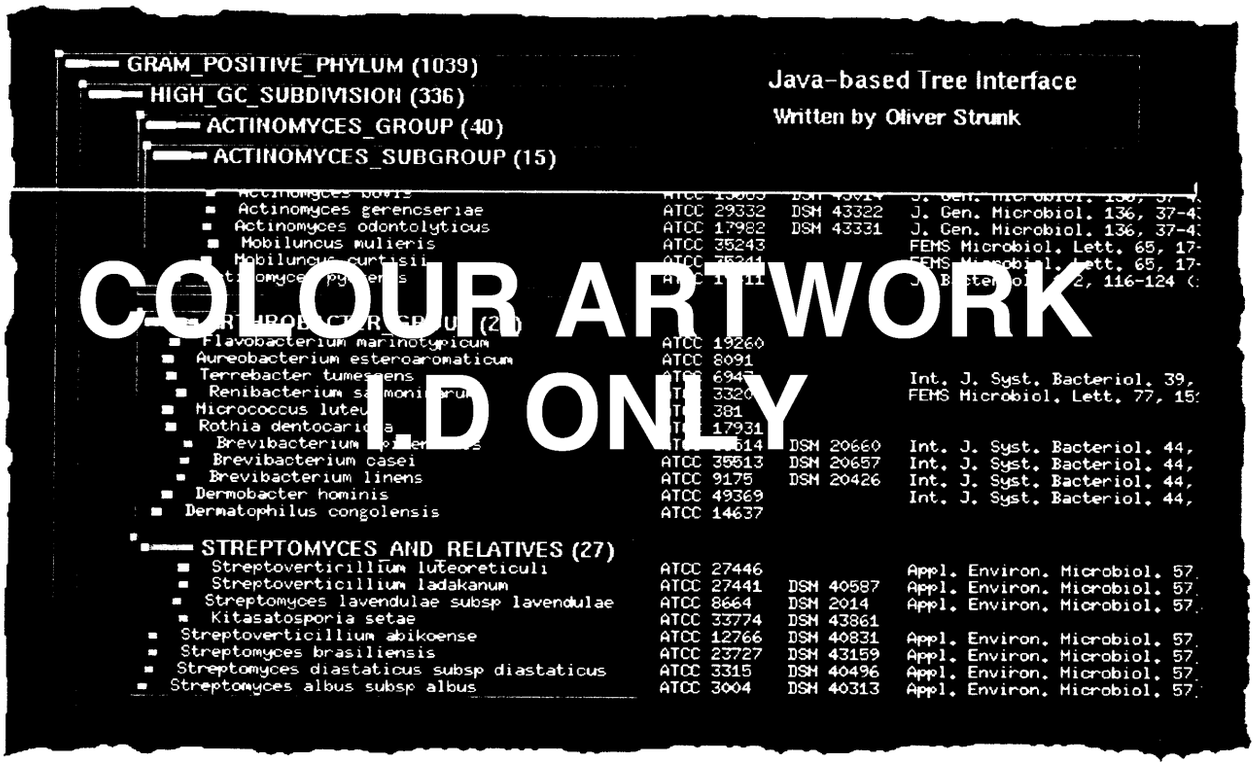

Figure 2 Prototype Web-based interface displaying the phylogenetic tree annotated with culture collection identifications and references. Strains located in BOTH collections are displayed on the tree as red nodes.

query a limited set of diverse data and display superimposed query results.

Potential benefits from an IMD

Currently considerable time is lost by both researchers and practitioners of microbiology searching for information on microorganisms and by failures to recognize concurrent features. Furthermore, new relationships or concepts are not possible to realize with the data in its current state of disarray. Practitioners such as those in the areas of pharmaceutical discovery, diagnosis, quality control, regulatory activities and patenting could save considerable time if they could access an electronic integrated microbial database.

An integrated microbial database could also help to fully realize the microbial world. Two major research directions are currently apparent for the field of microbial biology: genomic sequencing, which allows us to learn about a limited set of organisms in depth, and diversity discovery, which allows us to extend our knowledge to the many fascinating yet undiscovered organisms that make up our world. These two major branches of discovery can be linked for a more comprehensive understanding of the microbial world by an integrated microbial database, as illustrated in Figure 3. As we expand our knowledge in these two dimensions and by coupling it with an understanding of the interplay with environment, we can infer a much broader sphere of knowledge about the microbial world.

