

Relational Databases for Biologists

[†]Aaron J. Mackey and [‡]William R. Pearson*

[†]Department of Microbiology,

[‡]Department of Biochemistry and Molecular Genetics,
University of Virginia, Charlottesville, VA 22908

January 4, 2002

Motivation

With the explosion of genome data and protein, gene, and functional database resources, biologists need to organize and interact with large datasets that combine information from diverse sources. Many public databases include cross-references between their internal data and related data in external databases. For example, sequences in GenBank contain cross-references to source organisms in the NCBI's Taxonomy database, the Gene Ontology Consortium publishes associations between its ontological terms and SwissProt proteins and yeast and *Drosophila* genes, and the PFAM domain database provides links to SwissProt entries with PFAM domains. However, it is not yet possible, through any public interface, to access an integrated version of these cross-referenced datasets. The need to manage genome-scale data has prompted some sophisticated researchers to build their own local databases, integrating only those datasets specific to their research needs. We believe "boutique" relational databases must become more ubiquitous in research labs and core support facilities.

Here we discuss the design and use of relational databases for this purpose, and introduce the entity-relationship model of data management that allows one to exploit the cross-referencing information of public databases for novel analyses. We build successively more complicated database models, discussing database schema design decisions and example usages of each structure. Advanced concepts for modelling hierarchical and time-dependent data will be discussed. Database models covering multiple domains of bioinformatic-related data are included. We will also review a few publicly available database frameworks for bioinformatics applications. This material will comprise a half-day tutorial.

*Corresponding author; Phone: (434) 924-2818; FAX: (434) 924-5069; email: wrp@virginia.edu

Intended Audience

The ISMB conference attracts biologists interested in informatics as well as informaticians interested in biology; a similar cross-section of participants should be interested in ways to investigate biological data in the context of publically available databases. We hope to convince biologists that setting up their own laboratory database does not require having an informatics department, nor investing significant resources in a database product. We also hope to educate informaticists in the relative advantages and disadvantages of various database schema alternatives. We hope to show both groups what new kinds of analyses can be accomplished by managing data within a relational context.

The Authors

Dr. William R. Pearson is the author of the **FASTA** package for sequence similarity searching and Professor of Biochemistry and Molecular Genetics at the U. of Virginia. His laboratory has begun using relational databases as parts of experimental protocols, rather than infrastructure, for computational biology analyses.

Aaron J. Mackey is a graduate student of Dr. Pearson's, and has significant experience designing databases for scientific and commercial use. His graduate research focuses on observing evolutionary patterns of mutation indicative of positive selection for change, both within single gene families and between entire bacterial genomes. He holds an MS in Immunology (1998) from Washington University in St. Louis.

Tutorial Outline

- I. Introduction: Why Put Data Into a Relational Database?
 - A. (Most) Data is Relational
 - B. Exploiting Relations
 - C. Relational Database Foundations
 1. The Entity-Relationship Model
 2. Structured Query Language
 - D. Alternatives: Flat Files, Object Oriented Databases, XML Databases
 - E. Research Labs vs. Enterprise Databases
- II. Database Products
 - A. MySQL
 - B. PostgreSQL
 - C. Commercial products – Oracle, DB2, SQLServer

III. Design and Use of a Sequence Database

- A. A Simple Sequence Database
- B. Getting Data Out of the Database
- C. Adding Cross-Referencing Data
- D. Putting Features on Sequences

IV. Beyond Simple Relationships

- A. Storing and Retrieving Hierarchical Data
 - 1. Data Models for Trees and Graphs
 - 2. Cross-Referencing the NCBI Taxonomy Database
 - 3. Integrating Sequence with Gene Ontology
- B. Capturing Changes to the Database – The Desire for Historical Integrity
 - 1. Time is not Relational
 - 2. Making it Work By Brute Force: Data Snapshots
 - 3. An Alternative: Audit Trails

V. Relational Database Solutions for Biologists

- A. **seqdb**: Proteins, Domains, Ontologies and Taxonomies
- B. **bioperl-db**: Sequences with Features
- C. **Ensembl**: Sequence Assemblies with Features

VI. Experimentally Generated Data

- A. Sequence Homology Search Results – **egads**
- B. Extending the Result Database to Include Other Algorithms
- C. Integrating Experimental Results with Public Data