CENG 465
Introduction to Bioinformatics

Spring 2006-2007

Tolga Can (Office: B-109)
e-mail: tcan@ceng.metu.edu.tr

Course Web Page:
http://www.ceng.metu.edu.tr/~tcan/ceng465/

1

Goals of the course

- Working at the interface of computer science and biology
  - New motivation
  - New data and new demands
  - Real impact
- Introduction to main issues in computational biology
- Opportunity to interact with algorithms, tools, data in current practice

2

High level overview of the course

- A general introduction
  - what problems are people working on?
  - how people solve these problems?
  - what key computational techniques are needed?
  - how much help computing has provided to biological research?
- A way of thinking -- tackling "biological problems" computationally
  - how to look at a "biological problem" from a computational point of view?
  - how to formulate a computational problem to address a biological issue?
  - how to collect statistics from biological data?
  - how to build a "computational" model?
  - how to solve a computational modeling problem?
  - how to test and evaluate a computational algorithm?

3

Course outline

- Motivation and introduction to biology (1 week)
- Sequence analysis (4 weeks)
  - Analyze DNA and protein sequences for clues regarding function
  - Identification of homologues
    - Pairwise sequence alignment
  - Statistical significance of sequence alignments
  - Suffix trees
  - Multiple sequence alignment
- Phylogenetic trees, clustering methods (1 week)

4

Course outline

- Protein structures (4 weeks)
  - Analyze protein structures for clues regarding function
    - Structure alignment
  - Structure prediction (secondary, tertiary)
  - Motifs, active sites, docking
  - Multiple structural alignment, geometric hashing
- Microarray data analysis (2 weeks)
  - Correlations, clustering
  - Inference of function
- Gene/Protein networks, pathways (2 weeks)
  - Protein-protein, protein/DNA interactions
  - Construction and analysis of large scale networks

5

Grading

- 2 Midterm exams - 20% each
- Final exam - 30%
- Written assignments - 15%
- Programming assignments - 15%

6

Miscellaneous

- Course webpage
  - http://www.ceng.metu.edu.tr/~tcan/ceng465/
  - Lecture slides
  - Assignments
  - Announcements
  - Other relevant information
  - Reading materials
    - Your first reading assignment:
      - J. Cohen, *Bioinformatics – An introduction to computer scientists*.
- Newsgroup
  - metu.ceng.course.465

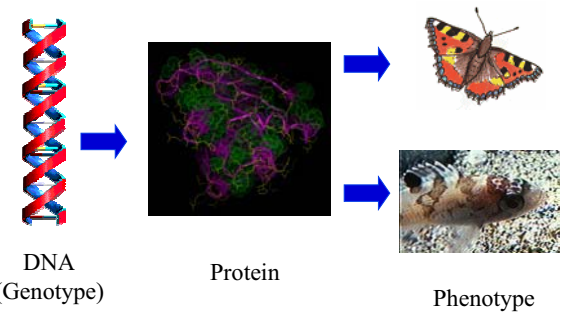
7

What is Bioinformatics?

- *(Molecular)* **Bio - `informatics`**
- One idea for a definition?
  Bioinformatics is conceptualizing **biology in terms of molecules** (in the sense of physical-chemistry) and then applying **"informatics" techniques** (derived from disciplines such as applied math, CS, and statistics) to understand and **organize the information associated** with these molecules, **on a large-scale.**
- Bioinformatics is a practical discipline with many **applications**.

