COMPUTATIONAL BIOLOGY AND MEDICAL INFORMATICS

Instructors: Robert J. Robbins, Ph.D.
955–9705 (also 301–903–6488)
rrobbins@gdb.org

Harold P. Lehmann, M.D., Ph.D.
614–0843
lehmann@welchgate.welch.jhu.edu

Office Hours: by appointment

Readings: Although no formal text is required, handouts will be provided during the course and it is expected that you will become familiar with the materials in the handouts. Extra copies of the handout material will be available in the Computer science department office.

Additional suggested readings (i.e., optional readings) are given in the bibliography that follows.

SYLLABUS

31 Jan:	The Computational Challenge of Biomedical Information Management
2 Jan:	Data Modeling and Database Design
7 Feb:	Data Modeling in Action: Computer–based Patient Record
9 Feb:	Data Modeling in Action: Managing Genetic Information in Biology Research
14 Feb:	Data Abstraction: Computational Genomics
16 Feb:	Data Abstraction: Semantic Structure of Medical Knowledge
21 Feb:	Diagnostic Systems
23 Feb:	Wrap up
6 Mar:	**PAPER DUE** (no extensions; late papers will be penalized)

HANDOUT MATERIAL

Materials handed out in class (papers, copies of overhead transparencies, etc.) are considered to be integral parts of this course and should be read by all students.

Handouts for Computational Biology

Robbins, R.J. 1992. Challenges in the human genome project. IEEE Engineering in Medicine and Biology, 11:25–34.

Robbins, R.J. 1995. Representing genomic maps in a relational database. S. Suhai (ed), Computational Methods in Genome Research. New York: Plenum Publishing.

Robbins, R.J. 1994. Database Fundamentals. (handout material)

Robbins, R.J. 1994. DNA as a Mass–Storage Device

Robbins, R.J. 1994. Molecular Biology Fundamentals. (handout material)

Handouts for Medical Informatics

(specified elsewhere)

COURSE REQUIREMENTS (FINAL)

The final written exercise for the course is due by 5:00 pm, Monday, 6 March. **Two copies** of the written material being submitted for this final assignment should be turned in to the Computer Science office on or before this date. Make sure that the material is clearly labeled as the final project for this course, and make sure that you actually hand the material to someone in the office with the comment that it is for this class. Do not merely place the material on some desk or table in the office. Carelessly treated papers can get lost and it is your responsibility to ensure that submitted materials reach the instructors. Be sure to keep a copy of your paper for your records (and in case something should happen to the paper after you submit it). If you wish, you may ask the CS office staff for a receipt when you hand in your assignment.

The purpose of this exercise is for you to demonstrate your understanding of data–modelling concepts and skills in developing a modest data model from conception to finished model. An adequate project should require no more than 5–10 pages. If this seems insufficient, speak to us.

Select a subject, either from the suggestions below or of your own design (in consultation with an instructor), and use E–R methodology to develop a database schema to accommodate data for a particular problem area. Possible areas are genetic maps, scientific literature, annotated and aligned nucleotide sequences. The model should include both graphical representations and a narrative commentary that includes a description of the problem domain and that documents the "business rules" and integrity constraints captured in the system. In particular, your report should have the following components:

- **Problem Definition:** A summary of the domain problem and of the need to be met by the data model.
- **Requirements Analysis:** An analysis of the details of the problem. What functionality will be delivered by the model?
- **Documented Data Model:** graphical representation of the model, prose description of the system's functionality, "business rules," and integrity constraints
- **Discussion of the adequacy of your model:** What extensions/improvements would you make if you had more time? Were there any aspects of your problem domain that you could not represent adequately with the tools available to you?

