Supporting Taxonomic Names in Cell and Molecular Biology Databases.

Jessie Kennedy

School of Computing

Napier University

10 Colinton Road

Edinburgh, EH10 5DT

Scotland

Tel:+44 131 455 2772

Fax:+44 131 455 2727

j.kennedy@napier.ac.uk

Supporting Taxonomic Names in Cell and Molecular Biology Databases.

Jessie Kennedy

Napier University

17/1/2003

j.kennedy@napier.ac.uk

Abstract

Groups of organisms require labels or names to refer to them, however the idea of a single static name index, although tempting for its simplicity, is both impractical and unadvisable as a basis for referring to organisms for which data has been collected and stored for analyses and sharing. The relevant issues are described and some of the challenges facing database researchers are discussed.

Introduction

The field of biological taxonomy involves taxonomists classifying and naming groups of organisms, which provides others, e.g. cell and molecular biologists with a framework for identifying, categorizing and referring to organisms. However, the process of discovering, classifying and naming all organisms on Earth is far from complete, and the continuing accumulation of knowledge results in revisions of existing classifications with associated changes in taxon concepts and names. Although we need names or labels to refer to things, we cannot simply assume a single, common reference classification, which uniquely categorises and names all organisms. The same organism may have at times been classified according to different taxonomic opinions and subsequently have several alternative names. Without halting the advancement of our knowledge of existing biodiversity, it is difficult to see how we can (in the foreseeable future) achieve a single, static index of species names, which will serve to provide unique identifiers for all organisms. Therefore, we must acknowledge this issue and deal with it adequately in biological information resources, which reference groups of organisms or taxa.

Biological databases are a relatively new medium for the storage of biological information. However, the emphasis on the design and development of these databases has primarily been in recording the data generated from experiments, such as nucleotide sequences, proteins, metabolic pathways, gene expression etc. [1,2,3,4], on particular groups of organisms, rather than the seemingly trivial reference to the source organism. Biologists interact with these databases using labels to refer to the specimens or organisms, including common names, generic names and species names. These are the same taxa and names used in the taxonomic literature but without reference to the taxonomic concept associated with the label. Biological taxonomy can provide the framework by which biological information is stored, retrieved, and exchanged, but it is necessary for biological databases to accurately represent the taxonomic constructs, rather than simply use an undefined label. The major challenges in biology are to answer the "bigger" questions, which require integrating data from different experiments (and hence databases). Therefore to ensure valid conclusions are drawn from any analysis which integrates data from different sources, it is vital that like is compared with like, however this cannot be guaranteed with an un-attributed name.

Several challenges for database research arise from the need to allow users to reference organisms by name while accurately representing the reality of the meaning and usage of taxonomic names. To represent taxonomic concepts adequately, the minimum information required is the full taxonomic name and reference to the author and publication in which the concept was described [5,6]. Therefore if a biologist is naming an organism (identifying it) he must cite the publication used for identification purposes. This publication will be either a taxonomic work (and hence will define the concept) or that publication should cite a taxonomic work in order to fix the taxonomic concept to the associated name. This will allow others to be sure of the concept associated with the name, however it will not allow them to automatically compare the concept to other concepts, unless they are experts in the taxonomic group concerned. In order to interpret the relationships between taxonomic names (concepts), one must know, not only the classification assumed by the original publication, but also the nomenclatural and taxonomic changes that relate that classification to others. There are 2 general ways that this can be done. If a sufficient description of the taxon concept has been captured [7] then it could be possible to automatically determine the similarity of concepts. However for most historical classifications there is insufficient information recorded to enable this to be done and therefore although this would be the most useful approach for the long term, it would only be a solution for future taxonomic revisions. A second mechanism is for taxonomic experts to explicitly define the relationships between taxa [8,9,10,14]. This approach is limited in that few other relationships can ever be inferred and little automation can be provided, therefore the process of determining the relationships between taxa will always require to be done manually. Both approaches require work by expert taxonomists, however even if this work was completed, there is insufficient support in existing database systems to take advantage of it.

Database Research Challenges

In order to model the reality of taxonomy and nomenclature, database management systems must provide support to store and manipulate the structures and properties of this type of data [11,12,13]. Classifications are hierarchies, however, when all revisions of classifications of groups of organisms are considered we have in effect a graph of overlapping hierarchies. There are many database research challenges in supporting taxonomy but perhaps the major challenge is modelling and manipulating large, distributed hierarchies and graphs of complex objects. Graph structures are fundamental basic structures which can be used to describe many biological data types in addition to the increasingly pervasive requirement for ontologies.

Currently database systems provide limited facilities for modelling graphs, although there are many research database systems which provide some of the functionality required. However, none to our knowledge provide all of the functionality required [12], nor are they in widespread use or provide the support expected for multi-user environments with large-scale data requirements.

- Most graph databases (or support for graphs in other databases) treat nodes simply as labels. We require to be able to store objects (e.g. specimens) and use them optionally in one or more graphs (e.g. classification hierarchies, type hierarchies, placement hierarchies). Therefore the objects (specimens) must be independent of the graphs in which they occur and the graphs must be able to support complex objects as opposed to labels. Therefore, we need database modelling concepts to allow us to describe objects and relationships, from which we can then compose hierarchies and graph structures.

