CISC 889 Bioinformatics (Spring 2004)
Lecture 1

Course Overview

Li Liao
Computer and Information Sciences
University of Delaware

Administrative stuff

- Syllabus and tentative schedule (check frequently for update)
- Office hours
- Collect student info (name, email, background:Language/AI-Machine Learning/Prob-Stat, why taking this course?)
- Each student introduce him/herself
- Introduce textbook and other resources
  - URLs, PDF/PS files, or hardcopy handout
  - A reading list
- Workload
  - 4 homework assignments (hands-on to learn the nuts and bolts)
    - Language issue: Perl is strongly recommended (A tutorial is provided)
  - 3 quizzes (terminologies, concepts, ...)
  - projects (gain in-depth knowledge on selected topics, prepare for research, talking points for job interview)
- Late policy: no makeup quizzes; 15% off per class up to three class mtgs.

Books

- **R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press, 1998.**
- Dan Gusfield. Algorithms on String, Trees, and Sequences. Cambridge University Press, 1997.
- João Meidanis & João Carlos Setubal. Introduction to Computational Molecular Biology. PWS Publishing Company, Boston, 1996.
- W.J. Ewens and G.R. Grant, Statistical Methods in Bioinformatics, Springer 2001.
- D.W. Mount, Bioinformaics: Sequence and Genome Analysis, CSHLP 2001.
- P. Baldi and S. Brunak, Bioinformatics, The Machine Learning Approach, The MIT press, 1998.

Bioinformatics: Bring Biology and Computing Together

Where do biology and computing meet?

Biology >> Computing

- Genetic algorithms
- Artificial neural network
- DNA computing
  - o Seminal work by Leonard Adelman [Science 266(1994)1021]

Biology << Computing

- String matching algorithms --> DNA/protein sequence similarity, genome sequencing
- Tree algorithms --> phylogeny of species
- Hidden Markov models -> Protein families
- Stochastic context-free grammar -> RNA sequence similarity
- Graph theoretic algorithms --> Protein families, genetic networks
- Pattern discovery algorithms --> motifs

Molecular Biology

Free materials:
- Kimball's biology
- Lawrence Hunter: Molecular biology for computer scientists
- DOE's Molecular Genetics Primer

Books:
- Instant Notes series: Biochemistry, Molecular Biology, and Genetics

What is life?
- 2nd law of thermodynamics: low entropy
- 7 characteristics
- living organisms:
  - eukaryotes - fungi, plants, and animals
  - bacteria, and
  - archea
- model organisms:
  - E.coli, C. elegans (Worm), S. Cerevisiae (Yeast), Drosophila (fruitfly), Arabidopsis (Weed), Mouse, Homo sapien (human)
- The tree of life: an evolutionary view

*** Exercise: Find info about these genomes: size, # of genes, who sequenced, and the database

CISC 889, S04, Lec1, Liao

Molecular Biology

The cell: structure, function/differentiation, and division.
- Genetic materials and the central dogma
- DNA, (4 nucleotides: A, C, T, G),
  - double helix structure, hybridization
- RNA,
  - Self-hybridization, palindromic structure
- Proteins (20 Amino acids: three letter code and one letter code)
  - Workhorse

*** Exercise: Memorize single letter code for the 20 amino acids

Genetic code: codon usage

Transcription and translation

Regulation/Metabolism

Wet Lab Technologies
- PCR
- Clone
- Microarray
- X-ray, NMR, and MS spectrometry

CISC 889, S04, Lec1, Liao

Challenges in Life Sciences

- Understanding correlation between genotype and phenotype
- Predicting genotype <=> phenotype
- Phenotypes:
  - drug/therapy response
  - drug-drug interactions for expression
  - drug mechanism
  - interacting pathways of metabolism

CISC 889, S04, Lec1, Liao

Paradigm shifted

- Traditional biology:
  - Make a hypothesis
  - Design experiment to test it
  - Run experiment
  - Conflict?
  - Revise hypothesis
- Data-driven paradigm:
  - Many labs conduct different genome-wide experiments.
  - Each experiment provides partial overview.
  - Resulting data is explored using computational techniques.
- Data mining

CISC 889, S04, Lec1, Liao

Topics

- **Mapping and assembly**
- **Sequence analysis (**Similarity -> Homology**):**
  - **Pairwise alignment (database searching)**
  - **Multiple sequence alignment**
  - **Gene prediction**
  - **Pattern (Motif) discovery and recognition**
- **Phylogenetics analysis**
  - **Distance based**
  - **Character based**
  - **Probabilistic**
- **Network analysis:**
  - **Metabolic pathways reconstruction**
  - **Regulatory networks (<-- Gene expression)**
- **Structure prediction (secondary & tertiary)**
  - **Comparative, Threading, Ab initio**

CISC 889, S04, Lec1, Liao

Algorithms

Design paradigms
- dynamic programming
- string and graphs
- approximate algorithms
- parallel

Applications
- Smith-Waterman
- Needleman-Wunsch
- BLAST and FASTA
- CLUSTALW
- MSA
- Consecutive one
- Suffix trees
- Vertibi algorithm (HMMs)
- Baum-Welch
- Nussinov
- Cocke-Younger-Kasami

CISC 889, S04, Lec1, Liao

Probability and Statistics

Chapter 11 in the text
- will cover relevant topics where they are needed

Basic concepts and formulas
- Poisson
- Gaussian
- Binomial
- Dirichlet
- Extreme limit
- Central limit

