Implementation Form – Graduate / Professional Certificates

This form must accompany a certificate proposal. It is used by administrative offices to better assist departments and programs with implementation.

Name of Graduate / Professional Certificate: Graduate Certificate in Bioinformatics
Faculty Program Director: Mark Craven
Primary Faculty/Staff Contact: Whitney Sweeney
Home Department/Academic Unit (Name/UDDS): Biostatistics and Medical Informatics/A531200
Approval Date: April, 2000
School/College: School of Medicine and Public Health
Approval Date: October 18, 2000
GFEC Approval Date: May 12, 2000
UAPC Approval Date: November, 2000
Implementation Term (typically the fall term after UAPC approval): Fall 2001

Year that first program review is scheduled (usually 5 years after implementation): We have not yet scheduled a program review

Plan Code (assigned by the Registrar’s Office): GCERT 106
Plan Descr (assigned by the Registrar’s Office): Bioinformatics
CIP Code (assigned by Academic Planning and Analysis): 261102 (Biostatistics)
Primary Divisional Disciplinary Assignment (assigned by APA for analysis purposes only): BIO

Curriculum:
____X____ Included in detail in the proposal
____X____ A list of required and elective courses is attached

Credit total required (should be between 9 and 12): 12 credits required
Credits required to be taken in residence at UW-Madison (must be at least 50%):
The assumption is that most, if not all credits will be taken while in residence at the UW-Madison, but at least 50% of them are required to be taken in residence.
Confirm that all core/required courses are approved through Divisional Committee: Yes
Confirm that courses in curriculum are offered on a regular basis and have space for students in this program: Yes

Projected enrollment: 5-10 students per year

What provisions have you made in the admissions process to gain consent from students’ degree/major program(s) to participate in the certificate program?
Whitney Sweeney is currently the coordinator of the Graduate Certificate in Bioinformatics. She will contact the degree/major program of each applicant to gain consent prior to formal acceptance into the Certificate Program.

Confirm that all courses numbered 300 or above Yes

Confirm that courses taken as pass/fail or audit are not allowed Yes

Confirm that special topics courses are only used if all instances count for the certificate: Yes

Will you use the typical minimum GPA requirement of 3.0 for all course work for the certificate? Yes

If no, specify other requirements:

Are courses taken Credit/No Credit allowed? No

If yes, specify limits:

Will exceptions to requirements be allowed? Yes

If yes, specify limits and process:

In very rare instances, exceptions to the requirements could be allowed. To do this, the instance would be brought to the attention of the Bioinformatics Capstone Committee and a unanimous decision would need to be made to grant the exception.

The department/program has a process in place to monitor student progress and to notify the Registrar’s Office when students complete the certificate requirements Yes

Program faculty and staff understand that a student’s graduation should not be delayed to complete the certificate. Yes

Specify overlap provisions – name degree/major or certificate programs that may not be earned along with the certificate. Note that majors take priority over certificates. (Students may not earn a graduate certificate if they are also earning a post-baccalaureate major/degree or PhD minor with the same name.)

Assessment plan – confirm that the proposal includes a plan that describes how the faculty will regularly evaluate student learning. Yes
Proposal for
A Graduate Certificate in Bioinformatics
Department of Biostatistics and Medical Informatics
May 1, 2000

1. INTRODUCTION

In the realm of biological and medical science, bioinformatics is becoming a central discipline. Bioinformatics, which is the application of computer science theory and methods to molecular biology, is placing a new demand on the training of graduate students in biology and in computer science. The educational objective of the proposed Graduate Certificate in Bioinformatics is to provide added formal training to pre-doctoral graduate students in molecular biology to improve their fundamental skills in bioinformatics. The goal is to allow them to have enough basic knowledge to continue their own research and to collaborate with computer scientists specializing in bioinformatics methods. As there is currently no active program which allows students at the UW-Madison to be trained in this area, this graduate certificate program in bioinformatics should not have an impact on any currently existing degree program.

2. NEED

The need for all researchers to have a basic competence in the area of bioinformatics has been eloquently expressed by the Working Group on Biomedical Computing, Advisory Committee to the Director, NIH in their June 3, 1999 report.

"In science and technology in the latter half of the 20th century, two fields have stood out for their speed of progress and their effect on society: biomedicine and computation. Increasingly, researchers spend less time in their "wet labs" gathering data and more time on computation. As a consequence more researchers find themselves working in teams to harness the new technologies. A broad segment of the biomedical research community perceives a shortfall of suitably educated people who are competent to support those teams. The problem is not just a shortage of computationally sophisticated associates, however. What is needed is a higher level of competence in mathematics and computer science among biologists themselves. To make optimal use of information technology, biomedical researchers need, first of all, the expertise to marry information technology to biology in a productive way."

This is the need that we are trying to address with this proposed Graduate Certificate Program in Bioinformatics.

3. PROPOSAL SUMMARY

The Graduate Certificate Degree in Bioinformatics consists of four courses for a total of 12 semester credits. We assume that, depending on the course and or research load of the student, they may complete the curricular requirements in 1-2 years at the most. This certificate program may also serve as a distributed minor.

At this point, we have identified a list of courses from which to choose that includes three required courses. One of these required courses consists of a choice between one of two Statistics courses, either of which will provide adequate background. The second course is specifically a bioinformatics course, and although it has been taught as a topics course in the past, it is being proposed as a regular course at this time. The third required course is a paired Biochemistry series which is currently taught in alternate years. These courses do not have to be taken in any particular sequence.
Basic Course Requirements:

- **Either Statistics 541 or 572.** Statistical concepts are basic to the methods and algorithms used in bioinformatics. Students are required to take one of two introductory statistics courses for graduate credits. Either course will give an adequate starting background.

**Statistics 541 - Introduction to Biostatistics** - is designed for the biomedical researcher. Topics include: descriptive statistics, hypothesis testing, estimation, confidence intervals, t-tests, chi-squared tests, analysis of variance, linear regression, correlation, nonparametric tests, survival analysis and odds ratio. Biomedical applications are discussed for each topic.

**Statistics 572 - Statistical Methods for Bioscience II** - is a design course aimed at CALS graduate students but the principles are quite applicable to molecular biology. Topics include: polynomial regression, multiple regression, two-way ANOVA with and without interaction, split-plot design, subsampling, analysis of covariance, elementary sampling and an introduction to the bioassay.

- **Computer Science 838 - Bioinformatics (Craven Course)**
Bioinformatics is an exciting new area that involves developing computational methods for managing and analyzing information about the sequence, structure and function of biological molecules and systems. The goals of this course are to provide an understanding of the fundamental computational problems in molecular biology and a core set of widely used algorithms in computational biology.
This is currently a Computer Science topics course and a proposal (see enclosed) is being submitted to make this into a regular course in the Department of Biostatistics and Medical Informatics.

- **Biochemistry 711/712 - Sequence Analysis (Palmenberg Course)** - This is a two-part course beginning with a lecture/discussion group course (711) and finishing with a hands-on laboratory course, taught at actual computer terminals, designed to complement and reinforce the sequence analysis concepts presented in Biochemistry 711. This course gives students a practical background in using many available software packages such as DNA-STAR for gene sequencing, etc. Students are provided with actual data and gain experience using this software.

Elective Courses:

In order to complete the 12 credits, students may select a fourth course from among the following. Additional elective courses are expected to be added as faculty are recruited. These courses are a beginning.

- **Statistics 542 - Fundamentals of Clinical Trials** - Intended for biomedical researchers interested in the design and analysis of clinical trials. Topics include definition of hypotheses, measures of effectiveness, sample size, randomization, data collection and monitoring, and issues in statistical analysis.

- **Computer Science 540 - Introduction to Artificial Intelligence**, teaches principles of knowledge-based search techniques; automatic deduction; knowledge representation using predicate logic, semantic networks, connectionist networks, frames, rules; applications in problem solving, expert systems, game playing, natural language understanding.
• **Computer Science 731** - Advanced Artificial Intelligence - Novel techniques within Bayesian Networks, Machine Learning and Data Mining, Planning and Computer Vision have proven useful for many real-world problems. This course will cover some of the most important recent algorithms from these areas and will illustrate their use with biomedical applications.

• **Computer Science 760** - Machine Learning. The intent of this course is to present a broad introduction to machine learning, including discussions of each of the major approaches currently being investigated. Class lectures will discuss general issues in machine learning, as well as present established algorithms. Computational approaches to learning, including: inductive inference, explanation-based learning, analogical learning, connectionism, and formal models, what it means to learn, algorithms for learning, comparison and evaluation of learning algorithms, cognitive modeling and relevant psychological results.

• **Computer Science 766** - Computer Vision - an introductory course to the basic concepts in computer vision including fundamentals of image analysis and computer vision, image acquisition and geometry, image enhancement, recovery of physical scene characteristics, shape-form techniques, segmentation and perceptual organization, representation and description of two-dimensional objects, shape analysis, texture analysis, goal-directed and model-based systems, parallel algorithms and special purpose architectures.

• **Medical Informatics Independent Study**
  Some students may find their needs are better met by an independent study with one of the faculty in the department, in collaboration with a biological faculty member.

• **Laboratory Experience 699**
  To be admitted, a graduate student must be working in the laboratory of a faculty member who is recognized as conducting research in molecular biology and has applied to be a part of this Bioinformatics Program.

4. **CERTIFICATE DEGREE STEERING COMMITTEE**

This program will be administered through the Department of Biostatistics and Medical Informatics and will be further developed and guided by a Steering Committee. Initially, this committee will consist of Jude Shavlik (Chair), David Page, Mark Craven, Ann Palmenberg, Michael Gould, David Schwartz, Fred Blattner, and David DeMets.

5. **ADMISSION REQUIREMENTS**

To begin, the Steering Committee will serve as the Admission Committee or may appoint a committee after the program gets underway.

To be eligible, students must be a graduate student in one of the biological or medical science graduate PhD programs in one of the laboratories of participating faculty.

6. **FEE STRUCTURE**
The fee structure for the courses offered through this graduate certificate program will be the same as it is for other graduate courses.

APPENDICES

1. Statistics 541
2. Statistics 572
3. Biochemistry 711/712
4. Bioinformatics 838
5. Statistics 542
6. Computer Science 540
7. Computer Science 731
8. Computer Science 760
9. Computer Science 766
Proposal for
A Capstone Certificate in Bioinformatics
Department of Biostatistics and Medical Informatics
June 1, 2000

1. INTRODUCTION

In the realm of biological and medical science, bioinformatics is becoming a central discipline. Bioinformatics, which is the application of computer science theory and methods to molecular biology, is placing a new demand on the training of post-doctoral fellows in biology. The educational objective of the proposed Capstone Certificate in Bioinformatics is to provide added formal training to postdoctoral fellows in molecular biology to improve their fundamental skills in bioinformatics. The goal is to allow them to have enough basic knowledge to continue their own research and to collaborate with computer scientists specializing in bioinformatics methods. As there is currently no active program which allows fellows at the UW-Madison to be trained in this area, this capstone certificate program in bioinformatics should not have an impact on any currently existing degree program.

2. NEED

The need for all researchers to have a basic competence in the area of bioinformatics has been eloquently expressed by the Working Group on Biomedical Computing, Advisory Committee to the Director, NIH in their June 3, 1999 report.

"In science and technology in the latter half of the 20th century, two fields have stood out for their speed of progress and their effect on society: biomedicine and computation. Increasingly, researchers spend less time in their "wet labs" gathering data and more time on computation. As a consequence more researchers find themselves working in teams to harness the new technologies. A broad segment of the biomedical research community perceives a shortfall of suitably educated people who are competent to support those teams. The problem is not just a shortage of computationally sophisticated associates, however. What is needed is a higher level of competence in mathematics and computer science among biologists themselves. To make optimal use of information technology, biomedical researchers need, first of all, the expertise to marry information technology to biology in a productive way."

This is the need that we are trying to address with this proposed Capstone Certificate Program in Bioinformatics.

2. PROPOSAL SUMMARY

The Capstone Certificate Degree in Bioinformatics consists of four courses for a total of 12 semester credits. We assume that, depending on the research load of the fellow, they may complete the curricular requirements in 1-2 years at the most.

At this point, we have identified a list of courses from which to choose that includes three required courses. One of these required courses consists of a choice between one of two Statistics courses, either of which will provide adequate background. The second course is a specifically bioinformatics course, and although it has been taught as a topics course in the past, it is being proposed as a regular course at this time. The third required course is a paired Biochemistry series which is currently taught in alternate years. These courses do not have to be taken in any particular sequence.

Basic Course Requirements:
APPENDICES

1. Statistics 541
2. Statistics 572
3. Biochemistry 711/712
4. Bioinformatics 838
5. Statistics 542
6. Computer Science 540
7. Computer Science 731
8. Computer Science 760
9. Computer Science 766
NEW COURSE PROPOSAL

DATE PREPARED: ________________

DEPARTMENT: Biostatistics and Medical Informatics

DIVISIONAL COMMITTEE: Physical Sciences

THIS COURSE PROPOSAL HAS BEEN APPROVED BY DEPARTMENT CURRICULUM COMMITTEE: XX YES ____ NO

1. COURSE NUMBER: 210-776

2. COURSE TITLE (limit to 68 spaces): Bioinformatics

3. CROSSLISTING DEPARTMENTS (attach supporting letters):
   (a) Computer Sciences
      (b) (c)
      (d)

4. IS THIS A "TOPICS" COURSE? ____ YES XX NO

5. PLANNED OFFERING: ____ Sem I, and/or XX Sem II, and/or ____ SS.

HOW OFTEN WILL COURSE BE OFFERED? Once a year

6. CREDITS: (a) Number: 3 credits. (Credits should be explained by number of scheduled contact hours for instruction, discussion, laboratory.)
   (b) Variable credit explanation: ____________________________________________

   (c) Can students take this course more than once for credit? ____ Yes XX No

   (d) Grading system: XX A-F or ____ Cr/No Credit

7. CAPSULE STATEMENT OF COURSE CONTENT FOR CATALOGS (maximum 40 words):
   Algorithms for computational problems in molecular biology. Topics covered include: dynamic programming for sequence and structure alignment, hidden Markov models for sequence modeling, clustering and classification methods for gene-expression data, phylogenetic tree construction, algorithms for predicting protein and RNA folding.

8. PREREQUISITES:
   One of the following courses: Computer Science 536, 538, 539, 540, or 577.

Will this course be open to freshmen? ____ YES XX NO

9. WHO WILL TEACH THE COURSE? (If nonfaculty, attach vita)
   Professor Mark Craven (Biostatistics and Medical Informatics), and alternatively
   Professor C. David Page (Biostatistics and Medical Informatics,
   Professor Jude Shavlik (Computer Science)

10. LEVEL OF COURSE: ____ ELEMENTARY ____ INTERMEDIATE XX ADVANCED
11. SHOULD COURSE BE REVIEWED FOR 100-CREDIT RULE? ____ YES ____ NO XX N/A
(For courses NOT in Letters and Science)

12. SHOULD COURSE SATISFY L&S LITERATURE REQUIREMENT? ____ YES XX NO ____ N/A
(For courses in Humanities only)

13. (a) SHOULD COURSE BE REVIEWED FOR L&S BREADTH REQUIREMENT? ____ YES XX NO
IF YES, INDICATE WHICH: ____ S ____ H ____ B ____ P ____ Z ____ N ____ OTHER
(b) SHOULD COURSE BE REVIEWED FOR ETHNIC STUDIES REQUIREMENT? ____ YES XX NO
(c) SHOULD COURSE BE REVIEWED FOR THE GENERAL EDUCATION REQUIREMENT? ____ YES XX NO
If so, indicate which and submit that request directly to Letters & Science.
COMMUNICATION ____ COM-A ____ COM-B QUANTITATIVE REASONING ____ Q R-A ____ Q R-B

14. DESCRIBE THE COURSE CONTENT (expand on capsule statement in #7):
The science of molecular biology is undergoing a revolution in how it is practiced. In the last decade, a vast amount of data (DNA sequences, protein sequences, etc.) has become available, and computational methods are playing a fundamental role in transforming this data into scientific understanding. Computational methods for analyzing genomic data are also playing an increasingly important role in medicine and drug design. Bioinformatics (computational molecular biology) is an exciting new area that involves developing computational methods for managing and analyzing information about the sequence, structure and function of biological molecules and systems.
The goals of the proposed course are to provide an understanding of the fundamental computational problems in molecular biology and a core set of widely used algorithms in computational molecular biology. The computational problems to be covered will include: pair-wise sequence alignment, multiple sequence alignment and sequence family modeling, finding genes in DNA, clustering and classifying gene expression data, phylogenetic tree reconstruction, protein structure prediction, RNA structure prediction and information access issues. The computational models and algorithms covered will include: dynamic programming, heuristic search, Markov chain models, interpolated Markov models, hidden Markov models, EM algorithms, clustering algorithms, supervised learning methods, and stochastic context free grammars.

15. EXPLAIN THE NEED FOR THIS COURSE (in particular, explain how this course contributes to strengthening your curriculum):
This course will serve a valuable role for two constituencies. First, it will educate graduate Computer Science students in the computational problems and algorithms that are central to genomics and molecular biology. There is currently great demand for computer scientists trained in this area. The course is valuable even for those Computer Science students who will not pursue research or a career in bioinformatics, however. Many of the algorithms covered in the course, such as hidden Markov models, are important computational models with applications outside of molecular biology, and they are not taught in other Computer Science courses.
The second constituency for the course consists of graduate students in the biological sciences who need to know about bioinformatics algorithms for their research or to pursue careers in industry. As with computer scientists, there is great demand for biologists who have training in this interdisciplinary area.