8

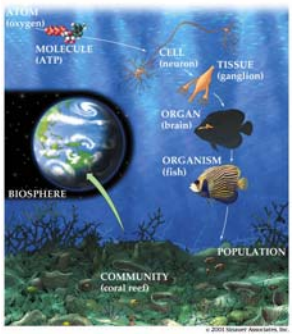
Introductory Biology



DNA (Genotype) — Protein — Phenotype

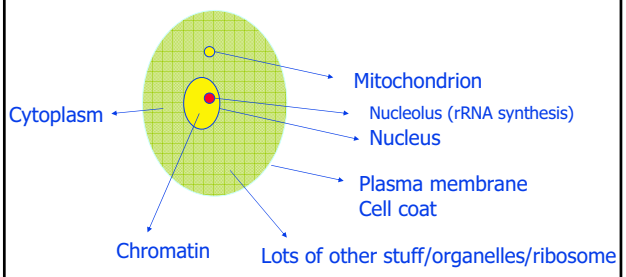
9

Scales of life

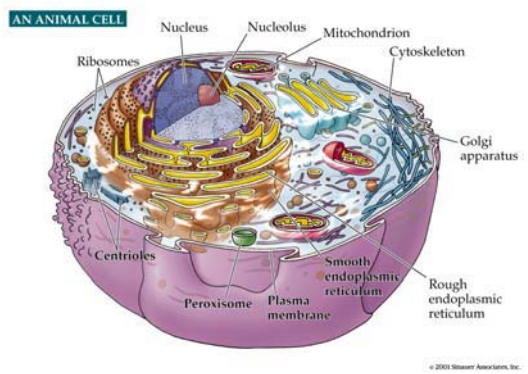


10

Animal Cell



11



12

Two kinds of Cells

- Prokaryotes – no nucleus (bacteria)
  - Their genomes are circular
- Eukaryotes – have nucleus (animal,plants)
  - Linear genomes with multiple chromosomes in pairs. When pairing up, they look like

Middle: centromere
Top: p-arm
Bottom: q-arm

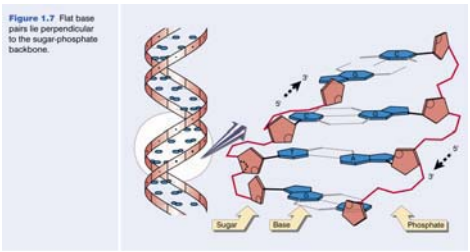
13

Molecular Biology Information - DNA

- Raw DNA Sequence
  - Coding or Not?
  - Parse into genes?
  - 4 bases: AGCT
  - ~1 Kb in a gene, ~2 Mb in genome
  - ~3 Gb Human

14

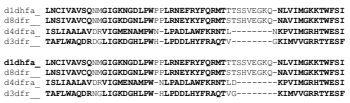
DNA structure



15

Molecular Biology Information: Protein Sequence

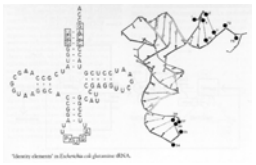
- 20 letter alphabet
  - ACDEFGHIKLMNPQRSTVWY but not BJOUXZ
- Strings of ~300 aa in an average protein (in bacteria),
  ~200 aa in a domain
- ~1M known protein sequences

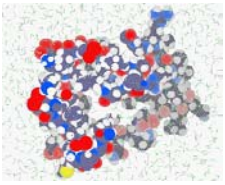


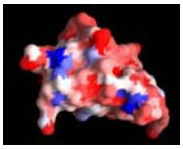
16

Molecular Biology Information: Macromolecular Structure

- DNA/RNA/Protein
  - Almost all protein



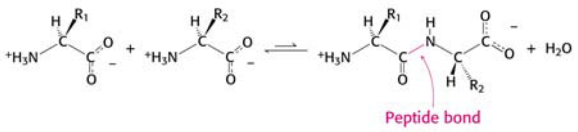




17

More on Macromolecular Structure

- Primary structure of proteins
  - Linear polymers linked by peptide bonds
  - Sense of direction

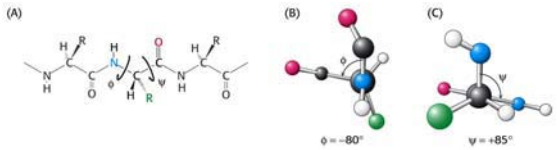


18

3

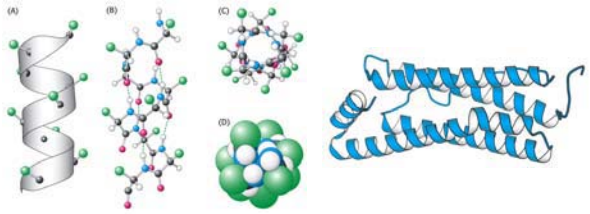
Secondary Structure

- Polypeptide chains fold into regular local structures
  - alpha helix, beta sheet, turn, loop
  - based on energy considerations
  - Ramachandran plots