Basic techniques/models
- Markov chain
- Markov Chain Monte Carlo (MCMC)
- Gibbs sampling
- maximum likelihood
- Bootstrap
- hypothesis test
- hidden Markov model
- Bayesian

CISC 889, S04, Lec1, Liao

IT skills

- **Working with unix system/Web server**
- **Programming (in PERL, Java, etc.)**
- **RDBMS: SQL, Oracle PL/SQL**

CISC 889, S04, Lec1, Liao

People

- International Society for Computational Biology (www.iscb.org) ~ 1000 members
- Severe shortage for qualified bioinformatians

CISC 889, S04, Lec1, Liao

Conferences

- ISMB (Intelligent Systems for Molecular Biology) started in 1992
- RECOMB (International Conference on Computational Molecular Biology) started in 1997
- PSB (Pacific Symposium on Biocomputing) started 1996
- TIGR Computational genomic, started in 1997
- ...

CISC 889, S04, Lec1, Liao

Journals

- Bioinformatics
- Journal of Computational Biology
- Genomics
- Genome Research
- Nucleic Acids Research
- ...

Current hot areas

- Protein function identification
- Microarray analysis (clustering, classification, combination)
- Data integration (Gene ontology, database interoperation)
- Natural Language Processing
- Inference and Simulation of genetic/metabolic networks

Industrial

- Oscar Gruss $ Son estimates bioinformatics be a $2 billion business in 5 yrs
- IBM:
  - BlueGene, the fastest computer with 1 million CPU
  - Blueprint worldwide collects all the protein information
  - Bioinformatics segment will be $40 billion in 2004 up from $22 billion in 2000
- SmithKline
- Celera
- …

New Sites

news

Nature **410**, 293 (2001); doi:10.1038/35066710

Singapore invests in bioinformatics

DAVID CYRANOSKI

[TOKYO] Plans for a new institute in Singapore could address one of the most acute skills shortages in science by producing up to 100 tr

The planned Bioinformatics Institute is part of Singapore's US$1 billion-a-year effort to turn the island into a powerhouse of biomedical research. Within five years, the institute should be delivering 100 masters degrees in bioinformatics. This is more than any other institution in the world, says Limsoon Wong, director of the Kent Ridge Digital Bioinformatics Laboratories, and one of the planners behind the institute.

According to Wong, training at the institute will go well beyond the curating of data. "We will be training people how to make predictions from the data concerning interaction between proteins, and how to use these data to drive experiments," he says.

The research and teaching institute will be housed temporarily at first, before moving to the planned 'biopolis' science park near the National University of Singapore, when it opens in two years time. The government has yet to announce its funding level, but the park is expected to start with a grant of around S$100 million (US$60 million).

The institute is likely to absorb the existing bioinformatics centre at the National University and to help service the nation's expanding genomics programme. Its own research programme will follow from the interests of the staff who will be recruited internationally.

Guna
direct
bioinf

Gunaretnam Rajagopal, a theoretical physicist at Cambridge University, has been named as the institute's deputy director, and starts wo
to build a world-class research organization, with strong encouragement, commitment and active support of the government of Singapor

naturejobs careers
Naturejobs **413**, 25 October 2001, 8 (2001); doi:10.1038/35101788

Science & Technology Networks in Scandinavia
December 12th - *Nature* supplement

Playing catch-up

ROBERT TRIENDL

Robert Triendl is a freelance writer based in Tokyo.

Japan's government is belatedly realizing that it needs to increase funding for training in bioinformatics, says Robert Trien field could hinder the country's efforts.

Well-trained bioinformatics specialists in Japan are not just rare — they are virtually non-existent. This is partly because of a lack of formal education in the subject, and the problem is systemic. With little formal recognition of bioinformatics as a field, graduate departments have until recently allocated only a limited number of students to existing bioinformatics teachers. The government recognized the need for more bioinformaticians as it scaled up the country's genomics efforts.

This year, Japan's Ministry of Education, Science, Sports and Culture started to upgrade bioinformatics education at national universities by creating additional staff positions and funding both undergraduate courses and graduate-level informatics training.

Kyoto University's new bioinformatics centre is a product of this new policy. The university is Japan's leading academic centre for bioinformatics research, but until a few months ago all its bioinformatics activities were concentrated in just one laboratory: Minoru Kanehisa's lab at the Institute for Chemical Research.

Tokyo University also plans to increase its bioinformatics education and training activities. And the private Keio University has set up a whole new campus focusing on systems biology and the dynamic modelling of biological systems such as human blood cells.

Part of the ministry's promotion and coordination fund, the programme will provide between US$1 million and US$2 million in additional funding over several years for undergraduate and graduate education in bioinformatics, systems biology, protein functional analysis and software development.

CBRC

Rese
Biolo
joinin
bioinf
towar

biotechnology | ARCHIVE | SEARCH | INFORMATION | CLASSIFIED | SUBSCRIBE

Naturejobs Biotechnology

MAR 2001

FULLTEXT PDF

FIGURES / TABLES

Careers and Recruitment

March 2001 Volume 19 Number 3 pp 285 - 286

University bioinformatics programs on the rise

Randy J. Zauhar

Randy J. Zauhar is associate professor of biochemistry and director of the graduate program in bioinformatics, Department of Chemistry and Biochemistry, University of the Sciences in Philadelphia, 600 S. 43rd Street, Philadelphia, PA 19104-4495 (e-mail: r.zauhar@usip.edu).