POSSIBLE PROJECT TOPICS

Schema–Design Problems:

1. *Genetic Maps*: Devise a system with sufficient complexity to accommodate all of the map objects described in Robbins, 1993, Representing Genomic Maps in a Relational Database.
2. *Scientific Literature*: Devise a system that could accommodate a very large collection of scientific literature. At a minimum, the system should allow indexed retrieves by author, by title, or by key word, or via relationship links to other objects in the database. The system should handle many different types of literature, such as personal communications, journal articles, monographs, edited works, chapters in edited works, symposium proceedings, abstracts, etc. The system should allow new types of literature to be added (or deleted) at will, without requiring a rewrite of any software that retrieves literature. Ideally, the system would be capable of returning a bibliography, which is a sorted list of formatted citations, in correct bibliographic sort order, not necessarily ASCII sort order. Also, the bibliography may sometimes have to contain attributes not present in the individual citations. For example, if J. Smith wrote three papers in 1991, these dates for the three papers should be 1991a, 1991b, and 1991c, where a, b, and c represent the position of the three papers in the final sort order. Note that Smith might actually have written four or more papers in 1991, so the assignment of a, b, or c would have to be done after a bibliography had been assembled in response to some query.
3. *Laboratory Information Management System (LIMS)*: Devise a system that could accommodate data being generated in a large research or clinical laboratory. At a minimum, the system should (1) help track materials as they pass through the laboratory, (2) schedule experiments and other laboratory activities, (3) track the data as they are generated, linking the results to the source of the materials, (4) allow the automated operation of analytical software against the data, automatically logging the results in the database. For a clinical system, such a system should also (a) alert laboratory staff when abnormal conditions occur and (b) accumulate data over time, to allow for the determination of base–line expectations and other trend analyses.
4. *Annotated and aligned nucleotide sequences*: If you have worked previously with sequence data, you may define your own problem and devise a solution. If you have never worked with sequence data, this is not the best project to pursue.

Data modeling problems:

Clinical information systems: For any of the following, create a data model that would deliver the desired functionality. Consider, in your discussion, the impact of a computer network on the implementation of your model.

1. *Integrating ADT (admission/discharge/transfer) and Lab systems.* In this problem, the hospital uses the information it has about where a patient is in the system and for how long (1) to keep track of its census (which is used to evaluate how much resources are left available for other patients and how many staff members (e.g., nurses) are needed to provide care) and (2) to generate a bill. The laboratory uses its system (1) to keep track of

its resources, (2) to provide an audit trail for quality assurance (e.g., to the hospital administrators), (3) to provide clinicians with online access to laboratory results, and (4) to generate a bill to the patient. Your job is to create the data model that would maximize sharing of information across these two units. Limit this problem in any way that makes sense (e.g., reduce patient descriptors to his hospital registration number).

This problem is the core Hospital Information System problem as confronted by database managers in the mid–80's, where the problem was integration of legacy systems.

2. *Clinical information system*. In this problem, the hospital would like to keep track of patient–centered data (e.g., history, physical, and lab data) that occur both inpatient and outpatient. Limit this problem (1) to complaints of cough, (2) to keeping track of the chest exam (how do the lungs sound), and (3) to chest x–ray results. Consider in your discussion the question of data acquisition and access: who puts the information in and who is allowed to see it.

   This problem is the core Clinical Information System of the 80's.

3. *Integrated Information Systems*: The core CIS problem of the 90's is to integrate questions (1) and (2). Try it, limiting (1) and (2) in any nontrivial way.

Semantics: For any of the following, think about a small part of the semantic network of UMLS or of another formalism, and construct a data model that supports the semantic network.

1. For instance, consider a network consisting of a procedure branch (performing a bronchoscopy (passing a tube down a patient's windpipe looking for a pathological process)), a diagnostic branch (performing a chest x–ray), and a descriptive branch (lung anatomy). Include the interrelations between these branches, as well as issues like synonyms. (I will be happy to supply the domain information; you have to ask me the right questions as part of this exercise!)

2. For instance, consider program debugging, rather than a medical domain. Again, consider a network consisting of a procedure branch (e.g., altering the code), a diagnostic branch (e.g., putting in break points), and a descriptive branch (e.g., define a small set of prototypical types of bugs).

3. The metathesaurus problem. Consider the problem of creating allowing an arbitrary semantic system to connect with any other arbitrary semantic system.

Diagnostic systems: For any of the following, think about a small example of the formalism, and construct a data model that supports the framework.

1. From the description of QMR, construct a small diagnostic network, using examples from the article or made up (but keeping within the structure of the pram). Construct the corresponding data model.