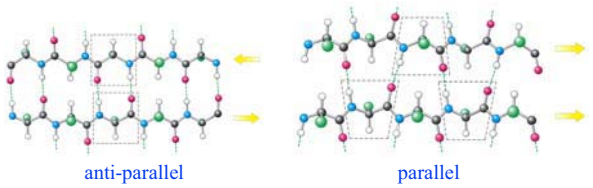
19

Alpha helix

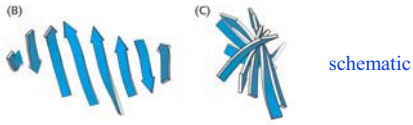


20

Beta sheet



anti-parallel parallel

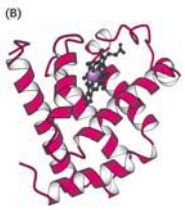


schematic

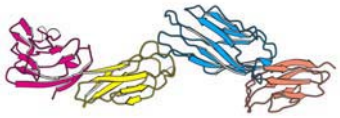
21

Tertiary Structure

- 3-d structure of a polypeptide sequence
  - interactions between non-local and foreign atoms
  - often separated into domains



tertiary structure of myoglobin

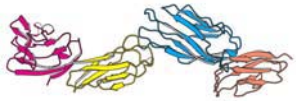


domains of CD4

22

Quaternary Structure

- Arrangement of protein subunits
  - dimers, tetramers



quaternary structure of Cro

human hemoglobin tetramer

23

Structure summary

- 3-d structure determined by protein sequence
- Cooperative and progressive stabilization
- Prediction remains a challenge
  - ab-initio (energy minimization)
  - knowledge-based
    - Chou-Fasman and GOR methods for SSE prediction
    - Comparative modeling and protein threading for tertiary structure prediction
- Diseases caused by misfolded proteins
  - Mad cow disease
- Classification of protein structures

24

4

Genes and Proteins

- One gene encodes one* protein.
- Like a program, it starts with start codon (e.g. ATG), then each three code one amino acid. Then a stop codon (e.g. TGA) signifies end of the gene.
- Sometimes, in the middle of a (eukaryotic) gene, there are introns that are spliced out (as junk) during transcription. Good parts are called exons. This is the task of gene finding.

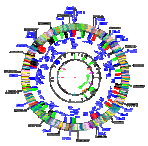
25

A.A. Coding Table

Glycine (GLY) GG*	Arginine (ARG) CG*
Alanine(ALA) GC*	Asparagine (ASN) AAT, AAC
Valine (VAL) GT*	Glutamine (GLN) CAA, CAG
Leucine (LEU) CT*	Cysteine (CYS) TGT, TGC
Isoleucine (ILE) AT(*-G)	Methionine (MET) ATG
Serine (SER) AGT, AGC	Phenylalanine (PHE) TTT,TTC
Threonine (THR) AC*	Tyrosine (TYR) TAT, TAC
Aspartic Acid (ASP) GAT,GAC	Tryptophan (TRP) TGG
Glutamic Acid(GLU) GAA,GAG	Histidine (HIS) CAT, CAC
Lysine (LYS) AAA, AAG	Proline (PRO) CC*
Start: ATG, CTG, GTG	Stop TGA, TAA, TAG

26

Molecular Biology Information: Whole Genomes



Genome sequences now accumulate so quickly that, in less than a week, a single laboratory can produce more bits of data than Shakespeare managed in a lifetime, although the latter make better reading.

-- G A Pekso, *Nature* **401**: 115-116 (1999)

27

1995

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

1997

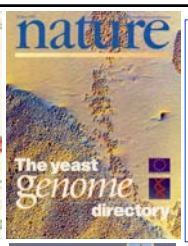
Eukaryote, 13 Mb, ~6K genes [Nature **387**: 1]

1998

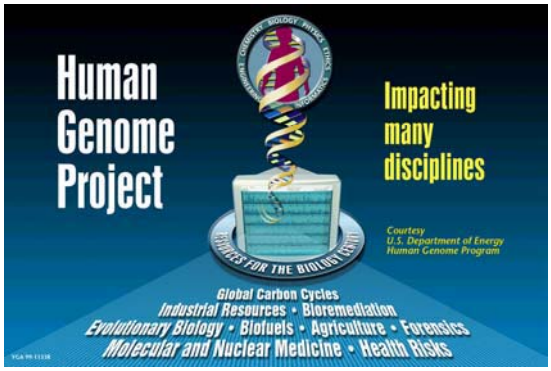
Animal, ~100 Mb, ~20K genes [*Science* **282**: 1945]

2000?