Fueled by strong demand from students and industry's need for trained bioinformaticists, universities are increasing their offerings in this fast-growing field.

Information is found in the arrangement of things with respect to each other. Whether we observe the arrangement of ink on the page to form words, or the order of nucleotide bases to form a gene sequence, we recognize that information is the key to understanding the world around us. Bioinformatics studies how information is stored, reproduced, and used by living systems. It is not an overstatement to say that bioinformatics is what biology is evolving to become in the 21st century.

The subject matter of bioinformatics has existed as long as there has been molecular biology, but it has emerged as a distinct discipline only in the last two decades. This was prompted by a veritable explosion of information triggered by

careers and recruitment

Nature **389**, 420 - 422 (1997) © Macmillan Publishers Ltd.

Running to catch up in Europe

HELEN GAVAGHAN

Helen Gavaghan is a science and technology writer based in Hebden Bridge, Yorkshire, UK.

Across Europe, the story is the same. Demand for those skilled in bioinformatics exceeds supply. Like biochemistry and biophysics befo the barriers between traditional academic fields, and demanding flexibility and a new way of thinking from its adherents.

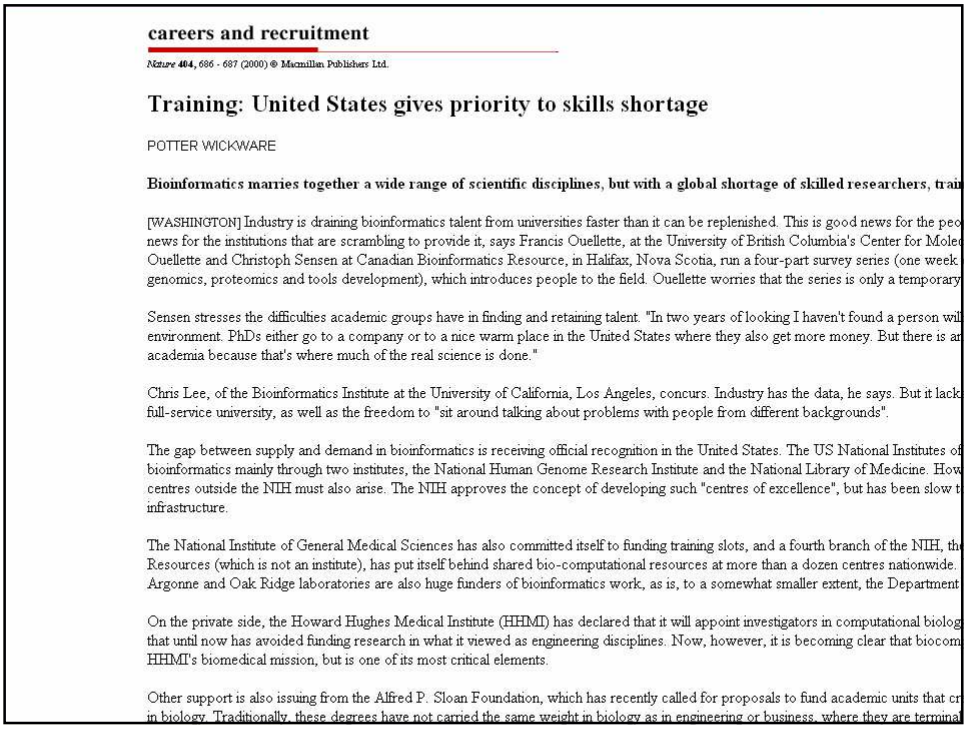
Computational biology has meant different things to different people. Not too long ago, says Hans Prydz of the University of Oslo's Biot handling NMR data or analysing Doppler echograms. Now renamed bioinformatics, it means looking for patterns in DNA and RNA, pr modelling proteins and mining massive databases that continue to grow. When the DNA database run by the European Bioinformatics In contained 700,000 nucleotides: now there are more than a billion.

Driven by the scientific and commercial importance of bioinformatics in genomics and drug discovery and development, governments, ur responding with varying degrees of vigour and success to the skills shortage and are seeking ways to cross the boundaries between disc physics, mathematics, computer science, statistics, protein chemistry, genetics and molecular biology.

At European level, the EBI, based near Cambridge (United Kingdom), is funded to the sum of about DM 9 million ($5 million) by memb Israel via their contributions to the European Molecular Biology Laboratory (EMBL) in Heidelberg, Germany. Contributions from the pl industries roughly double the institute's income. The EBI, an offshoot of EMBL, develops tools for bioinformatics, seeks innovative ways training courses for academics and industrialists. Initiatives with industry include the Industry Affiliates Initiative, which helps small and m and apply new techniques; the BioTitan Project, running nodes to enable faster access to databases; and the Biostandards project, funde European Union for promoting and developing standards.

National initiatives also exist, particularly in the United Kingdom and Germany. Says Andrew Lyall, responsible for bioinformatics at Gla in pretty good shape." There are two government-financed initiatives in the United Kingdom, both of which received a second lease of li

One of these schemes, supported by the Biotechnology and Biological Sciences Research Council (BBSRC), coordinates the UK bioin the scheme has concentrated on developing software that would enable biologists without information technology (IT) skills to use some their trade that are found on the World Wide Web. At a meeting earlier this month, the steering committee of the scheme decided to cha Brass, who runs a masters' degree course in bioinformatics at the University of Manchester and is a member of the committee, says, "W



careers and recruitment

Nature **404**, 686 - 687 (2000) © Macmillan Publishers Ltd.