2. Do the same for a belief–network—based program.

Miscellaneous:

1. Consider the notion of smart cards——portable, credit–card—sized electronic media that would store patient information. Construct a data model for the problem of communicating between two arbitrary hospitals. Focus on demographic information (age,

sex, race, address) and two pieces of clinical information (known allergies and a single chronic diagnosis). The goal here is to separate the communication from the HIS at any given hospital; you are therefore constructing a communication protocol.

2. Image database. Consider the data model for a system that allows users to search by the visual content of an image, e.g., "There's a yellow blob in the upper right–hand corner," or "Find me all images where the heart is on the wrong side of the chest." Limit the problem in any nontrivial way.

RECOMMENDED READINGS

These recommended readings are not required for the course, but are suggested as additional material for those interested in pursuing some course topics. They may also prove useful as background material for the final course assignment.

Recommended Readings: Computational Biology

Fields, C. 1992. Data exchange and inter–database communication in genome projects. *Trends in Biotechnology*, 10:58–61.

Frenkel, K.A. 1991. The human genome project and informatics. *Communications of the ACM*, 34:41–51.

Fuchs, R., Rice, P. and Cameron, G.N. 1992. Molecular biological databases—present and future. *Trends in Biotechnology*, 10:61–66.

Gilbert, W. 1991. Towards a paradigm shift in biology. *Nature*, 349:99.

Kingsbury, D.T. 1989. Computational biology for biotechnology: Part 1. *Trends in Biotechnology*, 7:82–87.

Lander, E.S., Langridge, R. and Saccocia, D.M. 1991. Mapping and interpreting biological information. *Communications of the ACM*, 34:33–39.

Mural, R.J., Einstein, R., Guan, X., Mann, R.C. and Uberbacher, E.C. 1992. An artificial intelligence approach to DNA sequence feature recognition. *Trends in Biotechnology*, 10:66–69.

Pearson, M.L. and Soll, D. 1991. The human genome project: A paradigm for information management in the life sciences. *The FASEB Journal*, 5:35–39.

Robinson, C. 1992. The genome race is on the road. *Trends in Biotechnology*, 10:1–5.

Recommended Readings: Biological Background

Colwell, R. R. (ed.) 1989. *Biomolecular Data: A Resource in Transition.* Oxford: Oxford Science Publications. 367 pp. ISBN: 0–19–854247–X

Gribskov, M., and Devereaux, J. (Eds.) 1991. *Sequence Analysis Primer.* New York: Stockton Press. 279 pp. ISBN: 0–7167–7002–4

Lesk, A. M. (Ed.) 1988. *Computational Molecular Biology: Sources and Methods for Sequence Analysis.* Oxford: Oxford University Press. 254 pp. ISBN: 0–19–854218–6

Ott, J. 1991. *Analysis of Human Genetic Linkage*. Baltimore: Johns Hopkins University Press. 302 pp. ISBN: 0–8018–4257–3.

Singer, M., and Berg, P. 1991. *Genes & Genomes: A Changing Perspective.* Mill Valley, California: University Science Books. 930 pp. ISBN: 0–935702–17–2.

Recommended Readings: Database Theory and Design

Batini, C., Ceri, S., and Navathe, S. B. *Conceptual Database Design: An Entity–Relationship Approach*. Redwood City, California: The Benjamin/Cummings Publishing Company, Inc. 470 pp. ISBN: 0–8053–0244–1

Codd, E. F. 1990. *The Relational Model for Database Management, Version 2*. Reading, Massachusetts: Addison–Wesley Publishing Company. 538 pp. ISBN: 0–201–14192–2

Date, C. J. 1983. *An Introduction to Database Systems, Volume II*. Reading, Massachusetts: Addison–Wesley Publishing Company. 854 pp. ISBN: 0–201–14474–3

Date, C. J. 1983. *Relational Database: Selected Writings*. Reading, Massachusetts: Addison–Wesley Publishing Company. 497 pp. ISBN: 0–201–14196–5

Date, C. J. 1990. *An Introduction to Database Systems, Volume I, 5th Edition.* Reading, Massachusetts: Addison–Wesley Publishing Company. 854 pp. ISBN: 0–201–51381–1