Human, ~3 Gb, ~100K genes [???]



Genomes highlight the Finiteness of the "Parts" in Biology



29

Gene Expression Datasets: the Transcriptome

Young/Lander, Chips, Abs. Exp.

Brown, µarray, Rel. Exp. over Timecourse

Also: SAGE; Samson and Church, Chips; Aebersold, Protein Expression

Snyder, Transposons, Protein Exp.

30

5

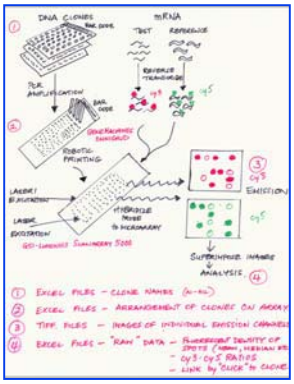
Array Data

Yeast Expression Data in Academia:
levels for all 6000 genes!

Can only sequence genome once but can do an infinite variety of these array experiments

at 10 time points,
6000 x 10 = 60K floats

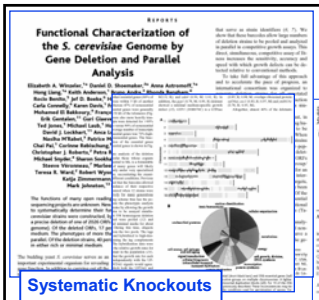
telling signal from background



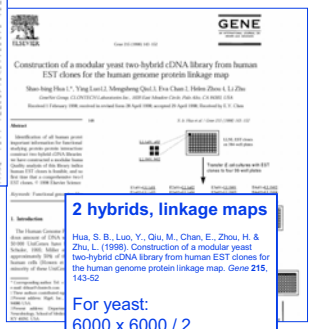
(courtesy of J Hager)

31

Other Whole-Genome Experiments



Functional Characterization of the *S. cerevisiae* Genome by Gene Deletion and Parallel Analysis



Construction of a modular yeast two-hybrid cDNA library from human EST clones for the human genome protein linkage map

Systematic Knockouts

Winzeler, E. A., Shoemaker, D. D., Astromoff, A., Liang, H., Anderson, K., Andre, B., Bangham, R., Benito, R., Boeke, J. D., Bussey, H., Chu, A. M., Connelly, C., Davis, K., Dietrich, F., Dow, S. W., El Bakkoury, M., Foury, F., Friend, S. H., Gentalen, E., Giaever, G., Hegemann, J. H., Jones, T., Laub, M., Liao, H., Davis, R. W. & et al. (1999). Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. *Science* **285**, 901-6

2 hybrids, linkage maps

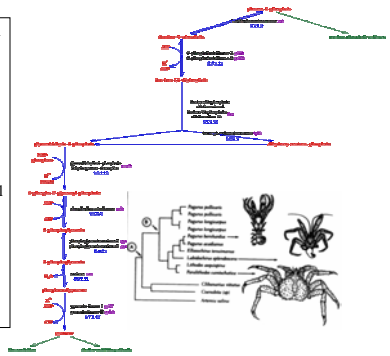
Hua, S. B., Luo, Y., Qiu, M., Chan, E., Zhou, H. & Zhu, L. (1998). Construction of a modular yeast two-hybrid cDNA library from human EST clones for the human genome protein linkage map. *Gene* **215**, 143-52

For yeast:
6000 x 6000 / 2
~ 18M interactions

32

Molecular Biology Information: Other Integrative Data

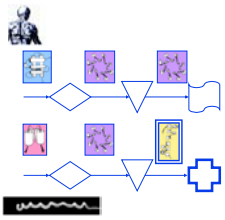
- Information to understand genomes
  - Metabolic Pathways (glycolysis), traditional biochemistry
  - Regulatory Networks
  - Whole Organisms Phylogeny, traditional zoology
  - Environments, Habitats, ecology
  - The Literature (MEDLINE)
- The Future....



