Molecular Biophysics & Biochemistry
447b3 / 747b3

Bioinformatics

Mark Gerstein

Class #1, 1/12/98

Yale University

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 “MIS” for Molecular Biology Information

Molecular Biology as an Information Science

- Central Dogma of Molecular Biology

```
DNA
 -> RNA
  -> Protein

   -> Phenotype
    -> DNA
```

- Molecules
  - Sequence, Structure, Function
- Processes
  - Mechanism, Specificity, Regulation

- Central Paradigm for Bioinformatics

```
Genomic Sequence Information
  -> Protein Sequence
   -> Protein Structure
    -> Protein Function
     -> Phenotype
```

- Large Amounts of Information
  - Standardized
  - Statistical

(idea from D Brutlag, Stanford)

Molecular Biology Information - DNA

- Raw DNA Sequence
  - ◊ Coding or Not?
  - ◊ Parse into genes?
  - ◊ 4 bases: AGCT
  - ◊ ~1 K in a gene, ~2 M in genome

```
atggcaattaaaattggtatcaatggttttggtcgtatcggccgtatcgtattccgtgca
gcacaacaccgtgatgacattgaagttgtaggtattaacgacttaatcgacgttgaatac
atggcttatatgttgaaatatgattcaactcacggtcgtttcgacggcactgttgaagtg
aaagatggtaacttagtggttaatggtaaaactatccgtgtaactgcagaacgtgatcca
gcaaacttaaactggggtgcaatcggtgttgatatcgctgttgaagcgactggtttattc
ttaactgatgaaactgctcgtaaacatatcactgcaggcgcaaaaaaagttgtattaact
ggcccatctaaagatgcaacccctatgttcgttcgtggtgtaaacttcaacgcatacgca
ggtcaagatatcgtttctaacgcatcttgtacaacaaactgtttagctcctttagcacgt
gttgttcatgaaactttcggtatcaaagatggtttaatgaccactgttcacgcaacgact
gcaactcaaaaaactgtggatggtccatcagctaaagactggcgcggcggccgcggtgca
tcacaaaacatcattccatcttcaacaggtgcagcgaaagcagtaggtaaagtattacct
gcattaaacggtaaattaactggtatggctttccgtgttccaacgccaaacgtatctgtt
gttgatttaacagttaatcttgaaaaaccagcttcttatgatgcaatcaaacaagcaatc
aaagatgcagcggaaggtaaaacgttcaatggcgaattaaaaggcgtattaggttacact
gaagatgctgttgtttctactgacttcaacggttgtgctttaacttctgtatttgatgca
gacgctggtatcgcattaactgattctttcgttaaattggtatc . . .


. . .   caaaaatagggttaatatgaatctcgatctccattttgttcatcgtattcaa
caacaagccaaaactcgtacaaatatgaccgcacttcgctataaagaacacggcttgtgg
cgagatatctcttggaaaaactttcaagagcaactcaatcaactttctcgagcattgctt
gctcacaatattgacgtacaagataaaatcgccattttttgcccataatatggaacgttgg
gttgttcatgaaactttcggtatcaaagatggtttaatgaccactgttcacgcaacgact
acaatcgttgacattgcgaccttacaaattcgagcaatcacagtgcctatttacgcaacc
aatacagcccagcaagcagaatttatcctaaatcacgccgatgtaaaaattctcttcgtc
ggcgatcaagagcaatacgatcaaacattggaaattgctcatcattgtccaaaattacaa
aaaattgtagcaatgaaatccaccattcaattacaacaagatcctctttcttgcacttgg
```

Molecular Biology Information: Protein Sequence

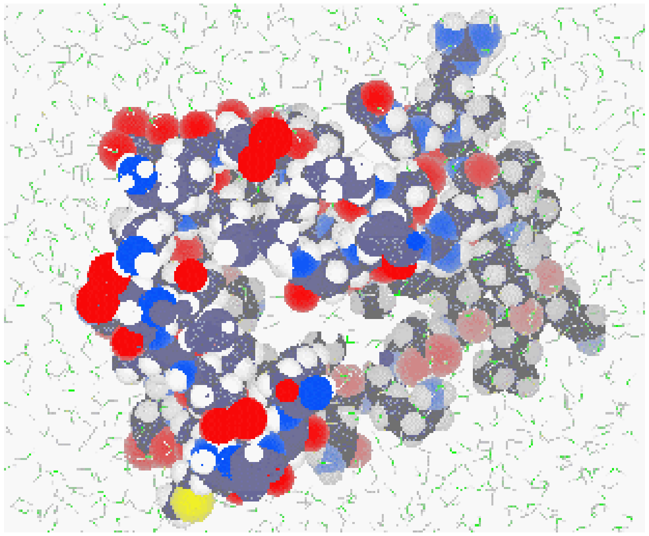
- 20 letter alphabet
  - ◊ ACDEFGHIKLMNPQRSTVWY but not BJOUXZ
- Strings of ~300 aa in an average protein (in bacteria), ~200 aa in a domain
- ~200 K known protein sequences