An integrated microbial database can allow us to address important and fundamental questions in microbial biology. Some of these questions include the following:

(1) Where are new species most likely to be found? We have at best only a very elementary understanding of the relationship between environment and both the extent of diversity and the microbial composition of particular environments. Better paradigms will allow us to predict where new diversity resides. An integrated database should contain both the organismal and environmental information that will allow such patterns to be revealed. At present the ecological data are the most deficient data component in microbiology and in need of serious attention if we are to better understand how diversity is patterned on earth.
(2) What are the patterns of phenotype with respect to phylogeny? For example, which lineages produce particular classes of chemical products; which lineages biodegrade particular classes of pollutant chemicals; or which lineages cycle elements important in sustaining biogeochemical cycles. Understanding these relationships can advance our understanding of these important processes, provide new tools to aid the discovery, as well as to enhance our understanding of gene exchange and evolutionary patterns in prokaryotes.
(3) Can the definition of a prokaryotic species be advanced? In principle genetic, phenotypic, and ecological information taken together should provide a species concept of similar biological meaning to that used for higher organisms. Polyphasic taxonomy is the accepted standard for prokaryotes. An integrated microbial database would provide the more extensive dataset needed for these goals, both in terms of numbers of examples as well as in the types of data. Fur-

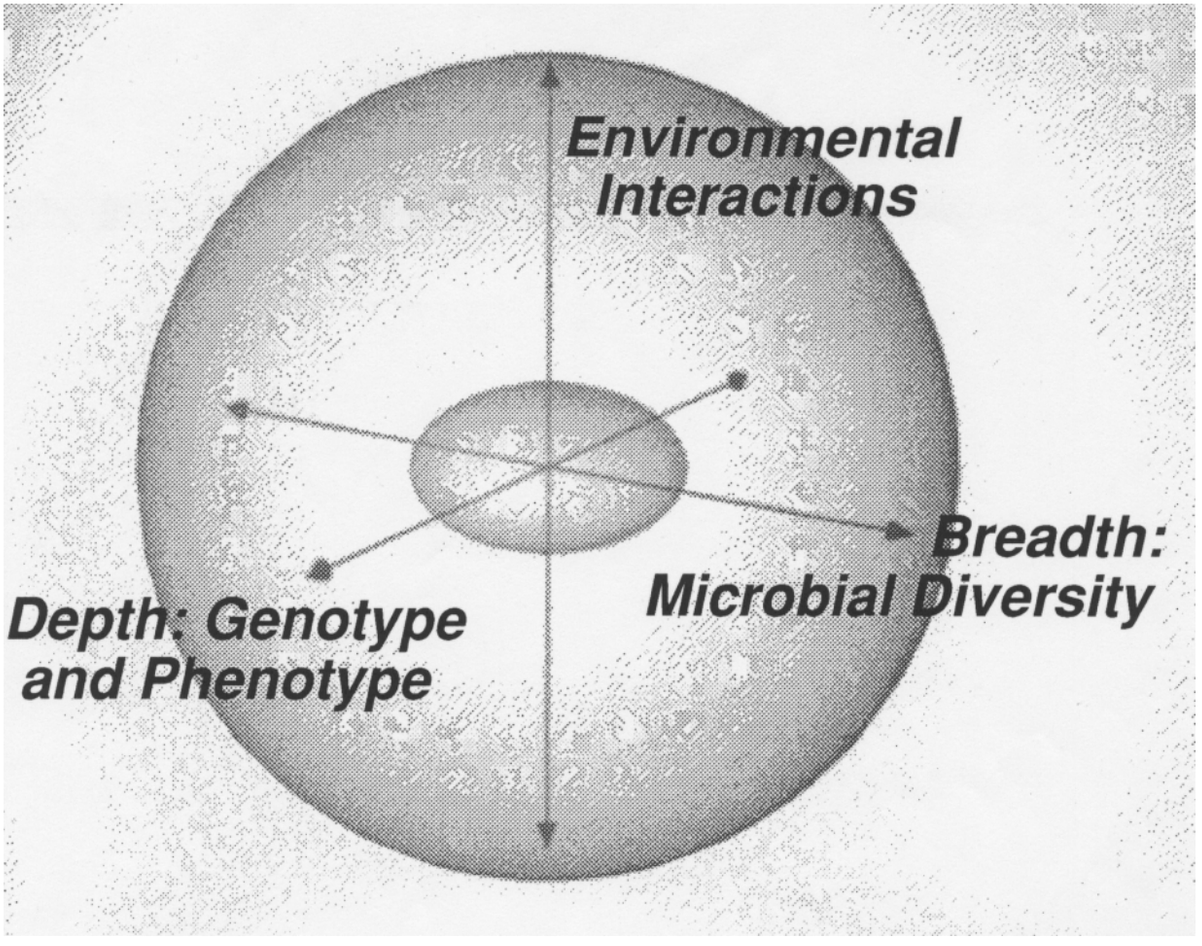

Figure 3 Microbial research proceeds in two major directions: diversity discovery (eg described strains, environmental rRNAs), which provides breadth, and detailed organism characterization (eg genome sequencing and metabolic reconstruction), which provides depth. An integrated microbial database will allow extension of knowledge from the current state by more easily inferring features of these and how they interplay with their environment—hence a sphere of knowledge (our current knowledge of the microbial world is represented by the inner, darker-shaded area).

thermore, these relationships could then be evaluated in a more quantitative manner.