16. RELATIONSHIP TO OTHER UW-MADISON COURSES/POSSIBLE OVERLAP WITH COURSES IN YOUR OWN OR OTHER DEPARTMENTS (attach correspondence from appropriate departments addressing question of overlap):
The most similar course at the UW-Madison is Biochemistry 711, and the associated lab course Biochemistry 712. The proposed course and Biochemistry 711/712 differ, however, in their intended constituencies, emphasis, and topics covered. Biochemistry 711/712 is aimed primarily at students in the biological sciences who do not necessarily have any formal background in computer science. The proposed course, on the other hand, is aimed primarily at students who have a strong background in computer science but do not necessarily have any formal background in biology. Thus, in the proposed course there is more emphasis on computer science and statistical concepts such as computational
complexity, search spaces, bias/variance trade-offs, etc. Additionally, there are some topics covered in the proposed course that have not been covered in Biochemistry 711/712 in the past. These include hidden Markov models, analysis of gene expression data, stochastic context free grammars, etc. Although the proposed course builds on concepts covered in other Computer Science courses (e.g., dynamic programming, machine learning, regular expressions and context free grammars), the material does not substantially overlap with any other Computer Science course.

17. WILL ANY COURSES BE DELETED AS A RESULT OF THIS PROPOSAL? (Complete separate course deletion form for each course to be deleted.)
No

18. PLEASE ATTACH A COURSE SYLLABUS AND READING LIST. The syllabus MUST indicate how students will be evaluated (assignments, term papers, exams).
See attached syllabus and reading list from the course when it was taught as a “special topics” course in the Fall 1999 semester through the Computer Sciences Department.

19. WILL THIS COURSE MEET A REQUIREMENT FOR THE MAJOR IN YOUR DEPARTMENT OR ANOTHER DEPARTMENT?
XX YES NO
If yes, please specify
The Department of Biostatistics and Medical Informatics does not currently offer a major (only an emphasis within the Statistics Department). This course will satisfy a requirement of the Graduate Certificate in Bioinformatics and the Capstone Certificate in Bioinformatics, both programs currently in the approval process.
This course should satisfy in part the “Computer Science Electives” requirement for Computer Sciences majors.

20. PLEASE ATTACH A COVER LETTER AND OTHER RELEVANT MATERIAL.
CS 838 - Bioinformatics (Fall 1999)

General Course Information

Instructor: Mark Craven  
  craven@biostat.wisc.edu  
Office: 5730 Medical Sciences Center (corner of Charter and University)  
Office Hours: 3-4:00pm Tuesday & Thursday, or by appointment

TA: Beverly Seavey  
  seavey@cs.wisc.edu  
Office: 3660 Computer Sciences  
Office Hours: 1-2:00pm Wednesday, 11:30-12:30pm Thursday

Prerequisite: CS 367 or equivalent

Meeting Time and Location: 1:00-2:15 TR, 3345 Engineering

Course Overview

The science of molecular biology is undergoing a revolution in how it is practiced. In the last decade, a vast amount of data (DNA sequences, protein sequences, etc.) has become available, and computational methods are playing a fundamental role in transforming this data into scientific understanding. Bioinformatics (computational biology) is an exciting new area that involves developing computational methods for managing and analyzing information about the sequence, structure and function of biological molecules and systems. The goals of this course are to provide an understanding of:

- the fundamental computational problems in molecular biology
- a core set of widely used algorithms in computational biology.

Students will also gain familiarity with the most important databases and servers used in the field (e.g. GenBank, SwissProt, MEDLINE,...).

Course Requirements

The grading for the course will be based on:

  - homework assignments: ~30%
  - midterm exam: ~30%
  - project: ~35%
  - class participation: ~5%
The homework assignments will involve a mix of written problems and programming.

Readings and Lecture Notes

Introduction to Molecular Biology and Bioinformatics
- required reading
  Primer on Molecular Genetics
  selected pages from Los Alamos Science issue on The Human Genome Project
- optional reading
  Sequencers Endorse Plan for a Draft in 1 Year (5/99)
  A New Five-Year Plan for the U.S. Human Genome Program (1993)
- lecture notes
  General Course Information (9/2)
  Introduction to Molecular Biology (9/2, 9/7)
  Introduction to Bioinformatics (9/7, 9/9)
- resources
  Genbank
  SWISS-PROT
  Protein Data Bank
  MEDLINE

Pairwise Sequence Alignment
- required reading
- lecture notes
  Pairwise Alignment (9/9, 9/14, 9/16)
  Substitution Matrices and Database Searching (9/16, 9/21)
- resources
  NCBI's BLAST server

Multiple Sequence Alignment, Hidden Markov Models and Sequence Motifs
- required reading


- lecture notes
  - Multiple Sequence Alignment (9/23)
  - Hidden Markov Models, Part 2 (9/30, 10/5)
  - Learning Sequence Motifs with EM (10/7)
  - More on Motifs(10/12)

- resources
  - PROSITE database
  - Blocks database
  - MEME server
  - Multiple Alignment Servers
  - SAM HMM code
  - HMMER HMM code
  - PFAM database of multiple alignments

Finding Genes in Genomic DNA

- required reading

- lecture notes
  - Finding Genes in Genomic DNA (10/12)
  - Finding Genes in Genomic DNA: The GRAIL System (10/14, 10/19, 10/21)

- resources
  - GRAIL server and the Computational Biosciences Section at ORNL
  - extensive bibliography on computational gene recognition

Gene Expression Data and Molecular Medicine

- required reading
  - T. Golub et al. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring [PDF] [HTML].
- optional reading
  N. Friedman et al. Using Bayesian Networks to Analyze Expression Data. Submitted to RECOMB 2000.
  NY Times article on pharmacogenetics
- lecture notes
  Clustering and Classifying Gene Expression Data, Part 1 (10/26, 10/28)
  Clustering and Classifying Gene Expression Data, Part 2 (11/4)
  Inferring Regulatory Networks from Gene Expression Data (11/8, 11/10)
  Knowledge-based Avoidance of Drug-Resistant HIV Mutants (11/10)
- resources
  Affymetrix
  Molecular Classification of Cancer data set from Lander’s group
  Yeast Cluster Analysis data set from Stanford
  Thorsten Joachims’ SVM code
Phylogenetic Trees
- required reading
- lecture notes
  Constructing Phylogenetic Trees (11/16, 11/23)
- resources
Protein Structure Prediction
- required reading
- lecture notes
  Protein Structure Prediction, Part 1 (11/23, 11/30)
- resources
  secondary structure data set from UC Irvine Machine Learning Repository
  Information Extraction from Biomedical Text Sources
  required reading
APPENDICES
FALL SEMESTER
Statistics 541
Introduction to Biostatistics

COURSE WILL COVER:

- Design of medical studies
- Parametric and non-parametric procedures for paired and two-sample data
- Counting and categorical data analysis methods
- Linear regression and analysis of variance
- Relative risk, odds ratio, logistic regression, survival analysis

This course is primarily intended for graduate students or post-doctorates in basic sciences and medical sciences, as well as interested faculty. It will emphasize study design and the choice of appropriate statistical procedures as well as the interpretation of medical research results.

SCHEDULE: Tuesday and Thursday, 4:00-5:15pm

LOCATION: K6/115 Clinical Science Center

INITIAL SESSION: 2 September 1999

REGISTRATION: Usual registration procedures apply.

TEXT: *Biostatistics: A Foundation for Analysis in the Health Sciences*, by Wayne W. Daniel, published by Wiley and Sons

INSTRUCTOR: Jeffrey A. Douglas, PhD, Assistant Professor of Biostatistics and Medical Informatics, 263-2917, K6/418 CSC, douglas@biostat.wisc.edu

TEACHING ASSISTANT: Haonan Tan, haonan@stat.wisc.edu, 262-7173, rm 4279

Computer Science and Statistics

CLASSES HELD: Sep 2,7,9,14,16,21,23,28,30
Oct 5,12,14,21,26,28
Nov 2,4,9,11,16,23,30
Dec 2,7,9,14
EXAMS: Oct 19 and Nov 18 at the regular time and place

FINAL EXAMS Dec 23 at 2:45 pm

MATERIAL COVERED ch1, ch2, ch3, ch4, ch5, ch6, ch7, 8.1, 8.2, ch9, 11.4, 12.4, 12.5, 12.6, 12.7, 12.8 13.4, 13.5, 13.6

HOMEWORK ASSIGNMENTS HW1 ch1: 3, 6 ch2: 2.5.2 (a-f), 2.5.6 (a-f)
HW2 ch3: 3.4.4, 3.4.6, 3.5.2 ch4: 4.3.4, 4.3.8, 4.4.2, 4.6.2, 4.6.6, 4.6.12, 4.7.4
HW3 ch5 5.3.4, 5.4.2, 5.5.6, 5.6.2 ch6 6.2.2, 6.3.4, 6.4.2, 6.5.4, 6.6.4, 6.7.2, 6.10.4
HW4 ch7 7.2.8, 7.2.12, 7.3.10, 7.4.2, 7.5.2, 7.6.4, 7.8.4
HW5 ch8 8.2.4 (do not do dot plots or box plots) ch9 9.3.2, 9.3.6, 9.4.4, 9.5.4, 9.7.4 ch11 review questions 7, 9, 10, 11, 12
HW6 ch12 12.3.2, 12.4.2, 12.5.2, 12.6.2, 12.7.2, 12.8.4 (I'll copy article) ch13 13.5.2, 13.6.2

GRADING HW 25%, EXAMS 40%, FINAL 35%
Statistics--Forestry--Horticulture 572
Statistical Methods for Bioscience II
Spring 2000

Instructor: Murray Clayton

Phone: 262-6459, 262-0530, or 262-1009 (secretary)

E-mail: clayton@stat.wisc.edu

Office Hours: In 4375 CSSC 11:00-12:00 Mondays,
1:30 - 3:00 Wednesdays, or by appointment

Texts: You will need a copy of Handouts for Stat/For/Hort 572 available at Bob's Copy Shop. (Please
bring this to class throughout the semester.)

The following texts are not required, but are listed as reference material that you might find useful.
They are on reserve at Steenbock Library, and have also been ordered at the University Book Store.

- Regression Analysis by Example, 2nd ed.) by Chatterjee and Price
- An Introduction to Statistical Methods and Data Analysis by Ott
- Statistical Methods) by Snedecor and Cochran
- Quick Start to Data Analysis with SAS) by Dilorio and Hardy

The following item is required --- if you took 571 last fall you should already have it; if not, then
copies are available at Bob's Copy Shop.

- Course Notes for Statistics/Forestry/Horticulture 571) by Nordheim and Clayton

Course Objective: As in 571, the goal is to provide students in bioscience with a thorough grounding
in modern statistical procedures. The emphasis will be on understanding underlying concepts rather
than on an extensive coverage of a wide range of topics. The development of the ability to interpret
results and to evaluate critically the methods used is of paramount importance. To a large extent the
assignments will involve the analysis of data sets that approach the “real-world” complexity of data
encountered in research. Substantial use will be made of the computer in conducting such analyses.

Assignments: Homework assignments are due each week on Friday, in your TA's mailbox, by 400
pm. Assignments will be returned in discussion section. Homework assignments should be well
organized and reasonably neat. You must show your work to receive credit. Late homework
assignments will be penalized unless extenuating circumstances exist. If possible, prior arrangements
should be made in such cases.

Exams: There will be an in-class midterm, a take-home midterm, and a final exam. The exams will
cover lecture materials, readings, and homework material. Exams will be open-book and open-notes.
The in-class midterm will take place on March 23, the take-home midterm will be distributed on April
27, and due May 4. The final exam will take place on Friday May 16 at 1225 pm. Missed exams will
not be permitted except when extenuating circumstances prevail.
Grading: The homework will count 20%, the in-class midterm will count 20%, the take-home midterm will count 30%, and the final will count 30%.

Discussion Sections: Attendance is strongly advised. Sections will begin meeting the second week of class.

Syllabus for Statistics-Forestry-Horticulture 572
Spring 2000

- Linear Regression (about 11 lectures)
- Review of simple linear regression
- Residual analysis—-which plots to make, what to look for, tests for outliers, corrective action
- Transformations
- Weighted least squares
- Pure error versus model error
- Relationship between regression and ANOVA models
- Multiple regression—-order of fitting, testing reduced models
- Polynomial regression
- Multicollinearity
- Stepwise fitting procedures
- Autocorrelation
- Comparing regression lines
- Experimental Design and Analysis of Variance (about 16 lectures)
- Definitions of experimental unit and experimental error
- Role of blocking
- Importance of randomization
- One-way completely randomized design—-how to implement, model, power, assumptions, diagnostics (residuals)
- More on contrasts—-orthogonal contrasts, orthogonal polynomials
- Use of regression ideas (analysis of covariance)
- One-way random effects model—-inference on variance components, inference about the mean
- Subsampling—-role in experimentation, allocation of resources
- Randomized complete block design—-how to implement, when to use, model, limitations, subsampling
- Latin square design—-how to implement, when to use, model, designs with several squares
- Multi-factor designs (fixed effects)—-meaning of interaction, role of blocking within these designs, models
- Multi-way random effects and mixed models
- Split-plot design
- Repeated measures—-basic issues, useful approaches

Selected additional topics (as time permits)
Course Description

Biochemistry 711

Sequence Analysis (lecture)

Biochem 711(aka Biochem 875-001 for fall '98), 2 credits, Fall Semester 1998, Tues. and Thurs. 9:55-10:15 am, Biochemistry Building, room 134.

Timetable call number: 72294 for fall '98

Instructor: Prof. Ann Palmenberg, Biochemistry Dept (262-7519).

Graduate level lecture course for aspiring molecular biologists and genetic engineers, designed to answer the questions; "What can you do with your sequence once you have it?" and "How can you put this information into realistic biological perspective?"

Topics (partial listing) will include overviews of: RNA, DNA and protein structure; mechanisms of genetic change; sequence generation methods; comparison and alignment algorithms; motif recognition; 2D predictions; phylogeny calculations; database searching; discriminating coding criteria; phenotypic selection; phylogenetic reconstruction; and other exciting stuff.

Sign up by touchtone registration (call# (not available yet), 2 cr). Prerequisites: (a) graduate status, (b) Genetics 466 or Biochemistry 501 or equivalent. Course will be graded on class participation, in-class quizzes and exams, and written reports on outside reading topics. First class: Thursday Sept 3, 1998.

NOTE: There will NOT be classes on Tues., Sept 8 or Thurs., Sept 10 1998. (Instructor is not available)

ENROLLMENT LIMIT OF 35 STUDENTS

Note: Biochem 711 will be offered ONLY in 1998, 2000, 2002, and subsequent even years. This course will not be offered in 1999, 2001, etc.

Reading Lists and Grading Requirements

Press here to see typical reading list and grading requirements

Current Syllabus

1998

Previous Syllabi

1994

1996

Last Modified July 22, 1998
TEXTBOOKS

Students who take courses typically complain about having to buy expensive textbooks that are not necessarily followed during class lectures. In reality, there is no single textbook or even a reference book that covers all concepts that we present during the semester. Still, there are some books that are very good places to read more about the topics we will present, and which you might find additionally useful in your future careers. These books can be ordered/purchased through the bookstore if you wish to have personal copies. At least one copy of each is also on reserve at the Steenbock Library. (Class lecture notes and handouts will also be put on reserve at Steenbock after each lecture. Please give us a few hours to get them to the library!)


You can also obtain through the UW Bookstore two other books containing more "heavy duty" sequence analysis concepts. If you EVER anticipate getting involved in the molecular aspects of sequence analysis during your career, you will find these to be excellent reference manuals for algorithms and techniques. You may purchase them or read the copies on reserve at Steenbock.


GRADING

Exams, book reports, quizzes and in-class participation (attendance and attention) will each be awarded "points." Your grade will depend upon your total accumulated score according to the following formula:

Exams (3x): 50%
Book reports (2x): 30%
Quizzes (1-2) +
homework: 10%
Class participation 10%

---
100%

At the end of the semester, cumulative scores are summed for all of your work according to these proportions and final grades awarded
according to the class curve ("average" B).

SEQUENCE ANALYSIS 1998

READING LIST FOR BOOK REPORTS

During this semester, you will be required to read and comment on (e.g.: write reports) on selected outside materials. The purpose of these readings is to increase your exposure to the classical foundations of evolutionary theory and also to make you aware of some contemporary hypotheses and controversies in the field. Some, but not all of these materials MAY be discussed in class, but realistically, there is no other educational process which can effectively substitute for actual contact with original books/articles. (Translation: "Reading this stuff is GOOD for you.")

Accordingly, we have procured copies of (nearly) 50 different titles which cover a multitude of topics related to this course, or which we feel might enhance your educational experience. PLEASE NOTE: these books are NOT library copies. They belong personally to ACP and are being made available to you as an expedient convenience for class purposes. Failure to treat these resources with proper respect could definitely lead to a prejudicial attitude at grade time.

The materials are divided into 4 categories according to general topic and each group has been assigned a referent color code:

Red Books: Sequence Analysis
Yellow Books: Classic Perspectives (history)
Green Books: Contemporary Viewpoints
Blue Books: Genetic Concepts

The assigned color for each book title is listed below. It is repeated with a coded sticker on the spine of the volume and also on the signout card inside each front piece.

WHAT YOU HAVE TO DO:

1. **DONT PANIC!** You do NOT have to read ALL the books on the list (though at some point in your scientific career, you might want to, or wish you had).

2. During the term of the semester, you will choose ONE selection from EACH OF TWO different categories (i.e. two different colors), read at least the minimum recommended passages/chapters/sections, then write a summarizing-type "book report" (see below). Within these guidelines, the specific choice and order of individual books is left entirely to you.

3. Be aware that only one, or at most two copies of each book selection is directly available through the course. Therefore, if you wish to take out a particular book, be sure to claim it early, because someone else will probably have it later! If you REALLY, REALLY, REALLY want a certain book for this assignment, but don't get it during the class exchange days, you are free to acquire and read your OWN copy (e.g. library or bookstore) if you can dig one up. *Exception: The Double Helix is such a wimpy, easy book, that only the student who checks out the single class copy, may write a report on this one.* Class copies can't be reserved in advance. You'll just
have to fight it out on exchange days.

4. Since there ARE only a limited number of titles and copies, exchange day (when reports and books are due, and new books are picked up) will be rigorously enforced. The syllabus lists due dates for each of the selections. You may turn in any report earlier than the due date but late reports/books will not be accepted and will be issued a suboptimal grade.

5. Reports should be succinct, 1000-1500 words each (about 4-5 pages) and with your name, the title of the book, and its color referent on the top of the first page (or at the top of the file). You may submit your reports in either physical or electronic format.