33

Organizing Molecular Biology Information: Redundancy and Multiplicity

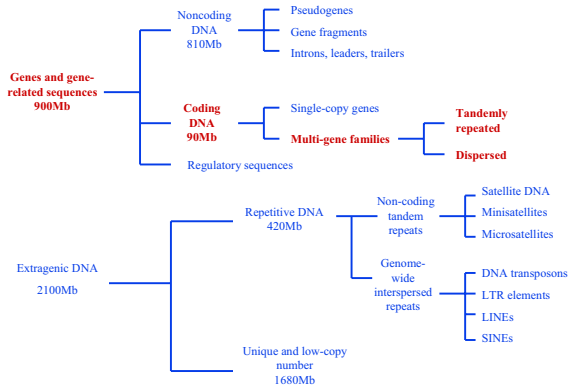
- Different Sequences Have the Same Structure
- Organism has many similar genes
- Single Gene May Have Multiple Functions
- Genes are grouped into Pathways
- Genomic Sequence Redundancy due to the Genetic Code
- **How do we find the similarities?.....**



Integrative Genomics - genes ↔ structures ↔ **functions** ↔ **pathways** ↔ expression levels ↔ regulatory systems ↔ ....

34

Human genome



35

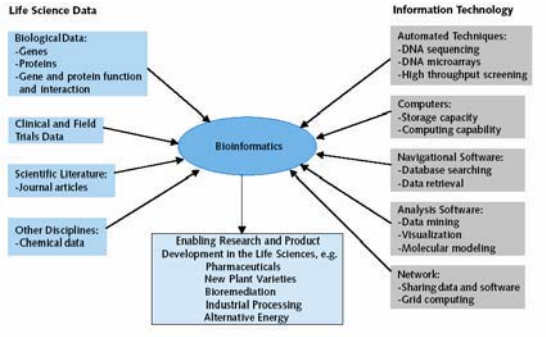
Where to get data?

- GenBank
  - http://www.ncbi.nlm.nih.gov
- Protein Databases
  - SWISS-PROT: http://www.expasy.ch/sprot
  - PDB: http://www.pdb.bnl.gov/
- And many others

36

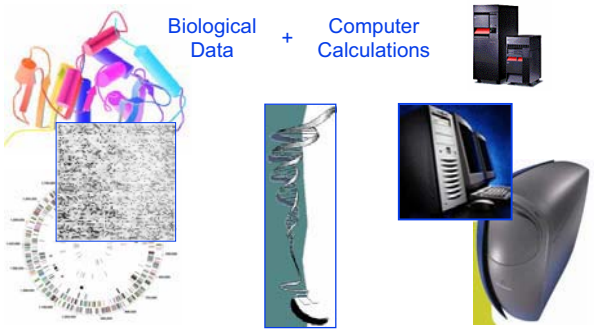
6

Figure 6.1. Bioinformatics Uses Information Technology to Manage and Analyze Information Generated by the Life Sciences



37

Bioinformatics: A simple view



38

Application domains

Table 6.2. Number of Survey Respondents Indicating Bioinformatics Research Activities by Application, 2002

Application	Number of firms in application	Conduct bioinformatics research
Human Health	780	247
Animal Health	144	37
Agricultural & Aquacultural/Marine	128	41
Marine & Terrestrial Microbial	41	19
Industrial and Agricultural-Derived Processing	132	45
Environmental Remediation and Natural Resource Recovery	41	12
Other Bio-defense	160	30

Note: The total number of firms that responded to the biotechnology survey was 1,031, and 304 of these firms indicated that they had some activity in bioinformatics. The number of firms by biotechnology application does not add up to the total number of firms that responded to the survey because firms were classified in an application if they indicated it as either a "primary" or "secondary" focus.

Source: Survey data from *Critical Technology Assessment of Biotechnology in U.S. Industry*, U.S. Department of Commerce, Technology Administration and Bureau of Industry and Security, August 2002.

39

Kinds of activities

	Conduct research on/in	Approved, marketed, or in production		Total
		Product(s)	Process(es)	
DNA-based				
Bioinformatics	29	2	1	30
Genomics, pharmacogenetics	29	3	2	30
DNA sequencing/synthesis/ amplification, genetic engineering	39	5	3	43
Biochemistry/Immunology				
Drug design & delivery	33	4	2	38
Synthesis/sequencing of proteins and peptides	27	3	1	30
Combinatorial chemistry, 3-D molecular modeling	18	1	0	19

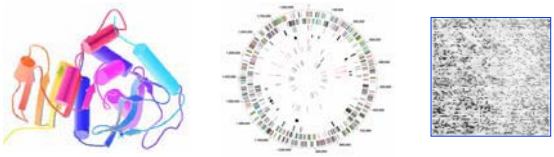
Note: The total number of responses to the biotechnology activity question was 1021. Percents do not add up to 100 percent because firms can have more than one activity.