Training: United States gives priority to skills shortage

POTTER WICKWARE

Bioinformatics marries together a wide range of scientific disciplines, but with a global shortage of skilled researchers, trai

[WASHINGTON] Industry is draining bioinformatics talent from universities faster than it can be replenished. This is good news for the peo news for the institutions that are scrambling to provide it, says Francis Ouellette, at the University of British Columbia's Center for Molec Ouellette and Christoph Sensen at Canadian Bioinformatics Resource, in Halifax, Nova Scotia, run a four-part survey series (one week genomics, proteomics and tools development), which introduces people to the field. Ouellette worries that the series is only a temporary

Sensen stresses the difficulties academic groups have in finding and retaining talent. "In two years of looking I haven't found a person wil environment. PhDs either go to a company or to a nice warm place in the United States where they also get more money. But there is ar academia because that's where much of the real science is done."

Chris Lee, of the Bioinformatics Institute at the University of California, Los Angeles, concurs. Industry has the data, he says. But it lack full-service university, as well as the freedom to "sit around talking about problems with people from different backgrounds".

The gap between supply and demand in bioinformatics is receiving official recognition in the United States. The US National Institutes of bioinformatics mainly through two institutes, the National Human Genome Research Institute and the National Library of Medicine. How centres outside the NIH must also arise. The NIH approves the concept of developing such "centres of excellence", but has been slow t infrastructure.

The National Institute of General Medical Sciences has also committed itself to funding training slots, and a fourth branch of the NIH, th Resources (which is not an institute), has put itself behind shared bio-computational resources at more than a dozen centres nationwide. Argonne and Oak Ridge laboratories are also huge funders of bioinformatics work, as is, to a somewhat smaller extent, the Department

On the private side, the Howard Hughes Medical Institute (HHMI) has declared that it will appoint investigators in computational biolog that until now has avoided funding research in what it viewed as engineering disciplines. Now, however, it is becoming clear that biocom HHMI's biomedical mission, but is one of its most critical elements.

Other support is also issuing from the Alfred P. Sloan Foundation, which has recently called for proposals to fund academic units that cr in biology. Traditionally, these degrees have not carried the same weight in biology as in engineering or business, where they are terminal

careers and recruitment

Nature 409, 963 (2001) © Macmillan Publishers Ltd.

Who makes the best bioinformaticians?

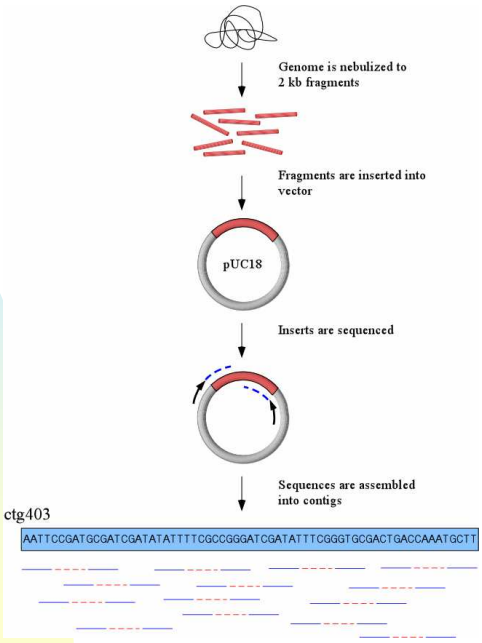
PAUL SMAGLIK

Paul Smaglik is editor of *Naturejobs*.

Bioinformatics careers can be divided into two paths: developing software, and using it. The field, catalysed by the rapid accumulation of genomic d attention as a salvation for jobs in biology. But that sentiment may not provide an accurate assessment of job opportunities, at least for career prosp For example, InforMax, one of the largest bioinformatics companies in the United States, generally doesn't hire biologists-turned-programmers, say chairman and chief executive officer of the company, based in North Bethesda, Maryland.

InforMax has about 95 programmers, almost all of whom come from a maths, physics or computer-science background. Titomirov says it is "much people with those skills about biology than to teach biologists how to code well. However, as the company turns to developing software to handle f and protein data, it may draw on more biologists to help design new software modules.

It is “much easier” to teach people with those skills about biology than to teach biologists how to code well.



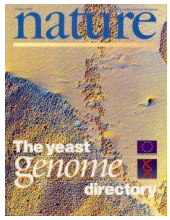
CISC 889, S04, Lec1, Liao

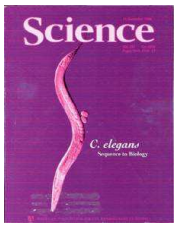
1995
Bacteria, 1.6 Mb,
~1600 genes
[Science 269: 496]

1997
Eukaryote, 13 Mb,
~6K genes
[Nature 387: 1]

1998
Animal, ~100 Mb,
~20K genes
[Science 282: 1945]

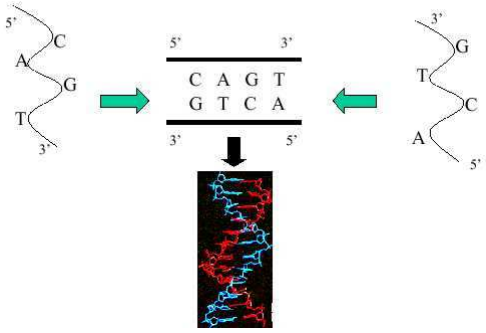
2001
Human, ~3 Gb,
~30K genes
[Science 291: 1304]
[Nature 409: 860]