PHYSICAL FORMAT: bring TWO legible, typed or computer printout copies (double-spaced) to class on the day the report is due. One will be graded and returned, the other will be filed.

ELECTRONIC FORMAT: send your report in ASCII format to aepalmen@facstaff.wisc.edu via e-mail so that it arrives as a dated message no later than 12:00 noon on the day the report is due. Your report will be graded, comments appended and a copy returned to you via e-mail. Please don't use an electronic format other than ASCII (delimited text), as I don't have the time to mess with non-generic word processor output or Binhex files!

6. The purpose of these reports is to convince me (ACP) that you have (1) read, (2) thought about, and (3) learned something from the chosen selections. The better you can convince me (in your report) of points 1-3, the higher your grade will be. The minimum recommended passages or sections for each book are outlined below, but you are certainly free to read the WHOLE thing if you so choose. (Possible bonus points may be awarded here, at my discretion, and depending upon the quality of your report and the difficulty of the material).

7. A word concerning plagiarism: DON'T! Academic dishonesty will be grounds for automatic failure of the assignment and the course. These reports are not tough assignments, but they must be your own original work, written in your own words. ACP has read all these books (believe it or not!) and has on file the second copy or an electronic copy of all previously submitted papers. The book reports should not be viewed as collaborative efforts to be carried out by or with other students, post-docs etc. However, if you are unsure about any technical points of writing or grammar, you may certainly ask others for help with these aspects, or contact the Student Writing Lab at 6171 Helen C. White Hall.

**RED BOOKS: (SEQUENCE ANALYSIS)**

Bishop, M.J. and Rawlings, C.J. (editors), 1987. Nucleic Acid and Protein Sequence Analysis: A Practical Approach. IRL Press, Oxford. (3X of 4) Excellent and detailed practical information about many of the sequence analysis programs and algorithms that are now in common use. From among chapters: 4, 6, 8-14, choose any 2 chapters and summarize the most important features of the described programs and the strengths and/or weaknesses of the algorithmic assumptions upon which they are based. Chapter 11, on RNA predictions is especially interesting.

Creighton, T.E. 1984. Proteins: Structures and Molecular Properties. W.H.Freeman & Co., New York. (1X). This text is commonly used in protein structure courses and is a superb compendium of the molecular forces that shape protein secondary and tertiary structure. Read chapter 3 "Evolutionary and Genetic Origins of Protein Sequences" and any other chapter except #1. On the basis of what you learn, why do you think it is so hard to accurately predict protein structures and/or functions from primary amino acid sequences? Can this ever really be done accurately? What kinds of information (biological, sequence or otherwise) would you like to have before predicting protein structure on the basis of primary data?

sequence analysis. Read part I "The Computer Approach." Identify and comment on many of the common pitfalls that typically waylay the naive sequence analyzer as he/she asks "What does my sequence mean?"

Doolittle, R.F. (editor) 1990. Molecular Evolution: Computer Analysis of Protein and Nucleic Acid Sequences. Methods in Enzymology, Volume 183. (Academic Press) (1X of 3). Excellent and complete summary of many topics covered in this course, as written by THE experts in the field. Select any 2 of the 7 sections, read and summarize the most important points or problems in the particular field.

Gribkov, M. and Devereux, J. (editors) 1991. Sequence Analysis Primer. Stockton Press, New York. (2X of 3). Excellent book offering the neophyte the necessary background to enter this exploding field, while helping the more seasoned researcher to fine tune their approach. If you really want to learn about sequence analysis, this is the place! The book is divided into 4 major sections. Read section 1 "DNA" and either section 2 "Protein" or section 3 "Similarity and Homology." Summarize the most important points or problems in the particular field.

Howe, C.J. and Ward, E.S. (editors) 1989. Nucleic Acids Sequencing. Oxford University Press, New York. (1X of 2). This is a REALLY technical "how-to" manual for hard-core sequencers. If you have ever had a problem in any aspect of sequence determinations, this book will tell you how to solve it at the bench. Read chapter 7 "Computing" and any other chapter that summarizes a technique that YOU might want to use in the lab. Summarize the experimental strengths of this approach, and outline the traps that unwary investigators might unwittingly succumb to.

Hunter, L. (editor). 1993. Artificial Intelligence and Molecular Biology. The MIT Press, Cambridge, Mass. (1X-P). An introduction by Joshua Lederberg is followed by 13 excellent chapters that attempt to bridge specific molecular biology problems, like protein and RNA structure prediction, with computer advances in neural networking and artificial intelligence. Read any 1 chapter and summarize the most important points (skipping over the mathematical points). Discuss the concept of artificial intelligence, and how it may or may not relate to a specific biological problem. Chapter 1, "Molecular Biology for Computer Scientists" is an excellent (but simple) review of basic genetics and biochemistry of proteins and nucleic acids.


MacLean, N. 1988. Oxford Surveys on Eucaryotic Genes. Oxford University Press, NY. (1X) Part of a series of books that provide a forum for authoritative reviews of particular genes or gene families. Each chapter is independent, written by key people in the field. The information content is dense. But, these are definitive summaries (as of 1988) and models for the types of comparative information that can be gleaned by studying sequences in combination with biochemistry. Read any 1 chapter (DNA supercoiling, cAMP kinases, RNA polymerases, etc.) and summarize how comparative sequence analysis contributed to the author's understanding of this gene family and its functions. These are model studies for how we should look at sequences.

Smith, D.W. 1994. Biocomputing: Informatics and Genome Projects. Academic Press, New York (1X-P). This survey book from the early 90's discusses many aspects of biocomputing, as covered in our class. The concepts and problems that still need to be solved are still valid today. Read chapter 4 (Comparative Sequence Analysis: Finding Genes) or chapter 6 (Phylogenetic Analysis and Molecular Evolution) or chapter 7 (Predictions of Protein 2D and 3D Structure) and summarize the important points.

Stone, E.M. and Schwartz, R.J. (editors) 1990. Intervening Sequences in Evolution and Development. Oxford University Press, New York. (1X-P). A series of 6 review articles (and an introduction) dealing with various concepts of exon splicing, and the putative role of this function in protein evolution. Each review is excellent and covers somewhat different aspects of the question. Read the introduction (chapter 1) and any other chapter of your choice. Briefly summarize the salient points and the author(s) view the world of exons and introns with respect to the molecular evolution of proteins.

von Heijne, G. 1987. Sequence Analysis in Molecular Biology: Treasure Trove or Trivial Pursuit. Academic Press, Inc., New York. (1X of 2). Another excellent basic text on sequence analysis (8 chapters). From among chapters 4, 5 and 6, select any 2 chapters and read. Summarize and comment on "What CAN you do with your sequence once you have it?"
GREEN BOOKS: (CONTEMPORARY VIEWPOINTS)

Dawkins, R. 1982. The Extended Phenotype. Oxford University Press, Oxford. (1X of 2). A collection of 14 chapters covering many ideas and concepts about evolution. Select any 4 chapters (preferably consecutive) and write 2-3 paragraphs on each describing (a) the main ideas and (b) your views on these ideas.

Dawkins, R. 1987. The Blind Watchmaker. W.W. Norton & Co., New York. (1X of 2). A collection of 11 chapters covering many important concepts and theories of evolution. Select any 4 chapters (preferably consecutive) and write 2-3 paragraphs on each describing (a) the main ideas and (b) your views on these ideas.

Dawkins, R. 1989. The Selfish Gene. Oxford University Press, Oxford. (2X-1P of 3). A collection of 13 chapters covering many interesting and exciting new concepts in evolution. Select any 4 chapters (preferably consecutively) and write 2-3 paragraphs on each describing (a) the main ideas and (b) your views on these ideas.

Dawkins, R. 1997. Climbing Mount Improbable. W.W. Norton & Co., New York. (1X). The jacket blurb says it all, "Dawkins is a genius of science popularization. If you have not read one of his books before, (this one) is a place to begin: it is nonstop mental and literary pleasure." A collection of 10 chapters covering some really interesting genetic observations (Check out the odd frog on pp97?). Select any 3 chapters (preferably consecutively) and write 2-3 paragraphs on each describing (a) the main ideas and (b) your views on these ideas.

Gould, S.J. 1973. Ever Since Darwin. W.W. Norton and Co., New York. (1X of 2). A collection of 33 essays divided into 8 sections. Read any 3 sections and write 2-3 paragraphs on each describing (a) the main theme and (b) why you agree or disagree with that theme. Note: section 8, "The Science and Politics of Human Nature" is particularly interesting!

Gould, S.J. 1980. The Panda's Thumb. W.W. Norton & Co., New York. (1X-P). A collection of 32 essays divided into 8 sections. Read any 3 sections and write 2-3 paragraphs on each describing (a) the main theme and (b) why you agree or disagree with that theme. Note: ladies in the class may be particularly interested in commenting on section 4, "Science and the Politics of Human Differences"

Gould, S.J. 1981. The Mismeasure of Man. W.W. Norton and Co., New York. (1X of 2?). A collection of essays divided into 7 chapters. Read any 3 chapters, including chapter 5 "The Hereditary Theory of IQ" and write 2-3 paragraphs on each describing (a) the main theme and (b) why you agree or disagree with that theme.

Gould, S.J. 1983. Hen's Teeth and Horse's Toes. W.W. Norton and Co., New York. (1X of 2). A collection of 30 essays divided into 7 sections. Read any 3 sections (including: EITHER section 3 "Adaptation and Development" OR section 5 "Science and Politics") and write 2-3 paragraphs on each describing (a) the main theme and (b) why you agree or disagree with that theme.

Gould, S.J. 1985. The Flamingo's Smile. W.W. Norton and Co., New York. (1X of 2). A collection of 30 essays divided into 8 sections. Read any 3 sections and write 2-3 paragraphs on each describing (a) the main theme and (b) why you agree or disagree with that theme. Use your own discretion to select sections of interest.

Gould, S.J. 1987. An Urchin in the Storm. W.W. Norton and Co., New York. (1X of 2). A collection of 18 essays divided into 5 sections. Read any 3 sections (including section 5 "In Praise of Reason") and write 2-3 paragraphs on each describing (a) the main theme and (b) why you agree or disagree with that theme. Use your own discretion to select the other two sections.
Gould, S.J. 1990. Wonderful Life. W.W. Norton and Co., New York. (1X of 2). A contiguous series of chapters chronicling the story of the Burgess Shale; the theory and ideas behind random nature of evolution that serendipitously led to the existence of the major taxa that are extant today. Read a minimum of pp 13-64, 79-102, 207-240, 277-319 (smaller sections within the major chapters) and describe (a) the importance of the Burgess Shale to modern evolutionary theory, (b) the importance of contingency (c) and the concept of the false order of the status of science.

Gould, S.J. 1992. Bully for Brontosaurus. W.W. Norton and Co., New York. (1X-P). A collection of 35 essays divided into 10 sections. This is considered by Gould himself and many others to be Gould's finest collection of essays. Read the following essays: Section 1 - essay 3 or 4, Section 3 - essay 8 or 9, Section 4 - essay 10 or 11, Section 5 - essay 15, Section 6 - essay 17 or 19, Section 7 - essay 21 or 22, Section 8 - essay 28 or 30. Write 1-2 paragraphs on each essay describe the main theme and your ideas & opinions on that theme.

Gould, S.J. 1995. Dinosaur in a Haystack. Random House, Inc., New York (1X-P). One of the more recent Gould books. continues along the same themes as he is famous for. This time, there are 34 short chapters, divided into 8 sections. Read all the short chapters within the same section or your choice, then summarize the main themes and your reactions to those themes. Section six on eugenics is especially provocative. See if you agree with any of these viewpoints!

Hubbard, R. and E. Wald. 1997. Exploding the Gene Myth. Beacon Press. Boston (1X-P). Popular science book that explores how genetic information is produced and manipulated by scientists, physicians, employers, insurance companies, educators and employers. To quote the dust jacket, "(They) have shown how the marriage of science and business has created that most treacherous of American progeny: commerce masquerading as human liberation." Or, "A much-needed antidote to the daily hype of biotechnology." Within 12 short chapters, the first 6 are historic background for non-scientists. Read any 3 of chapters 7-11, and briefly summarize the salient points. Your personal comments (agreement or outrage) on the issues raised in these chapters are warranted, since these are exactly the ethical and scientific issues you will be facing as a future scientists.

Kolata, G. 1998. Clone: The Road to Dolly and the Path Ahead. William Morrow and Co., Inc. NY (1X-P). The story of the scientists and science that resulted in the first cloning of a mammal from an adult somatic cell. The analysis focuses on the ethical implications, but the tone is written for the lay public. This is pretty light, read the whole book, comment on your personal opinions on the ethics and implications of this work for the future of bioengineering. Should human cloning be allowed? Is there a moral limit beyone which scientists should not cross, even if it is technically feasible?


Milunsky, A and Annas, G.J. (editors) 1985. Genetics and the Law. Plenum Press, New York (1X-P). Contains 35 short chapters, under 11 main headings, discussing the ethical and legal implications of recombinant genetics, gene therapy and related technologies, as they relate to current biotechnology and the law. Pick any 1 main heading, and discuss moral and scientific issues raised in those chapters that may confront researchers, doctors, administrators and lawyers of the present and future, as they strive to "improve the condition of human kind."

Nussbaum, M.C., and C.R. Sustein. 1998. Clones and Clones: Facts and Fantasies about Human Cloning (1X-P). A collection of 24 short essays divided into 5 parts, each discussing the moral, religious and scientific issues that now confront us in this new age of awesome biological technology. As young scientists, these issues, and the moral dilemas associated with them will lie squarely upon you. This is light reading, but it covers a lot of ground. Read all the essays associated with any one major section, and summarize the authors' main points, and your own personal and scientific reactions to their arguments. How does one restrict scientists from carrying out morally repugnant techniques, if some one is willing to pay for them? Where do you draw the line between personal scientific freedom, and ethical professional behavior?

to the dustjacket, the book "provides background for the startling headlines that quite possibly signal changes to all human life in the next century." Read chapter 8 (Moral Consequences of Molecular Biology) and any other one chapter and write 3-4 paragraphs on each describing (a) the main theme and (b) your opinions on these topics.

Wills, C. 1989. The Wisdom of the Genes: New Pathways in Evolution. Basic Books, a division of Harper Collins, USA. (1X of 2). This book deals with two questions: (1) Has the ability to evolve evolved? (2) Is there some way that the power of selection on individuals can be brought to bear on this ability to evolve? Read chapter 2 + 6 and 7, 9 or 10 and write 2-3 paragraphs on each describing (a) the main ideas and (b) your views on these ideas.

BLUE BOOKS: (GENETIC CONCEPTS)

Avers, C.J. 1989. Process & Pattern in Evolution. Oxford University Press, Oxford. (1X-P). Excellent detailed text on evolution with 11 chapters on separate topics. Read chapter 9 ("Phylogenetic Analysis") and any other one chapter. Comment on the accuracy of using phenotypic evidence or characters to reconstruct genotypic evolutionary histories. Summarize the important point(s) from the additional chapter that you select.

Brown, M.H. 1990. The Search for Eve. Harper & Row, New York. (1X of 2). Contemporary journalistic treatment of the hypothesis that all extant human lineages can probably be traced through a single, original maternal line, and that thus, there WAS a unique "Eve" at some point during evolutionary history. Read the whole book, its long, but light. Write a report summarizing the Eve hypothesis and your opinions on its validity.


Gould, J.L., and C.G. Gould. 1989. Sexual Selection. Scientific American Library, New York. (1X of 2). An easy to read collection of 9 chapters discussing the origin and nuances of sexual selection. Read chapter 1 and any other 3 chapters (you MAY want to read the whole thing). Comment on the concept that sex is war between two entities that share a common ancestry, for better or for worse! Is sex a good thing (from an evolutionary point of view)?

Hillis, D.M. and Moritz, C. (editors) 1990. Molecular Systematics. Sinauer Associates, Inc., New York. (1X of 2). Don't let the title or the size of this book freak you out (this is the big white one)! This is one of the best all around biotechnology methods books that was ever written. The only reason we don't use THIS as our class text, is it's hefty price! There's LOTS of good stuff in here. Of particular interest is Chapter 11 "Phylogenetic Reconstruction," by David Swoford et al. Read Chapter 11, skimming over the mathematical arguments (unless they really interest you) and concentrate on, "practical considerations, the nuts-and-bolts issues that (may) confront any investigator faced with the analysis of a new data set." Pick 2-3 questions of interest and briefly summarize how some important variables influence the selection of a "final" representative tree.

Kimura, M. (editor) 1992. Molecular Evolution, Protein Polymorphism and Neutral Theory. Springer Verlag. (1X of 2). Classic book supporting the hypothesis that those molecular changes that are less likely to be subjected to natural selection occur more rapidly in evolution. These concepts explain many of the actual variant populations observed at the molecular level in sequence analysis. Read chapter 1, which summarizes the theory, and any other chapter of interest (#9-12 are especially appropriate to sequence analysis). Summarize the concepts of neutral theory and whether or not you agree with them.
Kimura, M. 1985. The Neutral Theory of Molecular Evolution. Cambridge University Press, Cambridge. (1X of 2). An excellent and comprehensive text devoted to neutral theory. Read chapters 1, 2 and 7 (substitute another chapter if you wish) and summarize the concepts and how they may apply to sequence analysis.

Mayr, E. 1963. Animal Species and Evolution. Harvard University Press, Boston. (1X of 2). Classic reference for the concept and origin of biological species. Read chapters 1, 2 and any other 2 chapters of your choice. Summarize the basic principles and how these might be important to molecular evolution.


Smith, J.M. 1989. Evolutionary Genetics. Oxford University Press, Oxford. (1X of 2). Excellent text on evolution and the molecular forces that contribute to selection. Pick any three chapters (except #1) and succinctly summarize the basic concepts.

Stanley, S.M. 1979. Macroevolution. W.H.Freeman, San Francisco. (2X of 3). This is a classic and important text describing the hypothesis of quantum (or macro) evolution, based on principles of punctuated equilibrium. Read the first 3 chapters and any other 1 chapter that captures your fancy. Summarize each in 1-2 paragraphs.