(4) How unique are newly discovered organisms from what we already know, and which features characterize a given group of organisms? The ability to rapidly compare a new organism with existing knowledge using a robust query system and an integrated database is the most efficient manner to classify and assess the unique features of any set of organisms.

(5) What are the likely properties of organisms whose existence is suggested by the 16S rRNA sequence? Current rRNA sequence analysis of DNA extracted from nature suggests that most organisms are unknown and hence their ecological role and other properties remain obscure. An integrated microbial database, especially where phenotypic properties (including metabolism) can be linked to ribosomal sequence data, should allow one to infer properties of these organisms. For example, prediction of required substrates and end-products for an organism, or group of organisms, will be an outcome of metabolic reconstruction, making it easier to find cultivation conditions, construct bioreactors, etc. New services could be envisioned, such as: submit a partial rRNA sequence, and back comes not only its phylogenetic placement, but also what can be extrapolated from phenotypic data of relatives, from metabolic maps, or any data included in an Integrated Microbial Database System.

(6) What experiments should microbiologists do, and not do, when trying to characterize a microbe? If this microbe falls into a phylogenetic neighborhood where a given set of properties is clearly invariable, then very likely this set of properties is also present in the organism in question; at least a statistic can be made for the likelihood of its presence. Thus experimentation can be targeted towards key assumptions, and when the phylogenetic distribution of all functional units becomes known, the time saved for microbiologists should be substantial. Also, observations previously obtained by experimentation can be compared with results from interpretation and differences resolved.

(7) For a protein with certain desired properties, what oligonucleotides might be useful in amplifying the corresponding genes from nucleic acids isolated from the environment or identifying clones in an environmental clone library? Assuming a comprehensive enzyme

database is integrated, a query logic might proceed as follows: first retrieve the enzyme records with requested properties; follow links into the global alignment of the corresponding gene sequence; determine phylogenetic neighborhoods in which desired properties are clearly invariable (if any); for each of these neighborhoods, see if there are sub-sequences in the global gene alignment that uniquely characterize one or more of these neighborhoods. If present, perhaps a battery of such probes might capture the desired genes. Although this may not soon be a realistic approach, it illustrates the potential of data integration.

(8) Could the metabolic behavior of microbes be seriously addressed? Linkage of genomic sequences, phenotypic, metabolic and enzymatic databases such as EMP [5] will set the stage for the expansion of microbiology into meaningful full-scale simulations of the dynamic behavior of an organism, or group of organisms. This is our main motivation for creating functional overviews, and if this becomes possible, then it will clearly change microbial biology as we know it.

An integrated microbial database could be of high educational benefit with the addition of simplified interfaces, introductions and overviews. As part of CME's educational outreach, a 'Microbial Zoo' has been created [1], which links brief descriptions of commonly occurring microbes to cartoon sketches of familiar environments. By coordinating the developments, there could be a system which would be used by educators, students and scientists alike, offering desired levels of detail and functionality.

Availability and funding

An integrated microbial database should be freely available to all, including commercial entities. However, if sufficient funding cannot be raised, then a payment scheme may have to be devised, where clients pay for access. However, that idea would require registration, involve extra bureaucracy, and go against the idea of an open model. We believe such an information resource belongs to the international scientific community and to the public. Some critical data are currently proprietary and in some cases it may be cheaper to buy such data into the public domain than to fund their re-creation.

Acknowledgements

Research funding for the cited workshop and described development efforts was provided by the Center for Microbial Ecology, Michigan State University, under NSF grant No. BIR 9120006.

References

1 http://commtechlab.msu.edu/CTLProjects/dlc-me/zoo
2 http://rdpwww.life.uiuc.edu
3 http://www-genome.wi.mit.edu/informatics/abstracts/jamison.html
4 http://www.atcc.org
5 http://www.biobase.com/EMP
6 http://www.cme.msu.edu
7 http://www.cme.msu.edu/CME/PUBLICATIONS/publications.html
8 http://www.cme.msu.edu/PDE
9 http://www.cme.msu.edu/WIT
10 http://www.embl-heidelberg.de/srs/srsc
11 http://www.gbf-braunschweig.de/DSMZ/dsmzhome.htm
12 http://www.genome.ad.jp/manuscripts/GIW93/IL/GIW93102.html
13 http://www.mcs.anl.gov/home/compbio/PUMA/Production/pump.html
14 http://www.mph.msu.edu/Bergeys
15 http://www.wdcm.riken.go.jp