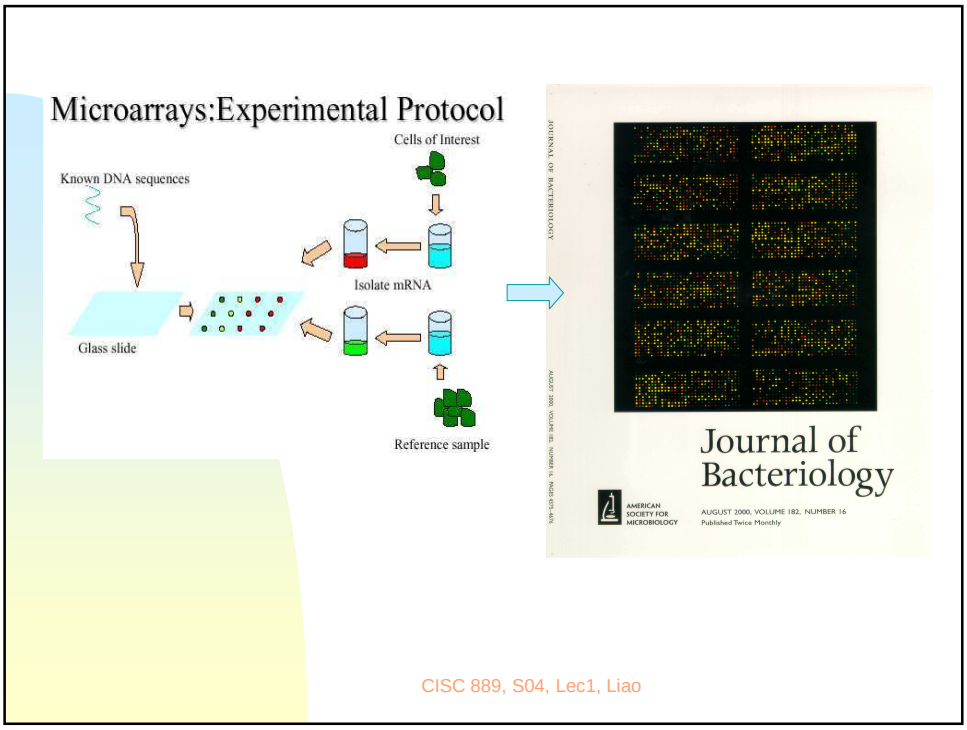


CISC 889, S04, Lec1, Liao

Microarrays: DNA Base Pairing

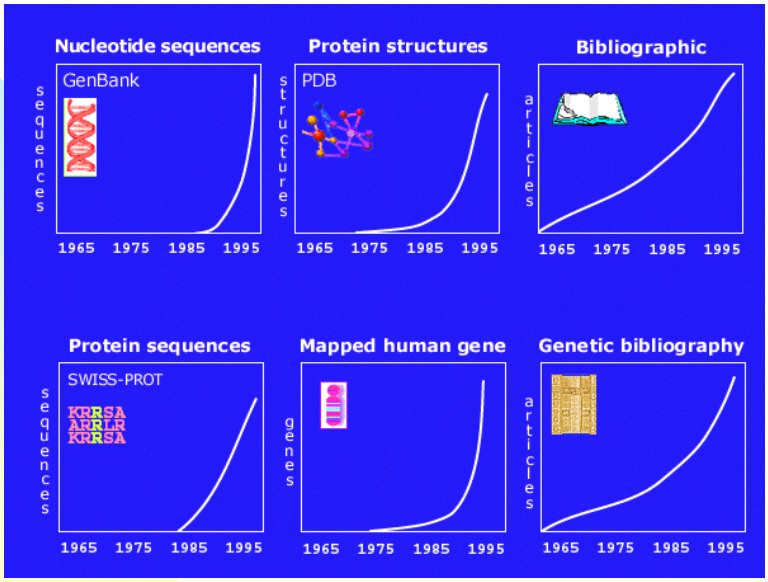


CISC 889, S04, Lec1, Liao

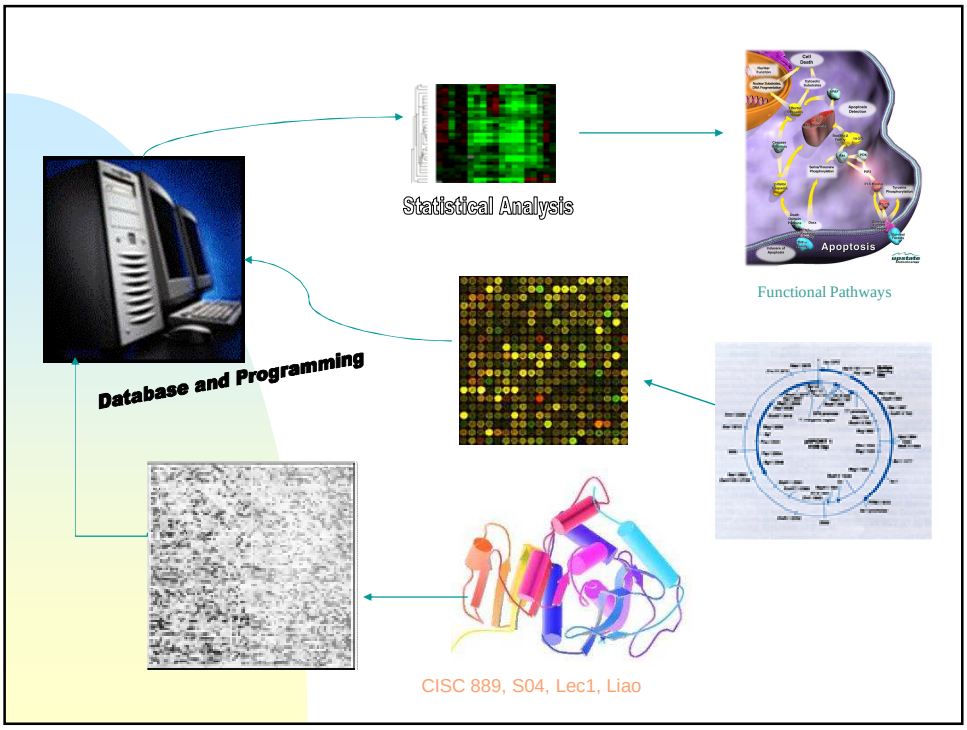


CISC 889, S04, Lec1, Liao

Information Explosion



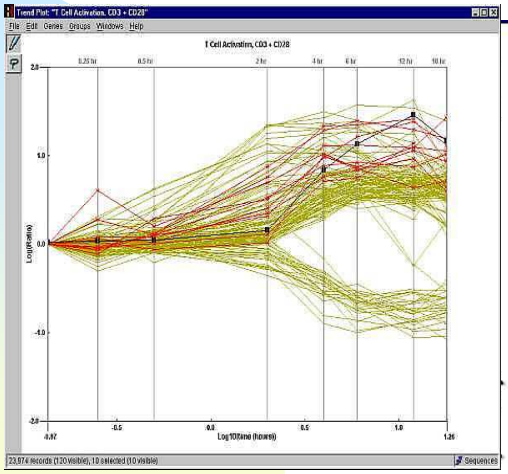
CISC 889, S04, Lec1, Liao

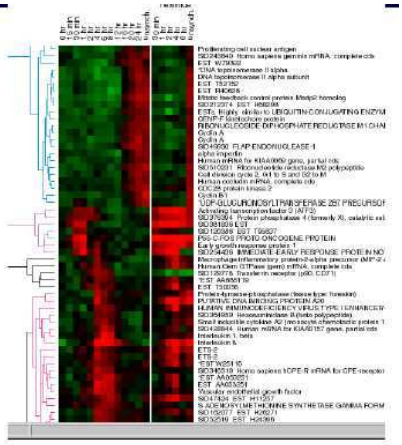


- **Genomics (structural, functional, comparative, pharmaco-, ...)**
- **Proteomics**
- **Metablomics**
- *****omics**
- **Biocomputing?**

CISC 889, S04, Lec1, Liao

Gene Expression Data Analysis



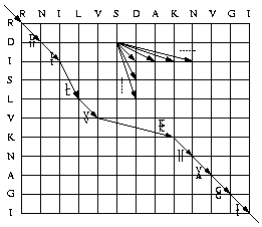


- A Central Theme:
  Sequences -> Functions
  - Similarity -> Homology
  - Beyond similarity
    - Structure (PDB, ab initio)
    - Phylogenetic Profiling
    - Gene expression profiling

- **Bioinformatics resources**
  - **Databases**
    - **Genbank**
    - **SwissProt**
    - **PDB**
    - **…**
  - **Tools**
    - pair-wise (FASTA, BLAST)
    - multiple sequence alignment (CLUSTAL W)
    - protein family database (Pfam)