Terzaghi, E. 1984. Molecular Evolution. Jones and Bartlett, Boston. (1X of 2). An anthology of key classical papers (publications) by the “greats” in the field that have contributed major concepts to molecular evolution. The representative papers are organized by common theme. Pick any 4 papers, read and summarize the salient features and discuss how these particular concepts are or are not still in vogue in current hypotheses.

YELLOW BOOKS: (CLASSICAL PERSPECTIVES)

Aldridge, S. 1998. The Thread of Life. Cambridge University Press, UK (1X-P). A real beginners book on biotechnology and molecular biology with a focus on how these technologies can be applied to non-scientific problems such as producing vegetarian cheese or cleaning up the environment. Easy to read, but not very challenging. Read any 4 (of 12) chapters, including preferably at least one chapter from the Biotechnology section. Summarize the main concepts and your reactions to how modern genetic engineering and/or biotechnology may have a significant future impact on our lives and our environment. Is this going to be a good thing?

Berra, T.M. 1990. Evolution and the Myth of Creationism. Stanford University Press, Stanford, CA. (1X-P). A lucid and insightful view of one of the most important ideas in the history of science. Read the whole thing (it’s short) and summarize the main concept and why you agree or disagree with the author’s premise.

Bishop, J.E. and Waldholz, M. 1990. Genome: The Story of Our Astonishing Attempt to Map All the Genes in the Human Body. Touchstone Press, New York. (1X-P). Title pretty well summarizes the contents. Read the whole thing, its long but (relatively) light. Do you believe that the genome initiative project will ever be truly successful? Will it really be worth the time, effort and money to see it through to completion?

Cook-Deegan, R. 1994. The Gene Wars., W.W. Norton & Co., New York (IX-P). A firsthand account of the protracted political and scientific struggle to launch the human genome initiative. The 5 main sections are divided into 20 subchapters. Read the first section “The Scientific Foundation,” and any other section that summarizes the primary contribution of a major “player” in the launching of this modern field (e.g. Watson, Gilbert, etc.). What were the political and scientific key obstacles in developing the genome initiative? Comment on whether you think the potential promise of this initiative is worth the cost? Why or why not?
Cravens, H. 1988. The Triumph of Evolution: The Heredity-Environment Controversy, 1900-1941. John Hopkins University Press, Baltimore. (2X of 3). A history (4 chapters) describing the conflict between science (evolution) and culture (religion). Read chapters 3 and 4 and summarize (a) the main "lesson" and (b) why you agree or disagree with that lesson.

Darwin, C. 1958. The Origin of the Species. New American Library, New York. (2X of 3). Darwin's classic report, upon which much of modern evolutionary theory is based. Read (reverently) chapters IV and V ("Natural Selection" and "Laws of Variation") and any other one chapter. For each chapter, write 2-3 paragraphs describing (a) the main "lesson" and (b) why you agree or disagree with that lesson, and (c) how these ideas are (or are not) presently accepted in modern science.

Darwin, C. 1989. Voyage of the Beagle. Penguin Books, London. (2X of 3). A personal chronicle of Darwin's observations during his journeys on the Beagle. Read the Introduction and any 3 (of 23) chapters. Describe the types of observations assimilated by Darwin during his journeys and how this information may have led to alterations or refinements of his theory on natural selection. Would YOU want to visit any of these locations today? Why?


An incredible summary of the history of seminal developments in the field of molecular biology and molecular evolution from one of the "fathers" of modern biology. The book contains 5 major chapters and a summary section. Read chapter 4 (The Gene), or chapter 5 (The Molecule), or the conclusion chapter (The Integon). Summarize and comment on the salient points and how these concepts may potentially influence your options for career choices in the future world of molecular biology.

Ridley, M. (editor) 1987. The Darwin Reader. W.W. Norton & Co., New York. (1X of 2). Sections have been extracted from Darwin's 9 most important books in which he explains his concepts of evolution. From among chapters 2-10, read any 2 chapters and write 3-4 paragraphs on each describing how these concepts influenced or set the course for modern evolutionary ideas.

Watson, J.D. 1968. The Double Helix. New American Library, New York. (1X of 3). The story of how the structure of DNA was resolved. Read the whole thing, it's short! Write a classic book report on who was who, and how they contributed to the story. Do you believe today, that Watson and Crick should be accorded most of the kudos accorded to the resolution of the DNA structure?

Preliminary Syllabus as of August 24, 1998 (to be updated during semester)

SEQUENCE ANALYSIS: Biochem 711

Fall Semester 1998, 2cr. Tues/Thurs, 9:55-10:45 a.m.,
Biochemistry Bld. 420 Henry Mall, room 132

Dr. Ann Palmenberg (ACP), Professor, Dept. of Biochemistry and Institute for Molecular Virology. Room 301 AH&BS Bld., 1655 Linden Drive; Ph:262-7519; e-mail: acpalmen@facstaff.wisc.edu

or: Room 390 Biochemistry Bld, Henry Mall (~Oct. '98)


Sept. 8, Tues: No Class

Sept. 10, Thurs: No Class

Sept. 15, Tues: Nucleotides, DNA and Mutation: The Basic Stuff of Sequences (ACP)

NOTE: Pick up 1st book on reading list.

Sept. 17, Thurs: Unusual DNAs, Restriction Mapping, Intro to Computer-Aided Methods (ACP) NOTE: possible in-class quiz on single letter codes.

Sept. 22, Tues: Amino Acids and Proteins (ACP)

Sept. 24, Thurs: Proteins Structure Prediction (ACP)

Sept. 29, Tues: Protein Structure Motifs (Dr. Jean-Yves Sgro)

Oct. 1, Thurs: Orfs, Urrs and Biological Codes (ACP)

Oct. 6, Tues: High Tech Sequencing Methods (Dr. Lloyd Smith, Dept of Biomolecular Chem., UW-Madison)

Oct. 8, Thurs: In class exam on material to date. Students wishing more time may begin at 9:30 am.

Oct. 13, Tues: PAMs and Related Substitution Matrices (ACP)

Oct. 15, Thurs: Sequence Comparisons with Dotplots (ACP)

Oct. 20, Tues: Global and Local Optimal Alignments (ACP)

Oct. 22, Thurs: Profiles and Multiple Sequence Alignments (ACP)

Note: 1st book report due, pick up 2nd book

Oct. 27, Tues: Databases and Database Searching (ACP)

Oct. 29, Thurs: BLAST/Fasta: The Search Tools (ACP)

Nov. 3, Tues: BLAST: The Fallout (how to use it) (ACP)

Nov. 5, Thurs: In class exam on material since previous exam Students wishing more time may begin at 9:30 am.

Nov. 10, Tues: Primer Design and PCR (Dr. Doug Storts, Promega Corp.)

Nov. 12, Thurs: DNA Forensics (Dr. Jim Schurman, Promega Corp.)

Nov. 17, Tues: Forestry 101 (Introduction to Phylogenic Trees) (ACP)

Nov. 19, Thurs: Phylogenic Reconstruction (ACP)

Nov. 24, Tues: Database Mining (Dr. Fred Blatner, Dept. of Genetics, UW-Madison)

Nov. 26, Thurs: Thanksgiving (no class)
Dec. 1, Tues: RNA Structures and Optimal Folding Algorithms (ACP)
Dec. 3, Thurs: Suboptimal RNA Algorithms and Prediction Confidence (ACP)

Note: 2nd book report due.

Dec. 8, Tues: TBA
Dec. 10, Thurs: TBA

Dec. 15, Tues: In class exam on semester material. Students wishing more time may begin at 9:30 am.

Guest Speakers '98 (confirmed)
Dr. Lloyd Smith, Department of Biomolecular Chemistry, UW-Madison
Dr. Fred Blattner, Department of Genetics, UW-Madison
Dr. Jean-Yves Sgro, Biotechnology Center, UW-Madison
Dr. Doug Storrs, Promega Corp., Madison
Dr. Jim Schurman, Promega Corp., Madison

Last Modified August 24, 1998
Course Description

Biochemistry 712

Sequence Analysis (laboratory)

Biochem 712, 1 credit, Fall Semester 1998, Tues. 1:20-3:30 pm, BNMC Computer Lab, 1240 Biotech Center

Instructor: Prof. Ann Palmenberg, Professor, Dept. Biochemistry and Institute for Molecular Virology (phone# 2-7519, acpalmen@facstaff.wisc.edu)

NOTE: There will NOT be a Lab on Tues., Sept 8. (Instructors are not available)

First lab: Tuesday, Sept 15 (students should purchase 712 labbook at Bob’s Copy Shop beforehand)

Description:

Graduate level, hands-on laboratory course, taught at actual computer terminals, designed to complement and reinforce the sequence analysis concepts presented in the didactic course, Biochemistry 711.

Topics will include: overview of UNIX including file and directory management; introduction to sequence analysis; database searching; pairwise and multiple sequence comparison methods; alignments; profiles and profile searching; pattern recognition; phylogenetic reconstruction; database information retrieval; desktop molecular graphics and modeling; RNA folding; and other neat stuff.

Sign up by touchtone registration (Biochem 875 for fall of ’98). Non-negotiable prerequisite: concurrent or previous registration in Sequence Analysis 711 (previously listed as Biochem or AH&BS 875). Lab will be graded on class participation, computer-based homework assignments, in-class exams/quizzes, and instructor evaluation of student progress and effort.

ENROLLMENT LIMIT OF 16 STUDENTS !!

Biochem 712 is offered ONLY in 1998, 2000, 2002, and subsequent even years. This course will not be offered in 1999, 2001, etc.

In alternate years (1999, 2001, etc.), there is an undergraduate-level sequence analysis computer laboratory course. A. Palmenberg and J.-Y. Sgro also offer several not-for-credit Instructional BioModule Courses on genetics computing and molecular graphics computing. For further information, check the ACP web page: www.bocklabs.wisc.edu/~acp
CS 838 - Bioinformatics (Fall 1999) - Topics Course

General Course Information

- Instructor: Mark Craven  
craven@biostat.wisc.edu  
Office: 5730 Medical Sciences Center (corner of Charter and  
University)  
Office Hours: 3-4:00pm Tuesday & Thursday, or by appointment

- Prerequisite: CS 367 or equivalent
- Meeting Time and Location: 1:00-2:15 TR, 3345 Engineering
- Archive of class e-mail

Course Overview

The science of molecular biology is undergoing a revolution in how it is practiced. In the last decade, a vast amount of data (DNA sequences, protein sequences, etc.) has become available, and computational methods are playing a fundamental role in transforming this data into scientific understanding. Bioinformatics (computational biology) is an exciting new area that involves developing computational methods for managing and analyzing information about the sequence, structure and function of biological molecules and systems. The goals of this course are to provide an understanding of:

- the fundamental computational problems in molecular biology
- a core set of widely used algorithms in computational biology.

Students will also gain familiarity with the most important databases and servers used in the field (e.g. GenBank, SwissProt, MEDLINE,...).

Course Requirements
The grading for the course will be be based on:

- homework assignments: ~30%
- midterm exam: ~30%
- project: ~35%
- class participation: ~5%

The homework assignments will involve a mix of written problems and programming.

---

Readings and Lecture Notes

- Introduction to Molecular Biology and Bioinformatics
  - required reading
    - Primer on Molecular Genetics
    - selected pages from Los Alamos Science issue on The Human Genome Project
  - optional reading
    - Sequencers Endorse Plan for a Draft in 1 Year (5/99)
    - A New Five-Year Plan for the U.S. Human Genome Program (1993)
  - lecture notes
    - General Course Information (9/2)
    - Introduction to Molecular Biology (9/2, 9/7)
    - Introduction to Bioinformatics (9/7, 9/9)
  - resources
    - Genbank
    - SWISS-PROT
    - Protein Data Bank
    - MEDLINE

- Pairwise Sequence Alignment
  - required reading
  - lecture notes
    - Pairwise Alignment (9/9, 9/14, 9/16)
    - Substitution Matrices and Database Searching (9/16, 9/21)
  - resources
    - NCBI's BLAST server

- Multiple Sequence Alignment, Hidden Markov Models and Sequence Motifs
  - required reading
    - "Multiple Sequence Alignment Methods" chapter from R. Durbin, S. Eddy, A. Krogh,


- lecture notes
  - Multiple Sequence Alignment (9/23)
  - Hidden Markov Models, Part 2 (9/30, 10/5)
  - Learning Sequence Motifs with EM (10/7)
  - More on Motifs (10/12)

- resources
  - PROSITE database
  - Blocks database
  - MEME server
  - Multiple Alignment Servers
  - SAM HMM code
  - HMMER HMM code
  - PFAM database of multiple alignments

- Finding Genes in Genomic DNA
  - required reading

- lecture notes
  - Finding Genes in Genomic DNA (10/12)
  - Finding Genes in Genomic DNA: The GRAIL System (10/14, 10/19, 10/21)

- resources
  - GRAIL server and the Computational Biosciences Section at ORNL
  - extensive bibliography on computational gene recognition

- Gene Expression Data and Molecular Medicine
  - required reading
    - M. Eisen, P. Spellman, P. Brown, D. Botstein. Cluster Analysis and Display of
- T. Golub et al. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring [PDF] [HTML].

- optional reading
  - N. Friedman et al. Using Bayesian Networks to Analyze Expression Data. Submitted to RECOMB 2000.
  - NY Times article on pharmacogenetics

- lecture notes
  - Clustering and Classifying Gene Expression Data, Part 1 (10/26, 10/28)
  - Clustering and Classifying Gene Expression Data, Part 2 (11/4)
  - Inferring Regulatory Networks from Gene Expression Data (11/8, 11/10)
  - Knowledge-based Avoidance of Drug-Resistant HIV Mutants (11/10)

- resources
  - Affymetrix
  - Molecular Classification of Cancer data set from Lander’s group
  - Yeast Cluster Analysis data set from Stanford
  - Thorsten Joachims’ SVM code

- Phylogenetic Trees
  - required reading

- lecture notes
  - Constructing Phylogenetic Trees (11/16, 11/23)

- resources

- Protein Structure Prediction
  - required reading

- lecture notes
  - Protein Structure Prediction, Part 1 (11/23, 11/30)

- resources
  - secondary structure data set from UC Irvine Machine Learning Repository
● Information Extraction from Biomedical Text Sources
   ○ required reading

---

**Guest Lectures**

- *DNA Sequencing and Assembly* (11/2), Prof. Fred Blattner, Dept. of Genetics, University of Wisconsin
- *Whole Genome Comparisons using Suffix Arrays* (12/2), Dr. Bob Mau, Dept. of Genetics, University of Wisconsin
- *Optimal and Suboptimal RNA Folding* (12/7), Prof. Ann Palmenberg, Dept. of Biochemistry, University of Wisconsin
- *The Protein Threading Approach to Protein Structure Prediction* (12/14), Prof. Rick Lathrop, Dept. of Information and Computer Science, University of California at Irvine

---

**Homework Assignments**

Homework assignments are due at the start of the class on the due date. Homeworks turned in after the start of the class are considered late. Each student will have a budget of two unpunished late days for the semester. After these are used, there will be a penalty of 10 percentage points per day that an assignment is late.

- Assignment #1
  Due: September 28
  Solutions
- Assignment #2
  Due: October 19
- Assignment #3
  Due: November 23

**Course Projects**

- 1-page project proposal due November 9
- final project reports due December 21
- project information

**Exam**

- in-class, Thursday, November 18
SPRING SEMESTER
Special 8 Week Offer

FUNDAMENTALS OF
CLINICAL TRIALS
Statistics 542

by

David L. DeMets, Ph.D.

This course is intended for medical researchers interested in the design, conduct, and analysis of clinical trials. The course will develop skills in being a more critical reader of the medical literature or provide the tools to design your own protocol. Examples of clinical trials from the literature throughout the course.

The course is based on the text "Fundamentals of Clinical Trials" by Friedman, Furberg, and DeMets (3rd edition). Topics include:

- Defining the hypothesis and evaluation measures
- Defining the population
- Basic study designs
- Rationale and techniques for randomization
- Sample size estimation
- Recruitment issues
- Data collection
- Assessment of outcome - including adverse effects, quality of life, and survival
- Data monitoring
- Analysis issues
- Reporting of results

SCHEDULE: Tuesday and Thursday, 4:45 pm - 7:00 pm for 8 weeks
First class: Tuesday, January 25, 2000

LOCATION: G5/113 Clinical Sciences Center, 600 Highland Avenue

MATERIALS: Class Lecture Notes - Available from Bob's Copy Shop
Selected clinical trials articles will be handouts.

CONTACT: David L. DeMets
Department of Biostatistics and Medical Informatics, K6/446, 600 Highland Avenue
Phone: 263-2947; e-mail: demets@biostat.wisc.edu
1. Classes

Classes are given in an 8 week block of two per week.

Each class is divided into two parts.
   a. Lecture on basic fundamentals
   b. Discussion of published RCT's

2. Material

   a. Required
      • RCT papers (provided in class)

   b. Optional
      • Lecture notes (provided in class)

3. Homework

Homework will be assigned and graded. Some assignments are given on, for example - randomization, sample size, and survival analysis. Solutions are given on the homework as a handout. Homework prepares you for writing your protocol project.

4. Projects

   a. RCT Critique
      A published RCT will be given near the end of the eight weeks of lectures. You are to critique this article, evaluating the strengths and weaknesses, using the fundamentals presented in the lectures. The critique should not be longer than 5 typed pages.

   b. Protocol
      You are to write a protocol with all of the components presented in some detail. (An example is provided.) You may select the disease or question that is of most interest to you. You may also form a partnership with one other classmate. Get my approval of your topic before you start. Caution: Do not wait until week 8 to start.

4. Grades

The grades are 80% based on the critique (30%) and the protocol(50%). Homework accounts for 20%.
STATISTICS 542
FUNDAMENTALS OF CLINICAL TRIALS

Over the past three decades, randomized clinical trials have become one of the basic research tools in medicine to evaluate the benefits and risk of new therapeutic or prevention strategies. These may be pharmacologic, biologic, device, a procedure or behavioral modalities. Despite this wide range of modalities and disease processes, the basic fundamentals of clinical trial design are applicable. This course, based on the text by Friedman, Furberg, and DeMets, introduces the basic fundamental concepts of clinical trial design without requiring technical statistical training beyond an introductory course.