Source: Survey data from *Critical Technology Assessment of Biotechnology in U.S. Industry*, U.S. Department of Commerce, Technology Administration and Bureau of Industry and Security, August 2002.

40

Motivation

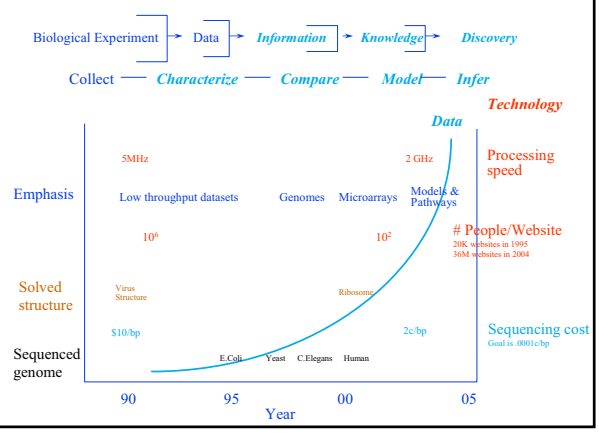
- Diversity and size of information
  - Sequences, 3-D structures, microarrays, protein interaction networks, *in silico* models, bio-images



- Understand the relationship
  - Similar to complex software design

41

Bioinformatics - A Revolution



7

Computing *versus* Biology

- *what computer science is to molecular biology is like what mathematics has been to physics .....*
  -- Larry Hunter, ISMB'94
- *molecular biology is (becoming) an information science .....*
  -- Leroy Hood, RECOMB'00
- *bioinformatics ... is the research domain focused on linking the behavior of biomolecules, biological pathways, cells, organisms, and populations to the information encoded in the genomes* --Temple Smith, Current Topics in Computational Molecular Biology

43

Computing *versus* Biology
looking into the future

- *Like physics, where general rules and laws are taught at the start, biology will surely be presented to future generations of students as a set of basic systems ..... duplicated and adapted to a very wide range of cellular and organismic functions, following basic evolutionary principles constrained by Earth's geological history.* --Temple Smith, Current Topics in Computational Molecular Biology

44

Scalability challenges

- Recent issue of NAR devoted to data collections contains 719 databases
  - Sequence
    - Genomes (more than 150), ESTs, Promoters, transcription factor binding sites, repeats, ..
  - Structure
    - Domains, motifs, classifications, ..
  - Others
    - Microarrays, subcellular localization, ontologies, pathways, SNPs, ..

45

Challenges of working in bioinformatics

- Need to feel comfortable in interdisciplinary area
- Depend on others for primary data
- Need to address important biological *and* computer science problems

46

Skill set

- Artificial intelligence
- Machine learning
- Statistics & probability
- Algorithms
- Databases
- Programming

47

Bioinformatics Topics
Genome Sequence

- Finding Genes in Genomic DNA
  - introns
  - exons
  - promotors
- Characterizing Repeats in Genomic DNA
  - Statistics
  - Patterns
- Duplications in the Genome
  - Large scale genomic alignment

48

Bioinformatics Topics Protein Sequence

- Sequence Alignment
  - non-exact string matching, gaps
  - How to align two strings optimally via Dynamic Programming
  - Local vs Global Alignment
  - Suboptimal Alignment
  - Hashing to increase speed (BLAST, FASTA)
  - Amino acid substitution scoring matrices
- Multiple Alignment and Consensus Patterns
  - How to align more than one sequence and then fuse the result in a consensus representation
  - Transitive Comparisons
  - HMMs, Profiles
  - Motifs

- Scoring schemes and Matching statistics
  - How to tell if a given alignment or match is statistically significant
  - A P-value (or an e-value)?
  - Score Distributions (extreme val. dist.)
  - Low Complexity Sequences
- Evolutionary Issues
  - Rates of mutation and change