Date, C. J. 1990. *Relational Database: Writings 1985–1989.* Reading, Massachusetts: Addison–Wesley Publishing Company. 528 pp. ISBN: 0–201–50881–8

Elmasri, R., and Navathe, S. B. 1989. *Fundamentals of Database Systems*. Redwood City, California: The Benjamin/Cummings Publishing Company, Inc. 802 pp. ISBN: 0–8053–6681–4

Nijssen, G. M., and Halpin, T. A. 1989. *Conceptual Schema and Relational Database Design*. Englewood Cliffs, New Jersey: Prentice–Hall, Inc. 342 pp. ISBN:0–13–167263–0

Stonebraker, M. (Ed.) 1990. *Readings in Database Systems.* Palo Alto, California: Morgan–Kauffman Publishers, Inc. 644 pp. ISBN: 1–934613–65–6

Teorey, T. J. 1990. *Database Modeling and Design*. Palo Alto, California: Morgan–Kauffman Publishers, Inc. 267 pp. ISBN: 1–55860–134–1

Tsichritzis, D. C., and Lochovsky, F. H. 1982. *Data Models*. Englewood Cliffs, New Jersey: Prentice–Hall, Inc. 380 pp. ISBN:0–13–196428–3

Ullman, J. D. 1988. *Database and Knowledge–base Systems, Volume I.* Rockville, Maryland: Computer Science Press. 631 pp. ISBN: 0–88175–188–X

Ullman, J. D. 1989. *Database and Knowledge–base Systems, Volume II.* Rockville, Maryland: Computer Science Press. 506 pp. ISBN: 0–7167–8162–X

Recommended Readings: Systems Design

Brooks, F. P., Jr. 1982. *The Mythical Man–Month*. Reading, Massachusetts: Addison–Wesley Publishing Company. 195 pp. ISBN: 0–201–00650–2

Gause, D. C., and, Weinberg, G. M. 1989. *Exploring Requirements: Quality Before Design.* New York: Dorset House Publishing. 300 pp. ISBN: 0–932633–13–7

Gause, D. C., and, Weinberg, G. M. 1990. *Are Your Lights On?* New York: Dorset House Publishing. 157 pp. ISBN: 0–932633–16–1

Norman, D. A. 1988. *The Psychology of Everyday Things*. New York: Basic Books, Inc. 257 pp. ISBN: 0–465–06709–3

Recommended Readings: Medical Informatics

Clinical Medicine. New York: Springer–Verlag.

Blum B (ed) 1986. *Clinical Information Systems*. New York: Springer–Verlag.

Clancey, W. J. and Shortliffe, E. H. (Eds.) 1984. *Readings in Medical Artificial Intelligence: The First Decade*. Reading, MA: Addison–Wesley.

Dick, R.S. and E.B. Steen (Eds.) 1991. *The Computer–Based Patient Record*. National Academy Press: Washington, DC.

Greenes, R.A. and E.H. Shortliffe. 1990. Medical informatics: An emerging academic discipline and institutional priority. *Journal of the American Medical Association*, 263:1114–1120.

Miller, P.L. 1984. A critiquing approach to expert computer advice: ATTENDING. *Research Notes in Artificial Intelligence, Vol. 1*. Marshfield, MA: Pitman.

Pearl, J. 1988. Probabilistic reasoning in intelligent systems: Networks of plausible inference. In Ronald J. Brachman (ed), *The Morgan Kaufmann Series in Representation and Reasoning*, San Mateo, CA: Morgan Kaufmann.

Pechura, C.M. and J.B. Martin (Eds.) 1991. *Mapping The Brain and its Functions*. National Academy Press: Washington, D.C.

Reggia, J.A. and S. Tuhrim (Eds.) 1985. Computer–Assisted Medical Decision Making. B.I. Blum (ed), *Computers and Medicine, Vols. 1 & 2*. New York: Springer–Verlag.

Shortliffe, E.H. and L.E. Perreault. 1990. *Medical informatics: Computer applications in health care.* Reading, MA: Addison–Wesley.