The course uses standard lecture material summarizing the fundamental concepts plus a series of published clinical trials to illustrate the concepts along the way. The course starts with an overview of clinical research, indicating the unique role that clinical trials play in the research spectrum which includes anecdotal observations, observational cohort studies and planned prospective experiments. The course focuses mostly on the comparative, also called Phase III, randomized clinical trial (RCT).

The fundamentals start with defining the question to be tested in the RCT. Many studies fail because the question is not well defined, either in concept or in the measurement of it. Since all of the design issues follow from the definition of the question, this is probably the most important step to be resolved. Once the question is carefully laid out, there are several standard RCT designs that can be selected. Central to most of these designs is the process of randomization in the assignment of the experimental or control strategy to individual participants. Several randomization techniques are outlined which achieve comparability of risk factors and other demographics between the experimental and control arms.

The size of the study is always a crucial component to the design of any study since cost and effort are a direct function of this sample size. Depending on the question, the nature of the primary outcome measurement for that question and the basic trial design, sample size concepts and calculations are outlined. For some situations, the answers can be presented simply in tables or graphs. Concepts such as significance level and power are conceptually described and their role in the sample size estimation as indicated.

Several operational issues are also covered including recruitment of participants, data collection and quality control. Many studies also fail because the amount of data collected exceeds what is necessary and what is affordable. The burden of collecting too much data that is not critical can consume resources and jeopardize the main goals of the study. Thus, this issue must be given careful consideration.

A brief presentation of survival analysis is included since this is such a common statistical technique in the analysis of many clinical trials, but not always taught in introductory statistical courses. One of the major activities during the conduct of any RCT is monitoring accumulating data for evidence of harm or early benefit, as well as timeliness and quality of data. While data must be monitored for ethical, scientific, and fiscal reasons, the process of repeated evaluation of
When data for evidence of harm or early benefit can lead to erroneous conclusions if special care is not taken. Statistical methods are presented which allow for early stopping if interim results are sufficiently convincing, but these methods control the false positive error rates to conventional levels; that is false positive error rates of 1% or 5%.

Once the trial is complete, the data must be analyzed and reported. Several common mistakes in analyzing the data are presented, such as not including all the participants in the analysis due to errors in verifying eligibility or due to patient non-compliance to therapy. While many of these reasons sound compelling at first, they can introduce substantial bias into the comparison of experimental and control arm outcomes and thus must be avoided. Finally, RCTs must be reported succinctly, but still be accurate and detailed to convey to the reader the essential components. Failures in proper reporting may damage the interpretation, or even worse, the credibility of the trial which may have been conducted and analyzed properly.

While the fundamentals described are very basic and straightforward, the challenge in the successful design, conduct, and analysis of an RCT is to implement all of these basic elements together.
<table>
<thead>
<tr>
<th>Class</th>
<th>Date</th>
<th>Topic</th>
<th>Chapter</th>
<th>Read Articles</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>01/25</td>
<td>Introduction</td>
<td>1-3</td>
<td>1 Epi Limits</td>
</tr>
<tr>
<td>2</td>
<td>01/27</td>
<td>Defining the Question</td>
<td>1-3</td>
<td>2 Guyatt</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3 Lawrie</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4 IONDT</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5 ONIT</td>
</tr>
<tr>
<td>3</td>
<td>02/01</td>
<td>Design</td>
<td>4,6</td>
<td>6 Fluid Resuscitation</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>7 CDP</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>8 4S</td>
</tr>
<tr>
<td>4</td>
<td>02/03</td>
<td>Design</td>
<td>4,6</td>
<td>9 ACTG (3 papers)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>10 MILIS</td>
</tr>
<tr>
<td>5</td>
<td>02/08</td>
<td>Randomization</td>
<td>5</td>
<td>11 HERS</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>12 DCCT</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>13 DCCT Protocol</td>
</tr>
<tr>
<td>6</td>
<td>02/10</td>
<td>Sample Size</td>
<td>7</td>
<td>14 GISSI</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>15 ISIS-I</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>16 PHS</td>
</tr>
<tr>
<td>7</td>
<td>02/15</td>
<td>Sample Size</td>
<td>7</td>
<td>17 ATBC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>18 CARET</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>19 PHS - Beta Carotene</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>20 Editorial</td>
</tr>
<tr>
<td>8</td>
<td>02/17</td>
<td>Baseline, Recruitment, Data Collection</td>
<td>8,9,10</td>
<td>21 NSABP-06</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>22 NSABP-06 Audit</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>23 Consent</td>
</tr>
<tr>
<td>9</td>
<td>02/22</td>
<td>Quality Control, Adverse Events, Quality of Life</td>
<td>11,12,13</td>
<td>24 NOTT</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>25 IPPB</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>26 CSSG</td>
</tr>
<tr>
<td>10</td>
<td>02/24</td>
<td>Survival</td>
<td>14</td>
<td>27 CASS</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>28 ECOG</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>29 PRAISE-I</td>
</tr>
<tr>
<td>11</td>
<td>02/29</td>
<td>Data Monitoring</td>
<td>15</td>
<td>30 BHAT</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>31 Tamoxifen</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>32 DRS</td>
</tr>
<tr>
<td>12</td>
<td>03/02</td>
<td>Data Monitoring</td>
<td>15</td>
<td>33 CAST</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>34 MERIT</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>35 VEST</td>
</tr>
<tr>
<td>13</td>
<td>03/07</td>
<td>Analysis Issues</td>
<td>16</td>
<td>36 ART ’78</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>37 ART ’80</td>
</tr>
<tr>
<td>14</td>
<td>03/09</td>
<td>Meta-Analysis</td>
<td>16</td>
<td>38 CARVEDILOL</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>39 Furberg, et al</td>
</tr>
<tr>
<td>15</td>
<td>03/21</td>
<td>Reporting/Close Out</td>
<td>17,18,19</td>
<td>40 FIT</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>41 PROSTATE</td>
</tr>
<tr>
<td>16</td>
<td>03/23</td>
<td>Class Discussion</td>
<td></td>
<td>42 Sample Protocol Project</td>
</tr>
<tr>
<td>4</td>
<td>04/06</td>
<td>Critique Due</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>04/20</td>
<td>Protocol Due</td>
<td>10</td>
<td></td>
</tr>
</tbody>
</table>

**3/11-19** SPRING BREAK

**15** 03/21 Reporting/Close Out

**16** 03/23 Class Discussion

**04/06** Critique Due

**04/20** Protocol Due

*REVISED 1/26/00*
CS 540 - Introduction to Artificial Intelligence

General Course Information

This course is offered each Fall and Spring semester.

Topics Covered

- Knowledge representation using predicate logic, semantic networks, neural networks, rules, Bayesian networks
- Deductive and probabilistic problem solving, planning, learning
- Applications in topics such as expert systems, game playing, vision, natural language understanding, speech recognition, robotics

Prerequisite: CS 367

CS 540 Pages of the Various Instructors

- Chuck Dyer (Spring '01)
- Jude Shavlik (Fall '00)

Local AI-Related Links

- PhD Qualifying Exam in AI
- U-Wisc AI Group
- U-Wisc Computer Vision Group
- U-Wisc Machine Learning Group
- U-Wisc Robotics Group
- U-Wisc Computational Biology (includes some AI)
- U-Wisc CS Dept

Graduate AI Courses at Wisconsin
• CS 731 - Advanced AI
• CS 760 - Machine Learning
• CS 766 - Computer Vision
• CS 780 - Robot Motion Planning

External AI-Related Links

Last modified: Tue Aug 20 15:00:00 1996 by Jude Shavlik
shavlik@cs.wisc.edu
CS 760 - Machine Learning (Fall 1999)

General Course Information

• **EXAM:** 7:15-9:15pm on December 14, 1999 in Room 1240 CS & Stats
  One-page of notes (8.5 x 11, both sides can be used) and calculators allowed.

• **Project reports** are due by 5pm on December 22. They should be 7-10 pages long, and a 12-point font should be used.

• **Instructor:** Jude Shavlik
  6357 CS & Stats
  shavlik@cs.wisc.edu
  Office Hrs: Tuesday/Thursday 3-4pm

• **Teaching Assistant:** Kim-Ee Yeoh
  1336 CS & Stats
  yeoh@cs.wisc.edu
  Office Hrs: Monday and Wednesday 1-2pm, and by appointment (send email)
  Office Phone: 265-9453

• Prerequisite: CS 540 or equivalent

• Archive of Class Email (send mail to this list)

Course Overview and Requirements

The intent of this course is to present a broad introduction to machine learning, including discussions of each of the major approaches currently being investigated. Class lectures will discuss general issues in machine learning, as well as present established algorithms. One secondary goal is to compare and contrast the various approaches, determining under which conditions each is most appropriate. Another is to relate these algorithms to human learning processes.
The work in the course will consist of 5-6 homework assignments (about one every two weeks), one exam (very late in the semester), and a course project. Your solutions will be partially automatically graded, so they must be written to run on the instructional Sun Sparcstations.

The homework assignments will generally be programming assignments that involve experimenting with machine learning algorithms and experimental methodology. Lab reports insightfully analyzing your experimental results will be required. Occasionally there may also be "paper-and-pencil" homework problems.

The final project can be a more ambitious experiment or enhancement involving algorithms used in the homeworks or the implementation of an additional algorithm. In either case, the implementation should be accompanied by a short paper (7-10 pages with a large font) describing the project. Projects not involving programming are also possible.

Homeworks will count for 35% of the grade, the "midterm" exam for 35%, the project 25%, and quality class participation 5%. The course will be graded on the conventional (A-F) system.

Lecture Notes

Digitized versions of the lecture transparencies are available as a new experiment this term. (Access is limited to University of Wisconsin sites.)

Reading Assignments

Assigned December 7, 1999:
Read Sections 2.1, 2.2, 2.3, 2.7, and 2.8 of Mitchell’s textbook.

Assigned November 30, 1999:
Read Chapters 11 and 12 of Mitchell’s textbook.

Assigned November 23, 1999:
Read Chapter 9 of Mitchell’s textbook.

Assigned November 11, 1999:
Read Chapter 7 of Mitchell’s textbook.

Assigned November 2, 1999:
Read Chapter 10 of Mitchell’s textbook.

Assigned October 26, 1999:
Read Chapter 13 of Mitchell’s textbook.

Assigned October 14, 1999:
Chapter 1 of the book "Advances in Kernel Methods: Support Vector Machines" by B. Schlkopf, C. Burges, and A. Smola
(A simpler intro from IEEE Intelligent Systems is also on-line)

Sections 1 & 2 (pages 1-14) of Machine Learning Research: Four Current Directions by T. Dietterich.

Assigned October 5, 1999:
   Read Chapter 4 of Mitchell’s textbook.

Assigned September 21, 1999:
   Read Chapter 3 of Mitchell’s textbook.

Assigned September 9, 1999:
   Read Sections 2.1-2.3 and Chapter 5 of Mitchell’s textbook.

Assigned September 2, 1999:
   Read Preface, Chapter 1, Section 6.1, 6.2, 6.9, & 6.10, and Section 8.1-8.2 of Mitchell’s textbook.

Homework Assignments

• Homework 4: Learning from Reinforcements - Q-Learning
  Due Thursday, 11/18/99

• Homework 3: Training Neural Networks
  Due Thursday, 10/28/99

• Homework 2: Inducing Decision Trees
  Due Tuesday, 10/12/99

• Homework 1: kNN, Naive Bayes, and Experimental Methodology
  Due Thursday, 9/23/99

• Homework 0: Creating Your Personal Dataset
  Due Thursday, 9/9/99

• Late policy on HWs:
  • HWs are due at 5 pm.
  • Each student will have FIVE "free" late days for use over the semester. Once these are exhausted, there will be a penalty of 10% per day (measured 5pm-to-5pm; weekends are free). "Days" are calendar days, not class days.
  • To make the TA’s job tractable, no HWs will be accepted more than one week late.

• Academic Misconduct

All examinations, programming assignments, and written homeworks must be done individually. Cheating and plagiarism will be dealt with in accordance with University procedures (see the
Academic Misconduct Guide for Students). Hence, for example, code for programming assignments must not be developed in groups, nor should code be shared. You are encouraged to discuss with your peers, the TAs or the instructor ideas, approaches and techniques broadly, but not at a level of detail where specific implementation issues are described by anyone. If you have any questions on this, please ask the instructor before you act.

Previous Exams (postscript until 1999; MS Word after)

- Fall 1999
- Fall 1998 | Fall 1997

Some ML-Related Links

- The UC-Irvine ML Dataset Archive | The UC-Irvine KDD archive | more datasets
- The WEKA Machine Learning Project (code for many ML algo’s, as well as some datasets)
- Journals: Machine Learning (the on-line page) | Data Mining and Knowledge Discovery | Journal of AI Research | AI Magazine | IEEE Neural Networks Council (several journals are connected to this page)
- Papers from the 1998 Intl. Conf. on Machine Learning (held in Madison)
- Knowledge Discovery in Databases
- Pointers to ML Courses
- Neural Network Resources
- Some ILP Stuff
- Some SVM Stuff
- Machine Learning Benchmarking
- International Society for Adaptive Behavior
- AI Bibliography Server | Neural Networks Bibliography Server (Austrian AI Institute)
- AI Resources (Canadian NRC Server)
- Aha’s Links to People in ML
- Reinforcement Learning Repository
- More External AI References

Help with Programming Assignments

- Some Java Tips
- The CS Lab’s Java page

Related Local Links

- U-Wisc ML Group
- U-Wisc ML & Math-Programming Group
- U-Wisc Comp Biology (includes some ML)
- U-Wisc AI Group
- U-Wisc CS Dept
- U-Wisc Library
- More local links
Instructor: Chuck Dyer

Telephone: 262-1965
E-mail: dyer@cs.wisc.edu
Office Hours: 2 - 3 Tuesdays and Thursdays, and by appointment

Teaching Assistant: Scott McElroy

Telephone: 262-1721
E-mail: mcelroy@cs.wisc.edu
Office Hours: 1 - 2 Mondays and Wednesdays, and by appointment

Student Photo Gallery

Final Scores and Grades

What’s New?

General Course Information

Introduction to the basic concepts in computer vision. First, an introduction to low-level image analysis methods, including image formation, edge detection, feature detection, and segmentation. Image transformations (e.g., warping, morphing, and mosaics) for image synthesis. Methods for reconstructing three-dimensional scene information using techniques such as shape from shading and depth from stereo. Active vision methods for scene recovery such as occluding contour detection by viewpoint control. Motion and video analysis. Three-dimensional object recognition.

- Schedule
11:00 a.m. - 12:15 p.m. Tuesdays and Thursdays in 1325 CS

- **Prerequisites**
  - CS 540, calculus, linear algebra, and C/C++ programming knowledge

- **Grading**
  - Midterm Exam: ~20%
  - Final Exam: ~25%
  - Homework assignments: ~35%
  - Project: ~20%

- **Syllabus**
- **Textbook**

- **Other Readings**
  - A collection of additional readings from journals and conference proceedings will be available online or at DOIT Documentation

**Supplementary Reading Sources**

- **Web Page:** [www.cs.wisc.edu/~cs766-1](http://www.cs.wisc.edu/~cs766-1)
- **Class E-Mail Alias:** cs766-1list@cs.wisc.edu
- **Class E-Mail Archive**

---

### Reading Assignments

- 4/22: Chapter 10 (except 10.3)
- 4/15: Chapter 8 (except 8.4.2, 8.5.2)
- Chapter 9
- Chapter 7 (except 7.4)
- 3/2: Chapter 5
- 2/18: Articles on view morphing, mosaics, and voxel coloring (see additional readings)
- 2/4: Chapters 3 and 4 (except 4.4), plus the article "Edge Detection" by D. Huttenlocher
- 1/26: Chapter 2 and Vista Documentation
- 1/19: Chapter 1

---

### Homework Assignments

---

### Lecture-Related Materials

---
image processing toolboxes are especially relevant. Matlab Tutorial (Univ Utah)

- **Test Images**
  Most test images will be put in the directory `/p/vision/images/` although they may require format conversion to be used. Some other images may be put in `~cs766-1/public/images/` Numerous image databases are also accessible via the WWW; for example, see the collection of test images at CMU

---

### Examinations

- **Midterm Exam:** Thursday, March 18 in class
  - Midterm Exam Solution
  - Histogram of Student Scores
- **Final Exam:** Tuesday, May 4 in class
  - Covers topics since the Midterm Exam, namely readings listed above in Chapters 5, 7, 8, 9, and 10 of the textbook, HW 4, the Williams and Shah paper, lectures and lecture notes.
  - Final Exam Solution
  - Histogram of Student Scores
- **Format:** Each exam will cover topics up through that time, including readings in the textbook, papers, and homework assignments. You may bring into each exam one (1) 8.5" x 11" sheet of paper with any notes you want on both sides. Exams will focus on main ideas and algorithms, not derivations or proofs. See old exams below for the types of questions that will be asked.
- **Old Exams**
  - Midterm Exam - Spring 1999
  - Midterm Exam Solution - Spring 1999
  - Final Exam - Spring 1999
  - Final Exam Solution - Spring 1999
  - Exam - Spring 1998
  - Exam Solution - Spring 1998
  - Exam - Spring 1997
  - Exam Solution - Spring 1997
  - Exam - Fall 1995
  - Exam Solution - Fall 1995
  - Exam 1 - Spring 1994
  - Exam 2 - Spring 1994
  - Exam 1 - Spring 1992
  - Exam 2 - Spring 1992

---

### Links of Interest

- Computer Vision Home Page (CMU) (highly recommended)
- Vision papers search engine (MIT)
- CVonline: Computer Vision Online Course Notes (Edinburgh)
Computer Accounts

- Accounts
  Course accounts are on the instructional Unix workstations (sols and vegas) running Solaris. You should use the color workstations if you are displaying images. Each class account has a disk space quota of 32 MB so you can store images for homeworks and your project. Be sure to delete old images and compress others (see \texttt{gzip(1)}), however, in order to save space.

- Email
  Email sent to cs766-1list goes to everyone in the class including the instructor and TA

- Printers
  To print images you should use one of the laserprinters, \texttt{laser1 - laser4}, which are located in room 1359. Alternatively, the generic printer name \texttt{laser} will send output to one of the four printers with the shortest queue. Caution: Before sending images to the printer, be sure to check the queue; if there are a lot of jobs being printed it is bad manners to send images to be printed because they take so long to print. Be considerate!