49

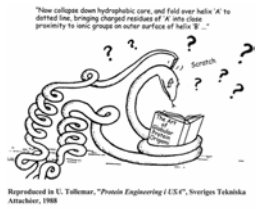
Computationally challenging problems

- More sensitive pairwise alignment
  - Dynamic programming is O(mn)
    - m is the length of the query
    - n is the length of the database
- Scalable multiple alignment
  - Dynamic programming is exponential in number of sequences
  - Currently feasible for around 10 protein sequences of length around 1000
- Shotgun alignment
  - Current techniques will take over 200 days on a single machine to align the mouse genome

50

Bioinformatics Topics Sequence / Structure

- Secondary Structure "Prediction"
  - via Propensities
  - Neural Networks, Genetic Alg.
  - Simple Statistics
  - TM-helix finding
  - Assessing Secondary Structure Prediction
- Structure Prediction: Protein and RNA



- Tertiary Structure Prediction
  - Fold Recognition
  - Threading
  - Ab initio
- Function Prediction
  - Active site identification
- Relation of Sequence Similarity to Structural Similarity

51

Topics -- Structures

- Basic Protein Geometry and Least-Squares Fitting
  - Distances, Angles, Axes, Rotations
    - Calculating a helix axis in 3D via fitting a line
  - LSQ fit of 2 structures
  - Molecular Graphics
- Calculation of Volume and Surface
  - How to represent a plane
  - How to represent a solid
  - How to calculate an area
  - Docking and Drug Design as Surface Matching
  - Packing Measurement

- Structural Alignment
  - Aligning sequences on the basis of 3D structure.
  - DP does not converge, unlike sequences, what to do?
  - Other Approaches: Distance Matrices, Hashing
  - Fold Library

52

Computationally challenging problems

- Alignment against a database
  - Single comparison usually takes seconds.
  - Comparison against a database takes hours.
  - All-against-all comparison takes weeks.
- Multiple structure alignment and motifs
- Combined sequence and structure comparison
- Secondary and tertiary structure prediction

53

Topics -- Databases

- Relational Database Concepts and how they interface with Biological Information
  - Keys, Foreign Keys
  - SQL, OODBMS, views, forms, transactions, reports, indexes
  - Joining Tables, Normalization
    - Natural Join as "where" selection on cross product
    - Array Referencing (perl/dbm)
  - Forms and Reports
  - Cross-tabulation
- Protein Units?
  - What are the units of biological information?
    - sequence, structure
    - motifs, modules, domains
  - How classified: folds, motions, pathways, functions?

- Clustering and Trees
  - Basic clustering
    - UPGMA
    - single-linkage
    - multiple linkage
  - Other Methods
    - Parsimony, Maximum likelihood
  - Evolutionary implications
- Visualization of Large Amounts of Information
- The Bias Problem
  - sequence weighting
  - sampling

54

9

Topics -- Genomics

- Expression Analysis
  - Time Courses clustering
  - Measuring differences
  - Identifying Regulatory Regions
- Large scale cross referencing of information
- Function Classification and Orthologs
- The Genomic vs. Single-molecule Perspective

- Genome Comparisons
  - Ortholog Families, pathways
  - Large-scale censuses
  - Frequent Words Analysis
  - Genome Annotation
  - Trees from Genomes
  - Identification of interacting proteins
- Structural Genomics
  - Folds in Genomes, shared & common folds
  - Bulk Structure Prediction
- Genome Trees

55

Topics -- Simulation

- Molecular Simulation
  - Geometry -> Energy -> Forces
  - Basic interactions, potential energy functions
  - Electrostatics
  - VDW Forces
  - Bonds as Springs
  - How structure changes over time?
    - How to measure the change in a vector (gradient)
  - Molecular Dynamics & MC
  - Energy Minimization

- Parameter Sets
- Number Density
- Poisson-Boltzman Equation
- Lattice Models and Simplification

56

General Types of "Informatics" techniques in Bioinformatics

- Databases
  - Building, querying
  - Schema design
  - Heterogeneous, distributed
- Similarity search
  - Sequence, structure
  - Significance statistics

- Finding Patterns
  - AI / Machine Learning
  - Clustering
  - Data mining
- Modeling & simulation
- Programming
  - Perl
  - Java/C/C++/..

57