    - patterns or motifs (Pratt, E-Motif, E-Matrix, MEME, SignalP,Psort, TopPred)
    - Gene identification (GeneScan, GeneMark, Glimmer)
    - Phylogenetic (Phylip, PAUP)
    - … CISC 889, S04, Lec1, Liao

Pair-wise alignment

$$score\begin{pmatrix} RDISLV---KNAGI \\ RNI-LVSDAKNVGI \end{pmatrix} = 6+2+5-g(1)+6+4-g(3)+5+2+6+5+5$$



CISC 889, S04, Lec1, Liao

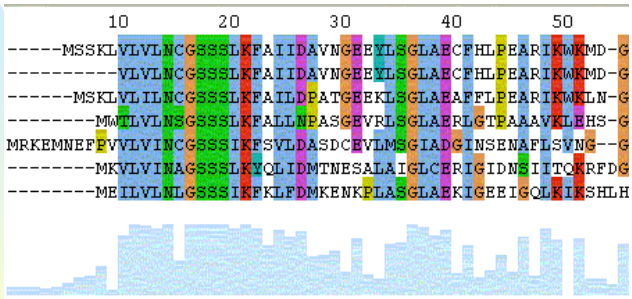
Substitution Matrices

- Importance of scoring matrices
  - Scoring matrices appear in all analysis involving sequence comparison. The choice of matrix can strongly influence the outcome of the analysis.
  - Scoring matrices implicitly represent a particular theory of evolution.
  - Understanding theories underlying a given scoring matrix can aid in making proper choice.
- Similarity vs. Distance
  - Elements of the matrices specify the weight to assign a given comparison.
- Types of scoring matrices
  - Identity
  - physical/chemical similarity (e.g., hydrophobicity)
  - Transition probability matrix (PAM, Blosum)