- Vision Software
  - Vista
    The Vista programming environment will be used in the homework assignments. The Solaris-compatible code is located in the directory \texttt{/p/vision/ip-tools/vista/} Man pages are in \texttt{/p/vision/ip-tools/vista/man/} and executables are in \texttt{/p/vision/ip-tools/vista/bin/} Add \texttt{/p/vision/ip-tools/vista/bin} to your path to use these. There is also a command called \texttt{cvw} in the bin directory, which is nice for displaying (and modifying) Vista images.
  - Xv
    \texttt{xv(1)} is an interactive image display program for the \texttt{X} window system that is very useful for displaying images in a variety of formats. Installed in \texttt{/s/std/bin/} and \texttt{/s/xv/bin/}
  - ImageMagick
    \textit{ImageMagick} is an X11 image display tool. Executables and man pages are installed locally in \texttt{/s/Imagick/} The command \texttt{display} is a very versatile alternative to \texttt{xv}. Add \texttt{/s/Imagick/bin} to your path to use these.
  - IMG*
    70 basic image processing operations invoked using Unix-like command lines. Code, executables and manual are in \texttt{/p/vision/ip-tools/imagstar/} Add \texttt{/p/vision/ip-tools/imagstar/bin} to your path to use these.
  - Khoros
    The Khoros image processing software development environment provides a set of basic image processing modules and a graphical programming language interface for rapid prototyping of simple image processing algorithms. The code is located in the directory \texttt{/p/vision/ip-tools/khoros} \texttt{/p/vision/ip-tools/khoros/bin/cantata} is the executable that starts up the interactive environment.
  - Netpbm
    A toolkit for conversion of images between a large variety of different formats. Based on the Pbmplus package. Installed in \texttt{/s/std/bin/} and \texttt{/s/pbm/bin/}
  - Matlab
    \textit{Matlab(1)} is a numeric computation and visualization environment. Signal processing and
The Biomedical Information Science and Technology Initiative

Prepared by the Working Group on Biomedical Computing
Advisory Committee to the Director
National Institutes of Health
June 3, 1999

CHARGE TO THE WORKING GROUP ON BIOMEDICAL COMPUTING

The biomedical community is increasingly taking advantage of the power of computing, both to manage and analyze data, and to model biological processes. The working group should investigate the needs of NIH-supported investigators for computing resources, including hardware, software, networking, algorithms, and training. It should take into account efforts to create a national information infrastructure, and look at working with other agencies (particularly NSF and DOE) to ensure that the research needs of the NIH-funded community are met.

It should also investigate the impediments biologists face in utilizing high-end computing, such as a paucity of researchers with cross-disciplinary skills. The panel should consider both today's unmet needs and the growing requirements over the next five years (a reasonable horizon for extrapolating the advances in the rapidly changing fields of computing and computational biology).

The result of deliberations should be a report to the NIH Director, which will be presented to the Advisory Committee to the Director. The report should include recommendations for NIH actions to support the growing needs of NIH-funded investigators for biomedical computing.

EXECUTIVE SUMMARY

In science and technology in the latter half of the 20th century, two fields have stood out for their speed of progress and their effect on society: biomedicine and computation. The charge of this Working Group is to assess the challenges and opportunities presented to the National Institutes of Health by the convergence of those two disciplines.

The principal obstacle impeding effective health care is lack of new knowledge, and the principal mission of the NIH is to overcome this obstacle. At this point the impact of computer technology is so extensive it is no longer possible to think about this mission without computers.

Increasingly, researchers spend less time in their "wet labs" gathering data and more time on computation. As a consequence, more researchers find themselves working in teams to harness the new technologies. A broad segment of the biomedical research community perceives a shortfall of suitably educated people who are competent to support those teams. The problem is not just a shortage of computationally sophisticated associates, however. What is needed is a higher level of competence in mathematics and computer science among biologists themselves. While that trend will surely come of its own, it is in the interest of the NIH to accelerate the process. Digital methodologies — not just digital technology — are the hallmark of tomorrow's biomedicine. The NIH therefore must find ways to discover, encourage, train, and support the new kinds of scientists needed for tomorrow's science.

To make optimal use of information technology, biomedical researchers need, first of all, the expertise to marry information technology to biology in a productive way. New hardware and software will be needed, together with deepened support and collaboration from experts in allied fields. Inevitably, those needs will grow as biology moves increasingly from a bench-based to a computer-based science, as models replace some experiments and complement others, as lone researchers are supplanted by interdisciplinary teams. The overarching need is for an intellectual fusion of biomedicine and information technology.

Invariably, scientists learn best by doing rigorous science. Indeed, the NIH mission is to do science, including teaching and learning. Socially meritorious goals of improving human health and preventing, detecting, diagnosing, and treating disease and disability are achieved most effectively when pursued within the overall context of rigorous science. This report and its recommendations focus, therefore, on science — both for its insights and as a path toward building an educated interdisciplinary workforce. The centerpiece of our recommendations is the proposal to inaugurate National Programs of Excellence in Biomedical Computing. It is in the context of those National Programs that the best opportunities can be created for doing and learning at the interfaces among biology, mathematics, and computation. With such new and innovative programs in place, scientists will absorb biomedical computing in due course, while supporting the mission of the NIH.

Recommendation #1:

The NIH should establish between five and twenty National Programs of Excellence in Biomedical Computing devoted to all facets of this emerging discipline, from the basic research to the tools to do the work. It is the expectation that those National Programs will play a major role in educating biomedical-computation researchers.

National Programs of Excellence in Biomedical Computing would advance research in particular areas of biomedicine, focusing on those in which computation is becoming increasingly essential. They would be funded in part through a new program, and in part through research grants from one or more of the Institutes that make up the NIH. The academic or research institutions at which the National Programs would be housed would be expected to contribute to the programs — and teaching would be an essential contribution.

National Programs could range in size. At a modest level, three to five researchers in complementary disciplines might receive $1.5 million a year to undertake the exploration of a single problem. Larger National Programs might bring together several problems and several technologies, perhaps in association with more than one institution or Institute, for up to $8 million a year. The NIH will determine the number and scope based on the applications and the grant process.

One important goal of the National Programs will be to develop and integrate the use of computational tools to meet the important challenges of biomedical research. These Programs are in keeping with the conclusions of the President's Information Technology Advisory Committee http://www.nih.gov/about/director/060399.htm
(PTAC) report in that it focuses on basic information technology research in the pursuit of insight into the issues facing biomedical research. Concurrently, the National Programs will create homes for interdisciplinary teams, and those teams will establish nurturing environments for exploration and education. In establishing National Programs, the NIH will send a powerful message, both in academe and within the NIH community itself, about the importance of computation and the value of interdisciplinary research.

Strong action by the NIH is required because the existing biomedical research and teaching structures of the universities and research institutions of this country inadequately value interdisciplinary efforts generally, and computation in particular. Few grant programs and fewer academic departments foster the kind of interdisciplinary work required to meet biomedical challenges, let alone educate students about them. National Programs specifically would include formal and informal instruction from the undergraduate through post-graduate levels, and incorporate a range of opportunities for scholars and researchers to participate.

Recommendation #2:

To make the growing body of biological data available in a form suitable for study and use, the NIH should establish a new program directed toward the principles and practice of information storage, curation, analysis, and retrieval (ISCAR).

The information that biomedical researchers are amassing in profuse quantities today from the Human Genome Project, clinical trials, statistics, population genetics, and imaging and visualization research creates enormous digital repositories of information. The scale of those databases swamps all the information collected before. They encompass multigigabyte, multivariate functions of three dimensions, of wavelength, and of time. The infrastructure needed to make them available is phenomenal: A single biomedical laboratory could produce up to 100 terabytes of information a year—about the same as the information in one million encyclopedias. In order to be useful, the data must be indexed and stored, and the challenges for data analysis and abstraction are formidable.

The creation and development of advanced databases and database technologies (methods for storing, retrieving, and analyzing biomedical data) is becoming more important in all biomedical fields. The emerging technology of bioinformatics helps researchers gather and standardize data from basic research and computer modeling, and combine and manipulate databases to tease out the knowledge they contain. The goal is a system of interoperable databases that will make available the fruits of the increased productivity enabled by computation.

That is particularly true in clinical research: As more information from clinical trials becomes available, the need for standardization and interoperability of clinical databases will increase dramatically. Coordinating knowledge gained from clinical trial data with new insights from genetic research could appreciably advance knowledge about the treatment of disease. A system of interoperable databases would allow clinical researchers to track any finding back to its basic science roots; conversely, a research scientist might track forward to postulate from hypotheses through potential applications based on innovative uses of existing data.

As the amount of data grows, the tools to compare and manipulate the data become more important. These tools form software bridges between databases that will allow researchers to link disparate information sources.

The NIH has been a leader in establishing databases of valuable information and making them available for study. Now it must organize and expand database resources internally and externally. Currently the agency uses contracts, grants, and cooperative agreements in bioinformatics, but no program focuses specifically on database development. Both the collection of the information, and the creation of the tools for storage, management, and access are increasingly important. Therefore, the NIH needs a program that will rally new and important bioinformatics efforts and build this vital part of the biomedical infrastructure.

Recommendation #3:

The NIH should provide additional resources and incentives for basic research (through R01 grants) to provide adequate support for those who are inventing, refining, and applying the tools of biomedical computing.

Biomedical scientists know best what they need, and they often need to take advantage of computational opportunities. However, in evaluating research grants and programs, reviewers and staff sometimes have been reluctant to provide support for computation and computational infrastructure at the level required. The computational infrastructure, of course, includes not only the hardware but also the people with the expertise to make good use of the hardware. It is time for the NIH to recognize the importance of both the tools and those who build them. In order to do that, the NIH needs to ensure that R01 grants may be used for biomedical computation. That is particularly important for grants that support environments rich in teaching potential as well as research excellence. Researchers who work with students should have the resources that will allow them to set an example of the use of biomedical computing.

As with any special emphasis or targeted funding, evaluation at three years is recommended.

Recommendation #4:

The NIH should foster a scalable national computer infrastructure. To assure that biomedical researchers can gain access to the computing resources they need beyond their desktops, the NIH should provide financial resources to increase computing capacity, both local and remote. The purpose of this recommendation is to establish a balanced infrastructure for all computational needs.

Biology is becoming increasingly complex and computation is becoming increasingly sophisticated. Today's biomedical computing needs resources that go beyond desktop computers to local clusters of processors, to mid-level facilities, and to the most powerful computers at national centers. Many biomedical researchers cannot do their work on their desktop computers alone. They need varying amounts of computing power at different times, and those resources should be made available. The infrastructure must be better balanced for a dynamic range of computational needs.

Powerful computers alone are not enough. The entire support system must be in place. Even researchers who can do their work on small clusters need access to the expertise to set up and manage those clusters, and need support from programmers who can write or adapt the necessary software. As the computing-power needs increase, so do the support needs.

The NIH should support facilities with mid-level computers where new algorithms and applications can be developed specifically for biological problems. The biomedical expertise at those facilities would support researchers seeking to adapt and apply the best computer technology to their work. For some applications, mid-level facilities could offer smaller versions of scalable systems that exist at the national supercomputer centers. Researchers might use those resources to test and develop code or design before moving to national supercomputer centers, or—in appropriate cases—to do their work on more powerful computational resources than they have in their laboratories. Mid-level facilities could be created through National Programs that focus on supercomputing science, or the resources could be made available through cooperative agreements with...
existing extramural centers as well as at intramural centers.

NIH scientists have long taken advantage of the national supercomputer centers run by the National Science Foundation and the Department of Energy for high-level computing. Because the number of biomedical researchers who can profit from using those facilities is increasing, the NIH should take a strong leadership position and help support the national supercomputer centers. Such NIH support would provide a welcome opportunity for a partnership between NSF and the NIH as the future of science unfolds in the 21st century.

CONCLUSION

The NIH can make a powerful contribution to the development of tomorrow's biomedical research community by increasing efforts to promote and support computational biology today. With the appropriate support in place, interdisciplinary research teams will coalesce for National Programs of Excellence in Biomedical Computing and ISCAR efforts. The natural byproduct of their emphasis on biomedical research will be a new generation of researchers who are skillful with computing, and who will have helped to create the computational tools they need to meet tomorrow's challenges.

As biomedical research becomes more computationally intensive, the Biomedical Information Science and Technology Initiative (BISTI) is essential if the NIH is to fulfill its mandate. This Initiative will be the means by which new techniques are developed, new knowledge is discovered, new research communities are created, and new ideas are disseminated to the institutions and people who can use them to solve the mysteries of life and health.

PREFACE

Methods that dramatically expand biological data also demand new modes of analysis and new ways to ask scientific questions.  
- Harold Varmus, NIH Director

Only the most rudimentary elements of biomedicine and computation were known in 1950. Development of the essential ideas and the technologies to implement them began with the discovery of the DNA structure and the construction of the first practical digital computers. Although there are intellectual connections between the two fields — DNA encodes the program for life — biomedicine and computation have advanced largely independently. And both have advanced with a rapidity that is unprecedented in history. The functional capacity of computing machines has doubled every 18 months, in accord with the prediction encapsulated in Moore's Law. At the same time, the increase in known genomic sequences — information relevant to our own genetic endowment — is being submitted to GenBank at a rate of more than 5,000 sequences (over 2 million nucleotides) per day. Computation has already transformed industrial society; a comparable biotechnological transformation is clearly on the horizon. Yet only in the last few years has it become clear that those two exponentially growing areas are now actually converging.

That convergence is already obvious in modern medicine. Medical diagnosis has been revolutionized by a suite of modern clinical-imaging methods including computed axial tomography (CAT scans), magnetic resonance imaging (MRI), and ultrasonography. Each of them is fundamentally a computational method. The rate of their development has generally been limited by the availability of affordable computation capacity; the physical methods and concepts were waiting for computation to catch up. In the basic science of genomics, the acquisition and analysis of genomic DNA sequence has computation at its heart. Without highly capable computers, algorithms, and software, DNA sequences would have little practical value, even if we could determine them without computation. Another obvious example of the convergence is protein structure determination: The rise of crystallographic and magnetic resonance methods is bound to Moore's Law. Today even the rudimentary visualization of a protein structure requires a computer with functional capacity unknown in 1960, unaffordable in 1980, and routinely available as a commodity today.

On the horizon are developments that will require and generate more data than science is currently prepared to utilize or assimilate. For instance, nanotechnology machines that function like minuscule test tubes and minuscule pumps will allow investigators to deliver suitable dosages of medicine responding to biological signals, and capture cellular-level information about disease. The chemist's pharmacological intuition is fast being replaced by high-throughput screenings, delivered at the rate of 50,000 or more tests a week, to track the exact effect of any drug or chemical substitute. Those advances are contingent on advances in computation.

The dominant trend in biomedical science and medical practice, as in every realm of science, is the increasing value and usage of computers. Computers in our laboratories are becoming as necessary and ubiquitous as laboratory instruments. The complexity of today's problems demand that the research scientist now spend less time doing experiments and more time figuring out what they mean. The data so painstakingly extracted in past years are now, through progress in biomedicine, produced in such volumes as to require computers just to read them. The scientist spends more and more time using the computer to record, analyze, compare, and display their data to extract knowledge. Libraries are being taken over by computers as well, and clinical practice is becoming increasingly computerized—not even considering electronic patient records and billing.

Despite all those well-known realities, the convergence of computation and biomedicine is poorly reflected in our universities and schools of medicine. Biomedical computing is not a recognized discipline, and despite the extraordinary demand for people with good education in both domains, computing, only a few cross-disciplinary training programs exist. Recognition of the convergence of biomedicine and computing is also quite limited among the agencies that fund biomedical and computation research. This Working Group was established to offer recommendations to remedy that situation at the National Institutes of Health.

MEETING THE POTENTIAL

Science rides on insight, that flash of understanding that suddenly gives a researcher a new way to explain a phenomenon. Insight itself comes from the hard work of gathering bits of information and ordering them, taking pieces of the puzzle and rearranging them until a new picture emerges. The process might be simple if the puzzle had a fixed number of pieces; in biology, hundreds of new puzzle pieces are added every day. To keep up with that flood of data, and to help order it, biomedical researchers are increasingly using computation. Computers are becoming puzzle-assembling tools.

But the computers, algorithms, and software, let alone the support infrastructure, are not keeping up with the exponentially rising tide of data in biomedical research. There is a consensus that much of the delay is in the lack of computational expertise in the clinics and the biomedical laboratories. Biomedical researchers need to know better how to use the powerful technology that both informs and advances their work, but the time spent developing that expertise should not come at the expense of time spent focusing on basic scientific problems. Today's researchers need

http://www.nih.gov/about/director/06396.htm
the option to work closely with colleagues who know the computing part of biomedical computing as well as the investigators who know the biomedical part. It is an inevitable (and welcome) mark of research progress and success that the problem space has grown too large to be tackled predominantly by lone researchers. A team might be able to turn data into databases, turn intuition into algorithms, turn processes into computer programs. It is a rare and unlikely individual, today or in the future, who can do all of those things solo at the state of the art.

For these reasons the primary recommendation of this Working Group is the establishment of National Programs of Excellence in Biomedical Computing.

With National Programs of Excellence bringing together interdisciplinary teams, researchers will be able to harness the power of tomorrow’s computers by collaborating to develop mathematical models, write software, and adapt systems. Team members can cooperate on algorithm development, software development, database development, and system development. They can make computers useful research tools, from high-performance systems in biomedical laboratories to ultra-high-performance systems in national centers. Such teams can help biomedical research move to a new horizon where new paradigms, ideas, and techniques can emerge. Biomedicine needs human power to utilize the computer power.

For many biologists, however, that human power is not available, making it hard for them to use even the tools now available to them. Many are bemoaning the lack of the human resources they need to use the computational resources that could be so helpful. The situation in biomedical research is the same as the situation in other research specialties: It now takes a cadre of experts. Just as every surgeon requires a team of nurses, medical technicians, and anesthesiologists, a computational biologist requires a team of software engineers, computer technicians, and biomedically trained algorithms to do the best work. The focus of the National Programs of Excellence in Biomedical Computing will be research; the subtext will be the opportunity to bring together related specialties and train a new generation of researchers whose skills cross-disciplinary boundaries.

