

Modelling Biological Data in Hierarchies

ISMB 2002 Tutorial Proposal (Half Day)

Graham J.L. Kemp

Department of Computing Science, Chalmers University of Technology,
SE-412 96 Göteborg, Sweden
kemp@cs.chalmers.se
<http://www.cs.chalmers.se/~kemp/>

Peter M.D. Gray

Department of Computing Science, University of Aberdeen,
King's College, Aberdeen, AB24 3UE, Scotland, UK
pgray@csd.abdn.ac.uk
<http://www.csd.abdn.ac.uk/~pgray/>

Are you about to load up a rat's nest of data from different sources? Are you baffled about how to structure it? Do you have users with apparently conflicting demands on how they want to visualise the data? Do you foresee interoperability problems? Do you have AI frame-based tools that need to work with very different data in a relational database? If so, then this tutorial is for you! The presenters have been working with bioinformatic data for about 15 years, and are happy to share their experience, but also to discuss ideas and contributions from the floor.

Objectives

After this tutorial, those attending will:

- understand fundamentally different kinds of hierarchy, including those formed from "sub-type" and "subpart" relationships;
- understand different uses of hierarchies in biology;
- be aware of the use of hierarchies in controlled vocabularies, and their use in defining ontologies;
- be aware of the standard representations of hierarchies in computer systems, and what they are suitable for;
- be aware of how hierarchies can evolve;
- understand how different hierarchies can co-exist, and how they can be used in combination.

Audience

Anyone working with, or about to start working with, biological databases — particularly those involved in their design. No previous experience with database design is necessary. Those interested in Bio-Ontologies might be interested in a comparison of different ways of storing the same biological data.

Motivation

In biological data sets we see many different examples of hierarchies that need to be represented in bioinformatics data resources, e.g. classification hierarchies, is-part-of hierarchies, and so on. The characteristics of these hierarchies can be very different to one another, as are the ways in which the hierarchies are used. Further, very different hierarchical representations can be designed by people building different kinds of computer systems for use with the same data.

Therefore, trying to use the same representation formalism to represent different kinds of hierarchy can result in inappropriate systems being designed for particular applications, or can lead to confusion as system designers try to coerce two or more different kinds of hierarchy into one.

From 1991-1994 we worked on an EU project on “Integrated knowledge base for protein structure and sequence”. In that project we found that hierarchies constructed by our project partners in object-oriented programs were very different to hierarchies in a database schema designed for the same kind of data. Both hierarchies were useful, but they were very different to one-another. The reason that they were so different was that they had been designed for different purposes.

From 1996-1999 we were investigators on “KRAFT: Knowledge Reuse and Fusion/Transformation”. That project, which was unrelated to bioinformatics, involved groups whose expertise was with databases and with ontologies. Again, the hierarchies designed by these groups were very different to one-another. Again, they had been designed for different purposes.

People who are used to UML class diagrams tend to concentrate on hierarchies formed by association links on the diagram, while people who are used to AI frame-based representations tend to concentrate on classification hierarchies. Neither is sufficient by itself, and we need insights from data modelling on how to combine them sensibly. Certain patterns are beginning to emerge, and we shall illustrate their usefulness in structuring biological data.

This tutorial will look at examples of different hierarchies, and how they are used, so that attendees will be better able to use the right hierarchy for the appropriate purpose. Almost as important is an understanding of how different hierarchies can fit together and co-exist, so that a single collection of data can be viewed and understood through alternative hierarchies.

This tutorial will **not** argue that one kind of representation formalism is better than any other — all are useful and all have value in bioinformatics. Rather, by looking at examples of hierarchical biological data, and the different ways in which these can be represented in computer systems, this tutorial aims to raise awareness of the different motivating factors for constructing hierarchies in object databases, in object-oriented programs, in hierarchical vocabularies, and in frame-based systems — and the different hierarchies that can result.

Examples will be drawn from several areas, including *protein structure*, *protein function* and *microarray expression data*.

Detailed Outline

1. Hierarchies in Biological Data (15 minutes)
 - 1.1 Classification hierarchies
 - 1.2 Part-subpart hierarchies
 - 1.3 Taxonomies
 - 1.4 Phylogenies

2. Representing Hierarchies in Computer Systems
 - 2.1 Database Schemas (45 minutes)
 - entity-relationship modelling
 - subtype-supertype relationships
 - relational implementations
 - object database models
 - UML class diagrams
 - single vs. multiple inheritance
 - subsets vs. subtypes
 - schema evolution
 - data interoperability

- 2.2 Frame Systems (15 minutes)
 - “is-a” hierarchies
 - generic and specific planes
 - inheriting and overriding default values
 - the Protégé system
- 2.3 Hierarchical Vocabularies (15 minutes)
 - terms in ontologies
- 2.4 Object-oriented Programming (15 minutes)
 - inheriting behaviour
 - code re-use
 - design patterns
- 3. Different Hierarchies can and must Co-exist
 - 3.1 Databases and OOP —examples (30 minutes)
 - 3.2 Databases and Hierarchical Vocabularies —examples (30 minutes)
- 4. Discussion (20 minutes)

About the Presenters

Graham Kemp joined Chalmers University of Technology, Sweden, as an Associate Professor in January 2002. Previously he had been a Research Assistant, Research Fellow and Lecturer at the University of Aberdeen, Scotland. He has been involved in bioinformatics and database research since 1987. He is one of the developers of the P/FDM object database management system, and has strong interests in protein structure. He has a BSc Honours degree in Computing Science (1987), and a PhD on “Protein modelling using an object-oriented database” (1991), both from the University of Aberdeen.

Professor Peter Gray is well known in the International Database Research community. He has been at Aberdeen University since 1968, researching in AI Logic and KR techniques applied to Databases. In 1995 he was European PC chair for the international VLDB conference in Zurich. He is particularly interested in techniques of program generation and program transformation applied to databases. In this decade he has led a team that developed the P/FDM database that builds on the original Multibase project using an object-oriented development of Shipman’s Functional Data Model for integration of heterogeneous data sources. Recently he headed the KRAFT consortium of three universities and BT who were building an evolving distributed network in which knowledge can be shared and fused using an agent-based architecture with mediators.

Relevant Recent Presentation Experience

Graham Kemp presented a short tutorial on “Databases Architectures and Schemas” at the BBSRC/EPSRC Bioinformatics Grantholders’ Workshop (Wellcome Trust Genome Campus, Hinxton, January 2000). He was a speaker at the European Science Foundation Training Course in Functional Genomics: Curation of Databases in Molecular Biology, (CODATA Secretariat Building, Paris, October 2001), where he presented on “Models of Database Interactivity”. In May 2002 he will give a 3-day course on “Designing and maintaining biological databases” (provisional title) in the Swedish National Research School in Genomics and Bioinformatics.

Peter Gray has taught knowledge representation courses at M.Sc. level for many years. This covered AI KR languages, both Prolog and LISP based, as well as comparison with database schemas and E-R models. Many of these ideas have been used with protein structure databases and he has given a one week course on this in Switzerland at an open international summer school sponsored by Hoffmann la Roche. He is currently working on knowledge representations suitable for knowledge reuse and exchanging knowledge between agents, as part of the AKT (Advanced Knowledge Technologies) Collaboration supported by UK Research Councils over 6 years. Many of the issues discussed in the tutorial are also very relevant to the exchange of complex scientific knowledge being captured in the AKT project.