Smith-Waterman algorithm

- Global similarity:
  - Needleman-Wunsch algorithm o(nm)
- Local similarity
  - $o(n^3 m^3)$?
  - Dynamic programming o(nm)

Multiple Sequence Alignment



CISC 889, S04, Lec1, Liao

Multiple Sequence Alignment

- Motivation
  - a multiple alignment gives a stronger signal for sequence similarities than a pairwise alignment, i.e. in
    - sequence assembly
    - molecular modeling
    - database search
    - primer design
  - ambiguities in the alignment may be resolved by a multiple alignment
  - looking into the relationship between the sequences leads to reconstructing a phylogenetic tree (see next section)

CISC 889, S04, Lec1, Liao

Multiple Sequence Alignment C'td

- Scoring
  - No good scoring method
  - A popular one is the sum-of-pairs (**SP**) function, which is defined as the sum of all pairwise scores between all pairs of letters in the columns of the multiple alignment
- Task
  - Given the sequences s1, s2,...,sn, find an optimal alignment A_{opt} which maximizes the alignment-score. Note that there may be several optimal alignments for a set of sequences, while, of course, the optimal alignment score is unique.

CISC 889, S04, Lec1, Liao

Multiple Sequence Alignment (C'td)

- Exact solutions
  - It was shown that finding an optimal multiple alignment with the SP measure is **NP-hard** with regard to the number of sequences Win a million dollars from (www.claymath.org).
- Approximation
  - progressive, e.g., Clustal W, and other heuristics to reduce searching space.
  - an alternative is HMMs

CISC 889, S04, Lec1, Liao

Pattern-based sequence analysis

- Motivation
  - Biological observation: There exist common sequence patterns or motifs in groups of functionally related proteins. Such patterns may be the result of common ancestry combined with conservative evolutionary pressure to maintain important residues at active sites and other functionally important parts of the protein
  - more sensitive (in contrast to sequence similarity)

CISC 889, S04, Lec1, Liao

- Tasks
  - Family characterization problem (Pattern discovery)
  - Family classification problem (Pattern recognition)

CISC 889, S04, Lec1, Liao

- Types of Patterns
  - substring patterns (like **TATAA**)
  - regular patterns (like the ones used in PROSITE database)
  - regular patterns plus simple repetitions
  - Caveat: Most pattern discovery use exact matching b/w patterns and sequences. An alternative to patterns is profiles or hidden Markov models, NN, SCFGA.
- PROSITE patterns
  - An example: **W-x(2)-[LI]-x(4,5)-R**

CISC 889, S04, Lec1, Liao

- Biological application
  - Gene functional class prediction
  - RNA splicing
  - Protein structure and function
  - Gene regulation

CISC 889, S04, Lec1, Liao

- Evaluation of performance
  - **False negatives:** sequences in the family that do not match the pattern
  - **False positives** sequences outside the family that do match the pattern
  - sensitivity vs. specificity (Ockham's razor)
  - statistical significance v.s. biological significance

- Evaluation of Patterns: To rank the patterns that brings us the most information about the sequences in S or that are the least likely (probable) to be conserved in S by chance.
  - **Minimum Description Length** (MDL) principle (from Machine Learning)
    - helps to penalize patterns over-fitting the example sequences.
  - **Statistical Significance** (See a good discussion in Califano et al's paper)
  - **Biological significance** is hypothesized from statistical significance, but not guaranteed.
    - can be tested using experimental techniques, e.g., site-directed mutagenesis.

Approaches to pattern discovery

- **Pattern driven:**
  - enumerate all (or some) patterns up to certain complexity (length),
    for each calculate the score, and report the best.
    E.g., MOTIF, ASSET, PRATT
- **Sequence driven:**
  - look for patterns by aligning the given sequences
    e.g., MacAW
- **Combined**
  - find candidate patterns by a pattern driven algorithm and combine
    or expand them in a sequence driven manner.

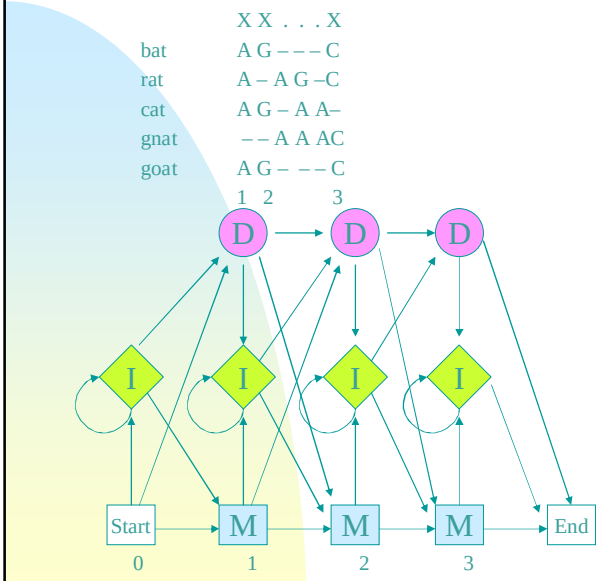
CISC 889, S04, Lec1, Liao

Protein Families

- Motivation
  - enhance signal/noise ratio for function prediction
  - biologically meaningful
- Approaches to protein family construction
  - clustering, e.g., K-means, Graph theory, pattern-based, etc.
  - supervised learning, e.g., HMMs, NN, SVM, etc.
- Types of Protein families
  - pattern-based e.g., PROSITE, e-Motif
  - Profiles, e.g., PRINTS, Blocks
  - Hidden Markov Models (HMMs) e.g., Pfam
  - Sequence Similarity, e.g, PIR, ProDom, COGs

