Final Exam Review Sheet

CS329e Elements of Biological Data Models, Fall ’08
Professor Daniel Miranker

Exam Time: Friday, December 12, 7:00 – 9:00PM, in the regular lecture room, Painter 3.14,

This is a comprehensive exam. This review sheet is intended only as a study guide concerning the breadth of the exam. You are expected to know all the terminology presented as covered in class, the text and the required supplemental reading. Again, individual terms in this document are indicative of the breadth, they do not represent the actual content.

Reading:
• Most of the database text, except the sections on SQL
• Papers

Topics:

1. What is data modeling?
   a. Terminology
      i. Data model
      ii. Three schema model
         1. external schema
         2. conceptual schema
         3. internal schema
      iii. Entity, attribute, identifier, relation/association
      iv. Logical model, Physical Model, DDL
      ...  
   b. What are the steps of a data modeling effort
      i. planning and analysis
      ii. conceptual design // logic without the details
      iii. logical design
      iv. physical design
      v. implementation

2. Basic Relational Database Concepts
   a. Schema(s)
   b. Content addressability
   c. Keys
      i. Candidate key
      ii. Primary key
      iii. Foreign key
3. Data Model Concepts
   a. Data modeling process
      i. Plan project
      ii. Determine requirements
      iii. Specify entities
      iv. … (see course slide handout)
   b. Data provenance
   c. Entity, attribute, relation/association
   e. Relations/associations properties
      i. Cardinality constraints
      ii. Aggregation
      iii. Identifying relationships
      iv. Inheritance (subtypes)

3. Requirements analysis
   a. Why is it important?
   b. How does considering requirements change a database design compared to the simple use of databases as extensions of laboratory notebooks?
   c. What is the sequence of questions to be resolved in a requirements analysis?
      i. Who
      ii. What
      iii. Why
      iv. If they could do it now, how do they do it?

4. Sequence Alignment, Dynamic Programming Algorithms, (be prepared to fill in a dynamic programming matrix by hand) *** change from the midterm
   a. Edit distance (global alignment)
   b. Local alignment (Smith-Waterman)
   c. Backtrace
   d. Chaining

6. Rosetta Analysis
   a. Goals?
   b. How is it done?
   c. Gene fusion
   d. Use of GO in the evaluation of the method (per project milestone 4)

7. MIAME, Gene Expression
   a. What are the aims of MIAME?
   b. What are the six components of MIAME?
   c. Data provenance
   d. Controlled vocabulary

8. Ontologies
   a. What is an Ontology?
   b. What purposes do they serve?
   c. Why are they important?
   d. GO (Gene Ontology)
      i. What is the purpose of GO?
      ii. How is it organized?
      iii. How was GO implemented?

9. Database Indexing
   a. How does a B+ tree compare to a Binary tree?
   b. Why are secondary indexes important?
   c. What is the heuristic for when to include a secondary index?