- Pattern matching is a common querying mechanism in graph databases, however patterns are typically simple paths in a graph. We require not only simple pattern matching but also patterns which allow matching of attributes of the nodes and edges in the paths of the graph.

- The levels in classification hierarchies are called ranks, however every classification does not make use of all possible ranks, although those that are used must appear in the given order. We need to be able to query by rank (level) in the graph where rank (level) is not semantically equal to depth, i.e. from a given node at a particular rank in one classification, a node at depth of 2 below in that classification will not necessarily have the same biological rank as a node at a depth of 2 below in another. Additionally to ensure the semantic integrity of the database we need to be able to specify constraints on the graph. e.g. nodes of a particular rank can only exist below other nodes in the hierarchy.

- Taxonomies are directed, therefore in queries we need to be able to traverse the graph or tree in a specified direction.

- The results of querying a graph could be a node or a sub-graph. If sub-graphs are returned the structure of the graph must be maintained, not simply the nodes.

- Having stored and being able to query our classifications we also need to be able to compare taxa or concepts. As discussed above this could be done in two ways, by capturing a definition of the concept in terms of for example its circumscription (members or child nodes of a given node) or by explicitly creating another edge between nodes that specifies explicitly the relationship between two taxa in different classifications (e.g. subset of) Both of these mechanisms have different graph query processing requirements.

We have built a prototype to support the functionality we require for taxonomic systems, but the system is currently not scalable for large systems. Nor has it been implemented on a platform with a sufficiently wide user base to encourage adoption of the approach. Providing this sort of functionality and support in commercial systems is also a major challenge in database research.

The development of taxonomy is a specialised field and the process is typically limited to small groups of organisms, therefore for pragmatic reasons there would need to be many autonomous taxonomic databases resolving parts of the overall taxonomic graph with an integrating database server providing a portal for all taxonomic names and synonym resolution. Any other biological database could then consult the taxonomy server for appropriate name and concept usage with possible synonymy or homonymy resolution with some indication of similarity or certainty of the relationship that could be used to guide the integration of data within and between databases. This does not mean that we require a taxonomic list server, forcing users to adopt a single view of the world, which might be possible in local regions or institutions but would certainly not be acceptable globally. Therefore, a taxonomy server supporting multiple views is essential to support the global sharing of data. Developing such a support mechanism is another major challenge.

References

1. STOESSER G., BAKER W., VAN DEN BROEK A., CAMON E., GARCIA-PASTOR M., KANZ C., *et al.* (2001). The EMBL Nucleotide Sequence Database. Nucleic Acids Res. 29:17-21

2. BARKER, W.C., GARAVELLI, J.S., HAFT, D.H., HUNT, L.T., MARZEC, C.R., ORCUTT, B.C. *et al*., (1998). The PIR-International Protein Sequence Database. Nucleic Acids Res. 27, 26, 27-32

3. OGATA, H., GOTO, S., SATO, K., FUJIBUCHI, W., BONO, H., AND KANEHISA, M. (1999). KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 27, 29-34

4. BRAZMA, A. *et al.* ArrayExpress — a public repository for microarray gene expression data at the EBI. Nucleic Acids Res. (in press)

5. YOON, N AND ROSE, J. (2001). An Information Model for the Representation of Multiple Biological Classifications. In: Alexandrov VN, Dongarra JJ, Juliano BA, Renner RS, Tan CJK, editors. Computational Science - ICCS 2001: International Conference, San Francisco, CA, USA, May 2001 Proceedings, Part 1. New York:Springer. 937-946.

6. YTOW, N., MORSE, D.R., ROBERTS, D. (2001). Nomencurator: a nomenclatural history model to handle multiple taxonomic views. Biological Journal of the Linnean Society, 73(1), 81-98.

7. PULLAN M.R, WATSON M.F, KENNEDY J.B, RAGUENAUD C, HYAM R. (2000). The Prometheus Taxonomic Model: a practical approach to representing multiple classifications. *Taxon* 49. 55-75.

8. VegBank. http://www.bio.unc.edu/faculty/peet/lab/PEL/vegbank/vegbranchhelp/VegBankInfo.htm

9. Moretax: http://www.bgbm.org/BioDivInf/Projects/MoreTax/standard_liste_en.htm

10. Universal Biological Indexer and Organizer: http://www.ubio.org/

11. RAGUENAUD, C., KENNEDY, J., BARCLAY P., (2000). The Prometheus Taxonomic Database. Bio-Informatics and Biomedical Engineering, Arlington, Virginia, USA, IEEE Computer Society Press. 63-70

12. RAGUENAUD, C. (2002). Managing Complex Data in an Object-Oriented Database, PhD Thesis, Napier University

13. RAGUENAUD, C., PULLAN, M., WATSON, M., KENNEDY, J., NEWMAN, M., BARCLAY, P. (2002). Implementation of the Prometheus Taxonomic Model: a comparison of database systems, Taxon, 51(1). 131-142

14. BEACH, J. H, PRAMANIK, S., BEAMAN, J. H. (1993). Hierarchic taxonomic databases. Ch. 15 in Fortuner, R., ed. Advances in Computer Methods for Systematic Biology: Artificial Intelligence, Databases, Computer Vision. Johns Hopkins Univ. Press, Baltimore. 241-256