CISC 889, S04, Lec1, Liao

Hidden Markov Model



Observed emission/transition counts

node position	0	1	2	3
A	–	4	0	0
C	–	0	0	4
G	–	0	3	0
T	–	0	0	0
A	0	0	6	0
C	0	0	0	0
G	0	0	1	0
T	0	0	0	0
MM	4	3	2	4
MD	1	1	0	0
MI	0	0	1	0
IM	0	0	2	0
ID	0	0	1	0
II	0	0	4	0
DM	–	0	0	1
DD	–	1	0	0
DI	–	0	2	0

CISC 889, S04, Lec1, Liao

Phylogenetic analysis

- **Phylogenetic tree** is a tree to represent the evolutionary relationship among species.
- **Task**

  Given a set of protein sequences, to infer their position during the evolution derived from a common ancestor. The evolutionary process is represented as a tree with proteins on the leaves and hypothetical ancestors on the nodes.

  - Caveat: the phylogenetic tree of a group of sequences does not necessarily reflect the phylogenetic tree of their host species.
    Orthologues: for speciation
    Paralogues: for duplication

CISC 889, S04, Lec1, Liao

Phylogenetic analysis, C'td

- Tree reconstruction
  - Distance-based
    - Un-weighted Pair Group Method with Arthmetric mean (UPGMA)
    - Ultrametric Trees.
    - Neighbor-joining
    - Additive Trees
  - Character-based: In general, an exact solution is NP-hard, and heuristics should be invoked.
    - Parsimony
  - Maximum Likelihood (ML): A probabilistic evolution model is needed

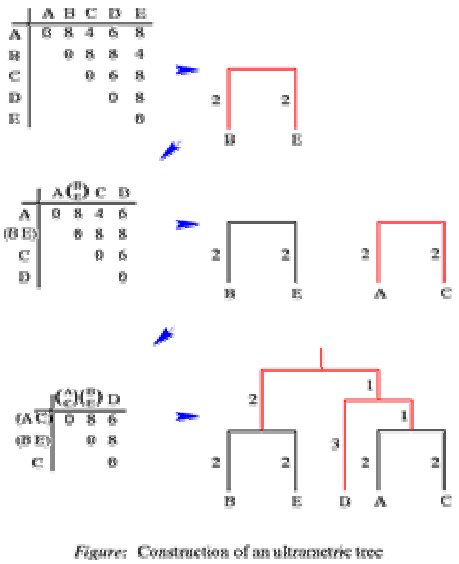
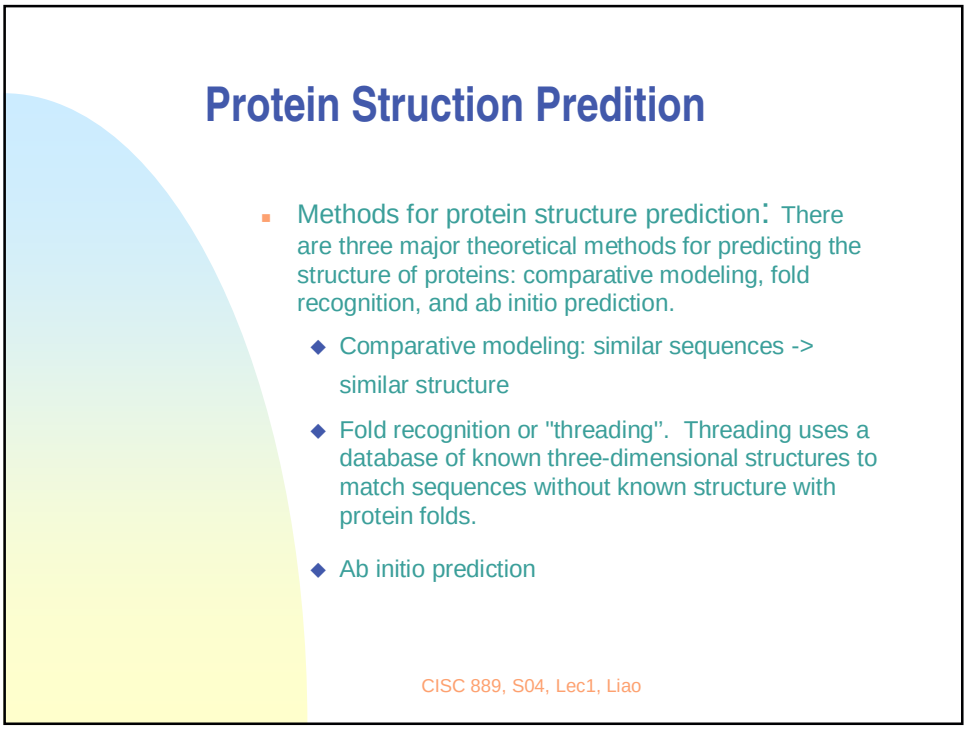


Figure: Construction of an ultrametric tree


Protein Struction Predition
Methods for protein structure prediction: There are three major theoretical methods for predicting the structure of proteins: comparative modeling, fold recognition, and ab initio prediction.
Comparative modeling: similar sequences -> similar structure
Fold recognition or "threading". Threading uses a database of known three-dimensional structures to match sequences without known structure with protein folds.
Ab initio prediction
CISC 889, S04, Lec1, Liao