The National Programs might focus on one or more of the following areas of biomedical computing: biology, medicine, algorithms, software, database research, or devices (e.g., image capture). The spectrum of research will be from the fundamental level of scientific discovery to usable tools to do science, all of which are vital to tomorrow’s biomedical research.

A Program of Excellence might be cross disciplinary or focused entirely on biology or medicine; it might be cross-institutional or at a single institution, or it may stand alone; it might pinpoint a single problem in the field, or several. The distinguishing features would include:

- A range of work, from fundamental discoveries to useful tools in biomedical computing.
- A plan for disseminating the results of the research-and-development effort, so that others can take advantage of the data that is produced, the tools that are created, and the science that is discovered.
- A full menu of education, ranging from formal undergraduate and graduate programs to courses and seminars for students and working researchers, visiting-scientist programs, "total-immersion" programs, one-week or two-week accelerated-training programs, and other innovative programs to help spread the knowledge gleaned in the course of research. That training would underline the scientific effort within the Program.

National Programs of Excellence in Biomedical Computing will answer the question of who will do computation tomorrow by educating students at all levels, with an emphasis on bachelor's and master's students to fill today's critical need for people with cross-disciplinary knowledge. Programs may be housed at a university or they may be freestanding and link to several universities; they will provide some new faculty positions and integrate and coordinate existing resources. They will offer short courses to biomedical researchers, and encourage visiting scientists.

THE NATIONAL PROGRAMS OF EXCELLENCE

Computation is becoming an enabling technology for biomedicine; some of the most exciting, challenging, and hardest problems posed to computing and computational scientists are emerging from the biomedical field.

Examples of the scope of the problems (and the cognate opportunities) abound:

SURGERY

Advanced medical-imaging systems give surgeons a better view of a patient's anatomy than they can see in front of them on the operating table. With computers that create three-dimensional models of real-time MRI scans, and programs that incorporate that model into a video of the operation in progress, surgeons can more precisely cut and suture, knowing both the extent of a tumor and its relationship to adjacent healthy tissue.

In other work, researchers are exploring the use of computer models to help surgeons decide whether to recommend surgery for stenosis, the narrowing of an artery. MRIs measure the flow of blood around a blockage, but they cannot measure the pressure on artery walls. Working together, surgeons, experimentalists and theoreticians, are building mathematical models of the pressure in the artery based on fluid dynamics.

Other researchers are exploring a computer-based virtual-reality interface with tactical feedback that would allow remote control of micro-surgical tools. Although that work is still in its early stages, it might eventually allow surgeons to perform microscopic surgery with minimal invasion, checking their progress and effectiveness with remote sensing devices, and thus reducing trauma and promoting healing.

A National Program devoted to the application of computing to surgery would concentrate the skills and knowledge of a range of experts on developing the hardware and software tools that are needed to bring computing into the operating room. It would also educate and train the physicians, bioengineers, programmers, and technicians who will develop and apply the new computer-based surgical techniques.

CLINICAL PRACTICE

In the not-too-distant future, clinicians will be able to match reconstructed images of a tumor deeply hidden in the body with a genetic characterization of a tumor, correlating the tumor's growth and metastatic involvement (the microcosm of the disease) with the patient's clinical response (the macrocosm of the disease). Imaging technologies might automate tissue-pathology analysis, leading to greater diagnostic accuracy.

Such work requires basic science research to amass the baseline data that allows that kind of exciting application of computationally based clinical medicine. A National Center focused on clinical practice could coordinate that kind of research and its direct application to human health.
Researchers use bioinformatics tools to create models that help them understand data in large problem spaces — from whole systems to whole organisms. That new understanding of the data helps them form hypotheses about biological systems. Scientists whose research once encompassed a single gene or a single protein are using bioinformatics to study integrated functions among tens of thousands of genes. In a now-classic example of the changes wrought by bioinformatics, a team of scientists discovered previously unknown sets of interrelationships when they did a standard fibroblast experiment on thousands of genes instead of the handful of genes that had been studied previously. They found a system far more complex than anyone had imagined. As biomedical researchers develop ways of dealing with large data sets, they can make leaps in understanding those more-complex systems.

The Human Genome Project will require tools that can handle information on three billion base pairs — DNA units. The HGP, when it is completed early in the next century, will give biology the equivalent of a periodic table of the elements for human systems. Tomorrow's researchers will be Kepler's to the Tycho de Brahes who are today sequencing the human genome. But with three billion base pairs and 100,000 genes in the human genome that could be involved in disease, biomedicine needs better techniques to store and identify genes and gene groups, and better methods to analyze them.

The study of the techniques and methods that allow researchers to collect and process data toward the understanding of the life and death of organisms is the essence of bioinformatics. It incorporates database creation, curation, and access. Some of the specific problems bioinformatics researchers are facing include:

- Standards. Terminology, syntax, and semantics need to be defined and agreed upon to allow integration of data.
- Curation. Database submissions need to be checked and cross-referenced to avoid the transitive propagation of error.
- Interoperability. Data should be as consistent as possible across databases so that researchers can compare and contrast it. For instance, three genomic databases (those concerned with the genomes of yeast, flies, and mice) are jointly producing a genetic ontology so that every biological process and function common to all three organisms can be referred to with the same words. Where databases are not consistent in schema, researchers need tools that will make transparent the querying and analysis across databases.

The database issue is in part a computational issue. To store and manipulate databases that have answers to biomedical questions hidden in thousands or hundreds of thousands of data points requires a level of sophisticated manipulation that grows more difficult as the volume of data grows. Moreover, the information needs to be presented in a format that humans can use: Reducing ten million data points to ten thousand still presents more information than a human mind can encompass. Writing the software that will turn those data points into models is a conceptual challenge.

Database issues are also systems issues. Biomedical researchers increasingly need databases of images and software as well as databases of numeric data. Those databases need to be housed on computers powerful enough to manipulate all the data quickly for many researchers at the same time.

Finally, there are research and policy issues. When are specialized databases appropriate, and how is that decided? How long should they be maintained, and by whom? What standards should apply? How should they be interconnected?

The Information Storage, Curation, Analysis, and Retrieval program this Working Group has proposed would give the NIH a way to support and advance databases and database development directly, either through grants or by establishing National Programs of Excellence focused on the special problems of data and its use. It would allow the NIH to reward proposals for research aimed at gathering and testing data, not just for research intended to test hypotheses.

**Infrastructure**

To deal with increasing amounts of biomedical data, the research community needs access to scalable computing systems. The need for computation is growing in bioinformatics analysis as well as in molecular dynamics and bioengineering simulations. The need is growing exponentially as the data from imaging and sequencing balloon and the use of computational simulations snowballs. Computational facilities are vital as biologists tackle more and more complex problems.

Researchers who five years ago spent little time on computers report that they now spend 90% of their research time in front of their monitors. Much of that change is because of the development of important biomedical databases such as those at the National Center for Biotechnology Information. Investigators have come to depend on those databases in their work. A study late last year showed that usage is increasing at 10% to 15% a month. In 1991 there were 195 searches a day. By 1994 that had increased to 5,000 a day. Last year there were 600,000 a day. At that rate, the NCBI databases will be used more than 25 million times a day by 2002. During the same period, the amount of determined DNA sequence had increased from 71 to 217 to 2,008 million base pairs. Sequencing the human genome (three billion base pairs) is expected to be completed sometime shortly after the turn of the century.

Those large databases require that researchers have available both the hardware and the software to manipulate them, either remotely or — when the application is unique — on their desktops. They also need to handle large datasets such as those used for imaging or simulations. A 3-D image that has a resolution of 1024 by 1024 by 1024 pixels contains at least a gigabyte of data. At least eight gigabytes of data are required for an image that is 2048 by 2048 by 2048, and clinical researchers and clinicians are demanding resolution beyond what the technology can offer today.

Biologists report problems finding funds for infrastructure support to maintain the computational resources in their laboratories: network routers, file servers, printers, and other facilities that are shared among many grantees. A great need is for people with the expertise to manage those systems and tailor them for biomedical uses. Those problems are exacerbated by the rapidly growing demand for local computer clusters where researchers can quickly turn around computational problems.

Some researchers have had to find novel ways to get the computational resources they need. One team used a major corporation's computers at night and on weekends to do its protein-folding analyses. In all, they used three times the computational resources that had been awarded for all their research projects for a year. Because the computing resources were made available, they were able to try new computational experiments, with good results. Unfortunately, such public-private partnerships are hard to put together, and so most research teams make do with inadequate equipment and power.

The unrelenting pressure on computational technology is evident in the increase in the usage of the nation's high-performance computing centers. At the National Science Foundation's supercomputer centers, for instance, out of the 50 largest projects in fiscal 1998, biomolecular computation consumed more resources than any other scientific discipline. That year the supercomputing cycles doubled, yet two-thirds of the requested cycles were turned down because of lack of sufficient resources. According to the NSF, 12% of all investigators who use their supercomputer centers are in biology, and they account for 25% of all cycles — an increase of 54% from fiscal 1997 to fiscal 1998. The biologists who used the NSF supercomputers used large amounts of time, not just a few hours, suggesting that for less-intensive applications researchers were able to find...
It is worth noting that fundamental discovery is the foundation for such advances in medicine, but because of the diversity of diseases as complex as cancer, the ultimate impact of a discovery on the treatment of human disease almost always requires studies in human populations, that is, clinical studies.

Weaknesses in computing support for clinical research — quality assurance, varying capabilities for electronic data capture, connectivity on the Internet, security and privacy of data, and high-speed communication between laboratories, to name a few — pose enormously expensive problems. This Working Group has not attempted to deal directly with those problems, but recommends that when NIH Institutes fund clinical research they be sensitive to the need for computing, connectivity, and high-speed links between laboratories studying the bases of disease.

NEUROBIOLOGY

Neurobiologists working on the brain's ability to process information are limited not by their ideas, but by the tools to create realistic models of brain function. Until recently, neurobiologists have been able to record only the activity of single cells; new technological advances allow them to record from hundreds or even thousands of cells at the same time. With that breakthrough, the focus has turned to creating the techniques that will allow monitoring and analysis of the large numbers of neurons involved in specific behaviors. The data and the computational power are available; neurobiologists need to address the bigger issue of manipulating their data. A neurobiology Program of Excellence could bring together expertise to apply the latest data-management tools to the study of how the brain controls motor movements or how it forms memories.

MEDICAL GENETICS

Geneticists are running analyses of large numbers of subjects against the enormous amounts of data now being released about the human genome, utilizing the data from hundreds of subjects and their family members to map disease genes within a region of 30-40 megabases of DNA — more than 100 megabytes of information on each person. Those analyses can take as long as six months on routine laboratory computers. To gain the advantage of a two-day turnaround on a supercomputer, geneticists must adapt their programs to the more powerful systems. Good research should not be hurried, but delaying progress because software is not available could delay the discovery of new findings, new treatments and new cures.

CLINICAL TRIALS

Much of the information that comes out of clinical trials is statistical in nature. While some statisticians have been involved in helping to interpret those results, with the vast amounts of data now being generated, the issues are becoming more interesting to statisticians as data problems. The statistical community is only now beginning to realize that it may have much to contribute. A National Program directed towards the display and understanding of high-dimensional data from clinical trials would involve statisticians, physicians, and computer scientists in the attempt to deal with specialized data of enormous complexity and size.

Such a National Program would not be strictly computational. From the statistician's perspective, some problems that are labeled computational are really problems of the analysis of complex data. That analysis requires computational support, to be sure, but the challenge is to create appropriate analytical tools, whether algorithmic or programmatic. That is certainly the case with genetic-array data on tumor cells, or pattern-recognition problems in some image reconstruction — the kinds of problems that engage clinicians as well.

RATIONAL DRUG DESIGN AT THE CELLULAR LEVEL

Biological chemists attempting to model entire cells are waiting for the data to catch up to the technology. When the human genome has been fully sequenced, with all the genes identified, biological chemists hope they can test their theories of drug activities on computer models of cells. While researchers know a great deal about drugs that simply inhibit enzymes, they are much less sure about drugs that have subtle effects on cellular function. Researchers might possibly chart the effect of drugs on genes themselves when they can model an entire cell. Microarrays and complex genomic databases might be used to help biological chemists identify drug side effects with minimal human or animal testing. Sophisticated, linked databases of chemical substances, gene and protein structures, and reaction pathways and physiological effects will be required to make that kind of drug design possible. It is part of the idea behind National Programs of Excellence to find ways to coordinate those disparate kinds of data.

CELL BIOLOGY

Why do some cells die, and others grow uncontrolled? In cells, what is aging, and what is cancer? Cell biologists believe the answer lies in the way proteins assemble in the cell. There, function seems to follow form: The shape of proteins determines what they will do. The secret of protein assemblies seems to be in the ability of adjacent proteins to pass enough information to reach a corporate consensus on action. To correlate the arrangement of the proteins with their functions, researchers need high-resolution images of protein structures, and they need to compare structures across functions. That is not a trivial task. It takes hundreds of thousands, maybe millions of cross-sections of cell structures captured by microscopy (electron, light, MRI microscopy) to create a clear picture of the structure. That work is impossible without computational tools to collect, process and interpret the data to help understand how biological systems works. A National Program might give researchers the computational equivalent of heavy machinery that they need to plow into such data-intensive science. By bringing together the machinery, the people who know how to collect, curate, and manipulate that data; and the scientists who are familiar with cell biology, the NIH could move researchers forward in understanding the life cycle of the cell, and the diseases that affect it.

A COMMON FOUNDATION

Sequencing the genome, image reconstruction, the simulation of biological systems and other emerging areas have all led to increased opportunity for understanding biology while illuminating the alarming gap between the need for computation in biology and the skills and resources available to meet that need. Much of what needs to be done in this new approach to biology will have to be done by people who are currently either not drawn into biology, have little standing in biology, or whose career opportunities are better in industry or in other scholarly disciplines. The NIH should act to increase the number of people who are trained in both biology and computation, and dignify that expertise within the biomedical research community.

At the same time, the NIH needs to ensure that computer power is available. While most biomedical researchers have the desktop systems they need, they do not have up-to-date local clusters, they do not have sufficient access to regional computing centers, and they do not have a viable plan for using national computing centers — particularly those that promise teraflop computers by the next century.

Biomedical computing is on a cusp. Its growth is inevitable, but the timetable is still unknown. A small push by the NIH could result in great changes in a short time. If the NIH does not act, change could take another five, ten, or twenty years.

Workforce Development
From the Principal Investigators who understand how to use computers to solve biomedical problems to the people who keep the computers
running, there is a shortfall of trained, educated, competent people. The NIH needs a program of workforce development for biomedical computing that encompasses every level, from the technician to the Ph.D. The National Programs of Excellence in Biomedical Computing would provide a structure for developing expertise among biomedical researchers in using computational tools.

Today the disciplines of computer science and biology are often too far apart to help one another. A computer-science student often stops studying other sciences after freshman biology or chemistry; a biology student, even one knowledgeable about computers, may not ever have had formal computer-science classes. Biomedical computing needs a better and more attractive meld of those disciplines. Today computer-science students have little incentive to learn about biomedicine. The barrier is not just the rigorous demands of computer science; it is also the relative rewards: The $50,000 to $80,000 a year that professional programmers earn makes the compensation associated with many research positions in biology laughable. This situation is even more risible when one includes the reality that substantial positions on NIH research grants are guaranteed for no longer than the grant award.

In the future, many biomedical scientists will have to be well educated in both biology and computer science. One-sided education will not work. The Department of Biological Structure at the University of Washington offers one of the few programs in biomedical computing. The computer-science side incorporates programming, data structures, simple computer architecture, databases, computer networks, basic artificial intelligence, knowledge representation, and qualitative modeling. On the biology side, the program emphasizes basic medical science with courses such as anatomy, histology, cell biology, biochemistry or molecular structure. Other courses provide the quantitative basis for the broad spectrum of biology, from basic mathematics through calculus, differential equations, linear algebra, and statistics.

Such cross-discipline education should be supported by the NIH grant system. Awards should be competitive with those for computer-science and physics education. Establishing such programs will not alone create an academic infrastructure for biomedical computing; research grants are needed to make a fundamental difference in academia. Grants to faculty members are more likely to change the focus of a Ph.D. program than any change in the job market for graduates.

Strong action by the NIH is required because the existing biomedical research and teaching structures of the universities and research institutions of this country inadequately value interdisciplinary efforts generally, and computation in particular. Few grant programs and fewer academic departments foster the kind of interdisciplinary work required to address biomedical challenges fully, let alone educate students about them. National Programs of Excellence would specifically include formal and informal instruction from the undergraduate through post-graduate levels, and incorporate a range of opportunities for scholars and researchers to participate.

Software Development

Biomedical computing needs software tools to take advantage of the hardware. Often that software is cobbled together by graduate students with little programming knowledge, for use by those whose expectations are bound by the immediate problem. The application may be used once, then abandoned when the problem is solved, the graduate student moves on, or the technology changes. The publication goes out, but the tools remain in the laboratory.

That system worked for years only because computing had not yet become an important tool for biologists. Now that biomedical research is more dependent on computers, the discipline cannot afford to waste the effort to produce one-off software that is used once and discarded. Software can be shared if it is correctly conceived, properly developed, and effectively promulgated. Such a process offers two benefits: Needed software will be made available, and time spent reinventing the same processes in one laboratory after another will be freed for basic research.

One important element in the system is the creation of software-development groups: software and computer engineers who can take laboratory-based software and "harden" it-standardizing it for more general use, testing it under various conditions, documenting it, supporting it, and upgrading it as technology changes. Currently the NIH generally does not support such efforts; grants from the NIH are typically available only to develop a working model, a prototype. Completing that software and distributing it is not possible under today's funding programs. It is a generally accepted rule in the software business that producing a working prototype is only 20% of the cost of making a commercial product. NIH funding mechanisms finance only that first 20%. Where software has shown itself to be valuable to a range of researchers in biomedical computing, the NIH needs to find ways to support its full development. That might be done through public-private agreements between research centers and industry, or through direct NIH funding.