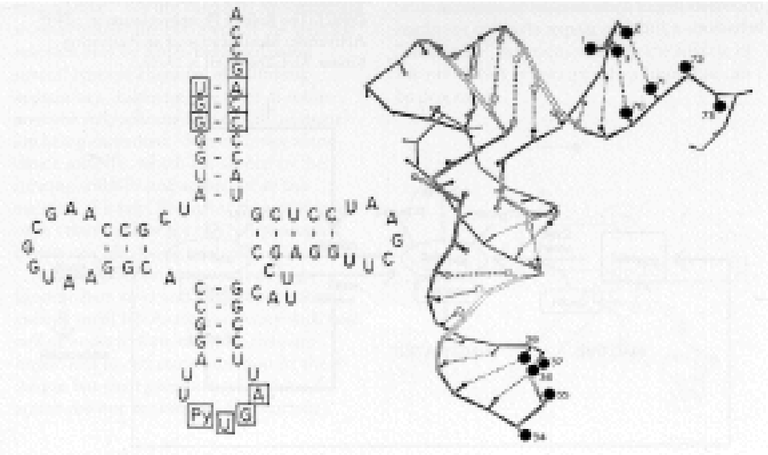
```
d1dhfa_ LNCIVAVSQNMGIGKNGDLPWPPLRNEFRYFQRMTTTSSVEGKQ-NLVIMGKKTWFSI
d8dfr__ LNSIVAVCQNMGIGKDGNLPWPPLRNEYKYFQRMTSTSHVEGKQ-NAVIMGKKTWFSI
d4dfra_ ISLIAALAVDRVIGMENAMPWN-LPADLAWFKRNTL--------NKPVIMGRHTWESI
d3dfr__ TAFLWAQDRDGLIGKDGHLPWH-LPDDLHYFRAQTV--------GKIMVVGRRTYESF

d1dhfa_ LNCIVAVSQNMGIGKNGDLPWPPLRNEFRYFQRMTTTSSVEGKQ-NLVIMGKKTWFSI
d8dfr__ LNSIVAVCQNMGIGKDGNLPWPPLRNEYKYFQRMTSTSHVEGKQ-NAVIMGKKTWFSI
d4dfra_ ISLIAALAVDRVIGMENAMPW-NLPADLAWFKRNTLD--------KPVIMGRHTWESI
d3dfr__ TAFLWAQDRNGLIGKDGHLPW-HLPDDLHYFRAQTVG--------KIMVVGRRTYESF

d1dhfa_ VPEKNRPLKGRINLVLSRELKEPPQGAHFLSRSLDDALKLTEQPELANKVDMVWIVGGSSVYKEAMNHP
d8dfr__ VPEKNRPLKDRINIVLSRELKEAPKGAHYLSKSLDDALALLDSPELKSKVDMVWIVGGTAVYKAAMEKP
d4dfra_ ---G-RPLPGRKNIILS-SQPGTDDRV-TWVKSVDEAIAACGDVP------EIMVIGGGRVYEQFLPKA
d3dfr__ ---PKRPLPERTNVVLTHQEDYQAQGA-VVVHDVAAVFAYAKQHLDQ----ELVIAGGAQIFTAFKDDV

d1dhfa_ -PEKNRPLKGRINLVLSRELKEPPQGAHFLSRSLDDALKLTEQPELANKVDMVWIVGGSSVYKEAMNHP
d8dfr__ -PEKNRPLKDRINIVLSRELKEAPKGAHYLSKSLDDALALLDSPELKSKVDMVWIVGGTAVYKAAMEKP
d4dfra_ -G---RPLPGRKNIILSSSQPGTDDRV-TWVKSVDEAIAACGDVPE----- IMVIGGGRVYEQFLPKA
d3dfr__ -P--KRPLPERTNVVLTHQEDYQAQGA-VVVHDVAAVFAYAKQHLD----QELVIAGGAQIFTAFKDDV
```

Molecular Biology Information: Macromolecular Structure

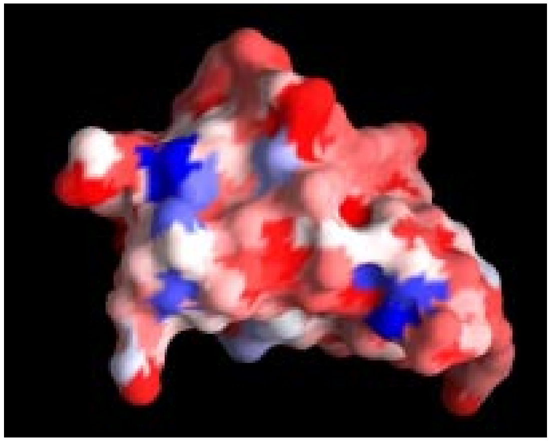
- DNA/RNA/Protein
  - ◊ Almost all protein

(RNA Adapted From D Soll Web Page,
Right Hand Top Protein from M Levitt web page)





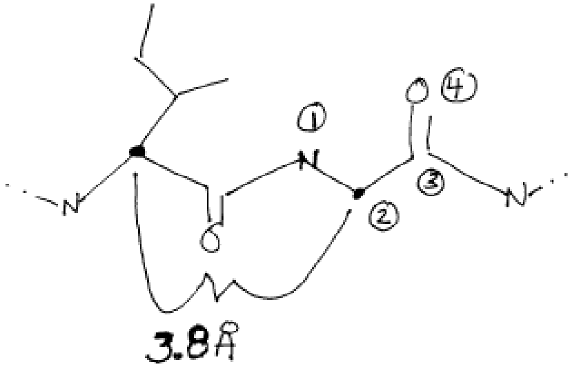
'Identity elements' in Escherichia coli glutamine tRNA.