Algorithms

The need for numerical computation continues to challenge the most advanced computers, so the design and application of new algorithms continue to be of major importance. Good algorithms make computers more effective. Algorithms are the mathematical expression of information in a specialized environment. They are the bridge between data and understanding.

Discovering algorithms that advance scientific knowledge requires a thorough grounding in computer science and mathematics, as well as a keen understanding of the particular problem domain. In biology, algorithm development is now done only by the most knowledgeable computational biologists; a small fraction of the Ph.D.s in the field. Yet algorithms encapsulate the hypotheses that drive science, and their development should be an integral part of biomedical-computing research. More expertise is clearly needed as biological data increase and more computational power becomes available. To put complicated biological applications on tomorrow's teraflop machines will require teams of people working for several years. Without new algorithms and software, the power of such computers will be wasted, and mid-level machines will flounder in a sea of data. Algorithm development, the process by which researchers harness computing power, is as necessary in biomedical computing as computer power. The NIH should put resources into algorithm research if it is to advance biomedical research.

However, those with a bent for mathematics and computer science and the tenacity to seek a Ph.D. now see little reward in biomedical computing. There are few academic positions in that field; research grants tend to support the biological and not the computational aspects of their work; and their salaries are based on standards in biology, not computer science. A Ph.D. in computer science or mathematics carries more prestige, offers more job options, and guarantees more money than a Ph.D. in biology. If the NIH does not act to make biomedical research more attractive to those who are knowledgeable in computational areas, as biology increasingly becomes an information science, there will not be enough people who can create algorithms for biomedical research.

Databases

Biomedical computing is entering an age where creative exploration of huge amounts of data will lay the foundation of hypotheses. Much work must still be done to collect data and to create the tools to analyze it. Bioinformatics, which provides the tools to extract and combine knowledge from isolated data, gives us new ways to think about the vast amounts of information now available. It is changing the way biologists do science. Analyzing biological, physical, and chemical data is new—mathematical biology has done that for more than a century—but because the advent of extensive databases and the tools to manipulate them gives researchers the ability to tease knowledge about living systems from complex biological data using modern computational tools. In large part because of the tools of bioinformatics, biology is becoming a data-driven science.
mid-size facilities to meet their needs. The pool of researchers changed, too: An analysis of the projects shows a 40% turnover in users. Together those facts suggest that supercomputers are broadly needed — and used — across biological disciplines.

For most supercomputer users, the access to computing cycles is only one of the benefits provided by a supercomputer center. The strength of the National Science Foundation’s supercomputer centers is as much in their support staff as in their hardware, and in the collegial interactions among supercomputer users. The opportunity to discuss problems and solutions is an important part of the centers’ gestalt. Most biomedical-computing researchers who use supercomputers have no colleagues doing similar work at their own institutions; today it is only at the supercomputer centers that they find colleagues — many of them in fields like physics, chemistry, and mathematics — with whom they can discuss their approaches. (National Programs, as they are developed, will also offer opportunities for biomedical researchers to work alongside colleagues in computer-rich environments, building new communities around common interests.)

The current levels of computing bring a variety of computational-biology problems within reach. However, to systematically study those systems — to really explore phase space, to understand not only how it works, but how those systems can be manipulated — requires computation at adequate resolution for sufficiently long periods of time, and also requires large numbers of related computations. For the biomedical promise of computation to be realized, tera-scale computing must become routine.

As more powerful computing becomes routinely available, more researchers will use it because the increased computing power will open up opportunities that did not previously exist, and biomedical researchers will move to exploit those opportunities. For that reason, any attempt to predict future needs based on current usage will result in a substantial underestimate.

IMPLEMENTATION

Because of the importance of this initiative across the NIH, and because of the basic emphasis on scientific research as a means to train scientists across disciplines and provide the tools for their work in the 21st Century, funding for the four parts of the Biomedical Information Science and Technology Initiative might be shared among the Institutes. National Centers of Excellence in Biomedical Computing, in particular, are good candidates for shared funding. Their basic educational purpose should encourage institutions to provide support for National Programs associated with their campuses.

To help the reviewers and staff who will be awarding grants under this initiative, this Working Group suggests the following review criteria for National Programs of Excellence in Biomedical Computing:

- Value to the biomedical community: Will the programs provide significant advances in the selected areas of research? Will the research provide foundations or infrastructure for other research? Will the research advance human health directly or indirectly?
- Cross-disciplinary focus: While the National Programs are not required to be formally multidisciplinary, does the program take advantage of the conjoining of biology and computation?
- Research results: Does the Program incorporate both fundamental discovery and the development of useful tools? Is there a viable plan for developing, refining, and applying those tools that includes contributions from software engineers and computer scientists or other appropriate collaborators? Is there evidence of widespread usefulness of those tools, with publications and patents that document usage of or pressing need for those tools?
- Dissemination of software, hardware, algorithms, or databases: Is there a plan for making tangible, useful output available to other researchers?
- A new approach: Does the National Program bring in new ideas and new personnel and resources, or is it an aggregate of existing facilities?
- Fiscal responsibility: Especially in a virtual or cross-institutional program, is there a well-defined sharing of responsibilities among the institutions so that there is a clear principle under which to assign funds (and overhead) on an annual basis?
- Training plans: Is there a full menu of education, ranging from formal undergraduate and graduate programs to courses and seminars for students and working researchers? Are there visiting scientist programs, "total immersion" programs, one to two week accelerated-training programs or other innovative programs to help spread the knowledge? How many students, post-docs, and working researchers are trained, and what is their placement after that training?
- Success indicators: Does the National Program educate people and forge tools in the process of doing basic research?

CONCLUSION

The National Programs of Excellence in Biomedical Computing and the teams they bring together are important because biomedical computing needs cross-disciplinary expertise. The result of those Programs will be individuals with broad knowledge that can be applied to biomedical issues — knowledge that incorporates the strengths of biology, computer science, and mathematics. In the short term, biomedicine will benefit from the team approach. In the long term, there will be individual biomedical researchers who can apply much of the expertise that biomedical computing needs. The Biomedical Information Science and Technology Initiative (BISTI), and particularly its National Programs of Excellence in Biomedical Computing, is a bootstrapping approach to that next level.

The Initiative will presage smaller changes, as well. NIH study sections may come to expect that a fair proportion of biomedical research will need computational resources, and may even suggest that researchers include provision of those resources in their grant applications. In academia, there inevitably will be some restructuring of academic departments of biology and biomedicine, and tenure and promotion decisions at universities may depend as much on computational achievements as on traditional biomedical research. Both changes will improve biomedical research. Biomedical computing offers promise of profound advances in understanding and improving human health. Its advent is assured: Biomedical researchers are increasingly using computers to collect, store, access, and explore new data about the human condition, and that ripple of change will soon be a tidal wave. However, although it is inevitable, the promulgation of this critical enabling technology could face delays of five to ten years without action by the NIH. These recommendations are intended to shape the socio-technical aspects of biomedical computing to realize more quickly the anticipated benefits.

Advisory Committee to the Director, NIH Working Group on Biomedical Computing
http://www.nih.gov/about/director/060398.html


Co-Chairs:
David Botstein, Ph.D.
Professor and Chair
Department of Genetics
Stanford University Medical School
Stanford, CA 94305

Larry Smarr, Ph.D.
Director, National Center for Supercomputing Applications
University of Illinois
Champaign, IL 61820

Biomedical Instrumentation, Imaging
David A. Agard, Ph.D.
Professor
Department of Biochemistry and Biophysics
University of California at San Francisco
San Francisco, CA 94143

Molecular Modeling and Simulation
Michael Levitt, Ph.D.
Chairman
Department of Structural Biology
Stanford University Medical Center
Stanford, CA 94305

Clinical Trials
David Harrington, Ph.D.
Professor
Department of Biostatics
Dana-Farber Cancer Institute
Harvard University
Boston, MA 02115

Digital Library
David J. Lipman, M.D.
Director
National Center for Biotechnology Information
National Library of Medicine
National Institutes of Health
Bethesda, MD 20894

Neurosciences
Gwen Ann Jacobs, Ph.D.
Co-Director
Center for Computational Biology
Montana State University at Bozeman
Bozeman, MT 59715

Heart Modeling
Charles S. Peskin, Ph.D.
Professor
Department of Mathematics
Courant Institute of Mathematical Sciences
New York University
New York, NY 10012

Surgical Decision Support
Christopher R. Johnson, Ph.D.
Director
Center for Scientific Computing and Imaging
University of Utah
Salt Lake City, UT 84112

Proteins
George Rose, Ph.D.
Professor
Department of Biophysics and Biophysical Chemistry
Johns Hopkins University School of Medicine
Baltimore, MD 21205-2196

Industry
Arthur Levinson, Ph.D.
President and Chief Executive Officer
Genentech, Inc.
So. San Francisco, CA 94080-4990

Genomics
Gerald M. Rubin, Ph.D.
Department of Molecular and Cell Biology
University of California at Berkeley
Berkeley, CA 94720-3200

Algorithms
Hamilton O. Smith, M.D
Investigator
The Institute for Genomic Research
Rockville, MD 20850

Population Genetics
M. Anne Spence, Ph.D.
Professor, Genetics
Department of Pediatrics
University of California Irvine Medical Center
Orange, CA 92868-3298

Information-Based Biology [Bioinformatics]
Shankar Subramaniam, Ph.D.
Professor
Departments of Biochemistry, Biophysics and Physiology
University of Illinois at Urbana-Champaign
Senior Research Scientist
National Center for Supercomputing Applications
Urbana, IL 61801

Liaison Members of the Working Group:

http://www.nih.gov/about/director/050599.htm
Robert R. Borchers, Ph.D.
Director
Division of Advanced Computational Infrastructure and Research
National Science Foundation
Arlington, VA 22230

Alan S. Graeff
Chief Information Officer, NIH
Bethesda, MD 20892

Mary E. Clutter, Ph.D.
Assistant Director for Biological Sciences
National Science Foundation
Arlington, VA 22230

Michael L. Knotek, Ph.D.
Program Advisor for Science and Technology
Office of the Secretary of Energy
U.S. Department of Energy
Washington, DC 20585

John Toole
Deputy Director
National Center for Supercomputing Applications
University of Illinois
Champaign, IL 61820

Margaret L. Simmons, Ph.D.
Associate Director, Program Development
National Partnership for Advanced Computational Infrastructure
San Diego Supercomputer Center
University of California at San Diego
La Jolla, CA 92093

http://www.nih.gov/about/director/060369.htm
Statistics courses (one is required)

**Biostatistics and Medical Informatics 541 (3 credits)** - Introduction to Biostatistics - is designed for the biomedical researcher. Topics include: descriptive statistics, hypothesis testing, estimation, confidence intervals, t-tests, chi-squared tests, analysis of variance, linear regression, correlation, nonparametric tests, survival analysis and odds ratio. Biomedical applications are discussed for each topic. Prerequisites: Math 221 or equivalent or instructor's consent.

**Statistics 572 (4 credits)** - Statistical Methods for Bioscience II - is a design course aimed at CALS graduate students but the principles are quite applicable to molecular biology. Topics include: polynomial regression, multiple regression, two-way ANOVA with and without interaction, split-plot design, subsampling, analysis of covariance, elementary sampling and an introduction to the bioassay. Prerequisites: Statistics/Forestry/Horticulture 571.

Biostatistics and Medical Informatics courses (both are required)

**Biostatistics and Medical Informatics 576 (3 credits, cross-listed as Computer Sciences 576)** - Introduction to Bioinformatics - The goals of this course are to provide an understanding of the fundamental computational problems in molecular biology and a core set of widely used algorithms. This is the first of two courses on bioinformatics. The topics it will cover include: pairwise sequence alignment, multiple sequence alignment, finding genes in DNA sequences, phylogenetic tree construction, and genome mapping and sequencing. This is currently being taught as a special topics course in Computer Sciences. Prerequisites: Math 222 and Computer Sciences 367.

**Biostatistics and Medical Informatics 776 (3 credits, cross-listed as Computer Sciences 776)** - Advanced Bioinformatics - The goals of this course are to provide an understanding of the fundamental computational problems in molecular biology and a core set of widely used algorithms. This is the second of two courses on bioinformatics. The topics it will cover include: probabilistic methods for sequence modeling, gene expression analysis, phylogenetic tree construction, protein structure prediction, RNA modeling, whole-genome analysis, and algorithms for exploiting biomedical text sources. Prerequisites: Biostatistics and Medical Informatics 576.

Elective Courses (choose one) - Additional elective courses are expected to be added as faculty are recruited.

**Biochemistry 711/712 (3 credits for 711, 1 credit for 712)** - Sequence Analysis - This is a two-part course beginning with a lecture/discussion group course (711) and finishing with a hands-on laboratory course, taught at actual computer terminals, designed to complement and reinforce the sequence analysis concepts presented in Biochemistry 711. This course give students a practical background in using many available software packages such as DNA-STAR for gene sequencing, etc. Students are
provided with actual data and gain experience using this software. Prerequisites: Graduate level standing.

**Biostatistics and Medical Informatics 542 (3 credits) - Fundamentals of Clinical Trials** - Intended for biomedical researchers interested in the design and analysis of clinical trials. Topics include definition of hypotheses, measures of effectiveness, sample size, randomization, data collection and monitoring, and issues in statistical analysis. Prerequisites: Statistics 541 or equivalent or instructor's consent.

**Bioinformatics Independent Study 799 (3 credits)** - Some students may find their needs are better met by an independent study with one of the faculty in the department, in collaboration with a biological faculty member. Independent study in another department may be substituted with prior approval.

**Computer Science 540 (3-4 credits) - Introduction to Artificial Intelligence** - Teaches principles of knowledge-based search techniques; automatic deduction; knowledge representation using predicate logic, semantic networks, connectionist networks, rules, machine learning, applications in problem solving, expert systems, game playing, natural language understanding. Prerequisites: Computer Sciences 367.

**Computer Science 545 (3 credits) - Natural Language and the Computer** - The course covers basic techniques and tools in natural language processing: generative grammars, parsing, dictionary construction, semantic networks, generation of text from a knowledge base, natural language interfaces, and machine translation. Prerequisites: CS 536 or CS 537 or 564 or consent of instructor.

**Computer Science 564 (3 credits) - Database Management Systems: Design and Implementation** - What a database management system is; different data models currently used to structure the logical view of the database: relational, hierarchical, and network. Hands-on experience with relational and network-based database systems. Implementation techniques for database systems. File organization, query processing, concurrency control, rollback and recovery, integrity and consistency, and view implementation. Prerequisites: CS 367 and 354.

**Computer Science 577 (3 credits) - Introduction to Algorithms** - Survey of important and useful algorithms for sorting, searching, pattern-matching, graph manipulation, geometry, and cryptography. Paradigms for algorithm design. Techniques for efficient implementation. Prerequisites: CS 367, and CS 240, or consent of instructor.

**Computer Science 731 (3 credits) - Advanced Artificial Intelligence** - Novel techniques within Bayesian Networks, Machine Learning and Data Mining, Planning and Computer Vision have proven useful for many real-world problems. This course will cover some of the most important recent algorithms from these areas and will illustrate their use with biomedical applications. Prerequisites: Computer Sciences 540.

**Computer Science 760 (3 credits) - Machine Learning** - The intent of this course is to present a broad introduction to machine learning, including discussions of each of the major approaches currently being investigated. Class lectures will discuss general issues in machine learning, as well as present established
algorithms. Computational approaches to learning, including: inductive inference, explanation-based learning, analogical learning, connectionism, and formal models, what it means to learn, algorithms for learning, comparison and evaluation of learning algorithms, cognitive modeling and relevant psychological results. Prerequisites: Computer Sciences 540.

**Computer Science 766 (3 credits) - Computer Vision** - an introductory course to the basic concepts in computer vision including fundamentals of image analysis and computer vision, image acquisition and geometry, image enhancement, recovery of physical scene characteristics, shape-form techniques, segmentation and perceptual organization, representation and description of two-dimensional objects, shape analysis, texture analysis, goal-directed and model-based systems, parallel algorithms and special purpose architectures. Prerequisites: Computer Sciences 540.

**Industrial Engineering 617 (3 credits) - Health Information Systems (previously offered as IE 691)** - Introduction to health information systems and health informatics. Major topics include clinical information systems, formal language and vocabularies, telemedicine, image technology and public health informatics. Lectures by local and national experts will be followed by instructor-facilitated discussion examining how industrial engineering tools and perspectives could improve the quality, efficiency and effectiveness of health information. Prerequisites: Senior or Graduate Standing, or instructor's consent.

**Mathematics 605 (3 credits) - Stochastic Methods for Biology** - This course is intended to provide a rigorous foundation for stochastic modeling of biological systems. The mathematical emphasis is in stochastic analysis and simulation. Biological applications include epidemiological phenomena, biochemical reaction networks and population dynamics. Prerequisites: Math/Stat 431, Math/Stat 309 or Stat 311, or consent of instructor.

**Mathematics 606 (3 credits) - Mathematical Methods for Structural Biology** - This course will provide a rigorous foundation for mathematical modeling of biological structures. Mathematical techniques include ordinary and partial differential equations, 3D Fourier analysis and optimization. Biological applications include protein folding, molecular dynamics, implicit solvent electrostatics, and molecular interactions. Prerequisites: Math 340 or 341, CS 302, or consent of instructor.

**Mathematics 608 (3 credits) - Mathematical Methods for Continuum Modeling in Biology** - This course is intended to provide a rigorous foundation for mathematical modeling of biological systems. The mathematical emphasis is on partial differential equations, particularly reaction-diffusion and transport equations. Biological applications include bacterial chemotaxis, spatio-temporal ecological dynamics, and cell-level reactions. Prerequisites: Math 322, Math 415, Math 514, or consent of instructor.

**Mathematics 609 (3 credits) - Mathematical Methods for Systems Biology** - This course is intended to provide a rigorous foundation for mathematical modeling of biological systems. Mathematical techniques include dynamical systems and differential equations. Applications to biological pathways, including understanding of bistability within chemical reaction systems, are emphasized. Prerequisites: Math 340 or 341, Math 415 or consent of instructor.