Molecular Biology Information: Protein Structure Details

- Statistics on Number of XYZ triplets
  - ◊ 200 residues/domain -> 200 CA atoms, separated by 3.8 A
  - ◊ Avg. Residue is Leu: 4 backbone atoms + 4 sidechain atoms, 150 cubic A
    - => ~1500 xyz triplets (=8x200) per protein domain
  - ◊ 10 K known domain, ~300 folds

```
ATOM      1  C   ACE     0       9.401  30.166  60.595  1.00 49.88      1GKY  67
ATOM      2  O   ACE     0      10.432  30.832  60.722  1.00 50.35      1GKY  68
ATOM      3  CH3 ACE     0       8.876  29.767  59.226  1.00 50.04      1GKY  69
ATOM      4  N   SER     1       8.753  29.755  61.685  1.00 49.13      1GKY  70
ATOM      5  CA  SER     1       9.242  30.200  62.974  1.00 46.62      1GKY  71
ATOM      6  C   SER     1      10.453  29.500  63.579  1.00 41.99      1GKY  72
ATOM      7  O   SER     1      10.593  29.607  64.814  1.00 43.24      1GKY  73
ATOM      8  CB  SER     1       8.052  30.189  63.974  1.00 53.00      1GKY  74
ATOM      9  OG  SER     1       7.294  31.409  63.930  1.00 57.79      1GKY  75
ATOM     10  N   ARG     2      11.360  28.819  62.827  1.00 36.48      1GKY  76
ATOM     11  CA  ARG     2      12.548  28.316  63.532  1.00 30.20      1GKY  77
ATOM     12  C   ARG     2      13.502  29.501  63.500  1.00 25.54      1GKY  78
. . .
ATOM   1444  CB  LYS   186      13.836  22.263  57.567  1.00 55.06      1GKY1510
ATOM   1445  CG  LYS   186      12.422  22.452  58.180  1.00 53.45      1GKY1511
ATOM   1446  CD  LYS   186      11.531  21.198  58.185  1.00 49.88      1GKY1512
ATOM   1447  CE  LYS   186      11.452  20.402  56.860  1.00 48.15      1GKY1513
ATOM   1448  NZ  LYS   186      10.735  21.104  55.811  1.00 48.41      1GKY1514
ATOM   1449  OXT LYS   186      16.887  23.841  56.647  1.00 62.94      1GKY1515
TER    1450      LYS   186                                              1GKY1516
```

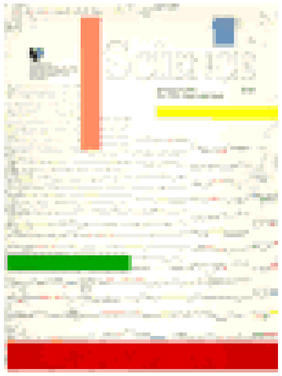


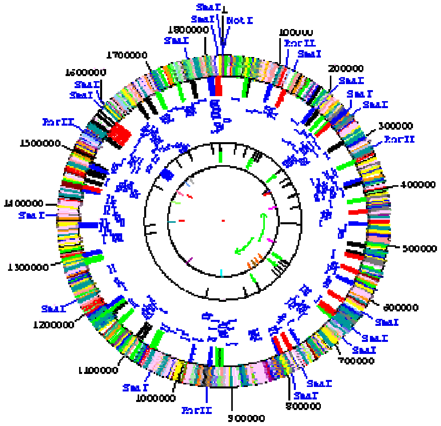
Molecular Biology Information: Whole Genomes

- The Revolution Driving Everything

  Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., Merrick, J. M., McKenney, K., Sutton, G., Fitzhugh, W., Fields, C., Gocayne, J. D., Scott, J., Shirley, R., Liu, L. I., Glodek, A., Kelley, J. M., Weidman, J. F., Phillips, C. A., Spriggs, T., Hedblom, E., Cotton, M. D., Utterback, T. R., Hanna, M. C., Nguyen, D. T., Saudek, D. M., Brandon, R. C., Fine, L. D., Fritchman, J. L., Fuhrmann, J. L., Geoghagen, N. S. M., Gnehm, C. L., McDonald, L. A., Small, K. V., Fraser, C. M., Smith, H. O. & Venter, J. C. (1995). "Whole-genome random sequencing and assembly of Haemophilus influenzae rd." *Science* 269: 496-512.

  (Picture adapted from TIGR website, http://www.tigr.org)





- Integrative Data

  1995, HI (bacteria): 1.6 Mb & 1600 genes done

  1997, yeast: 13 Mb & ~6000 genes for yeast

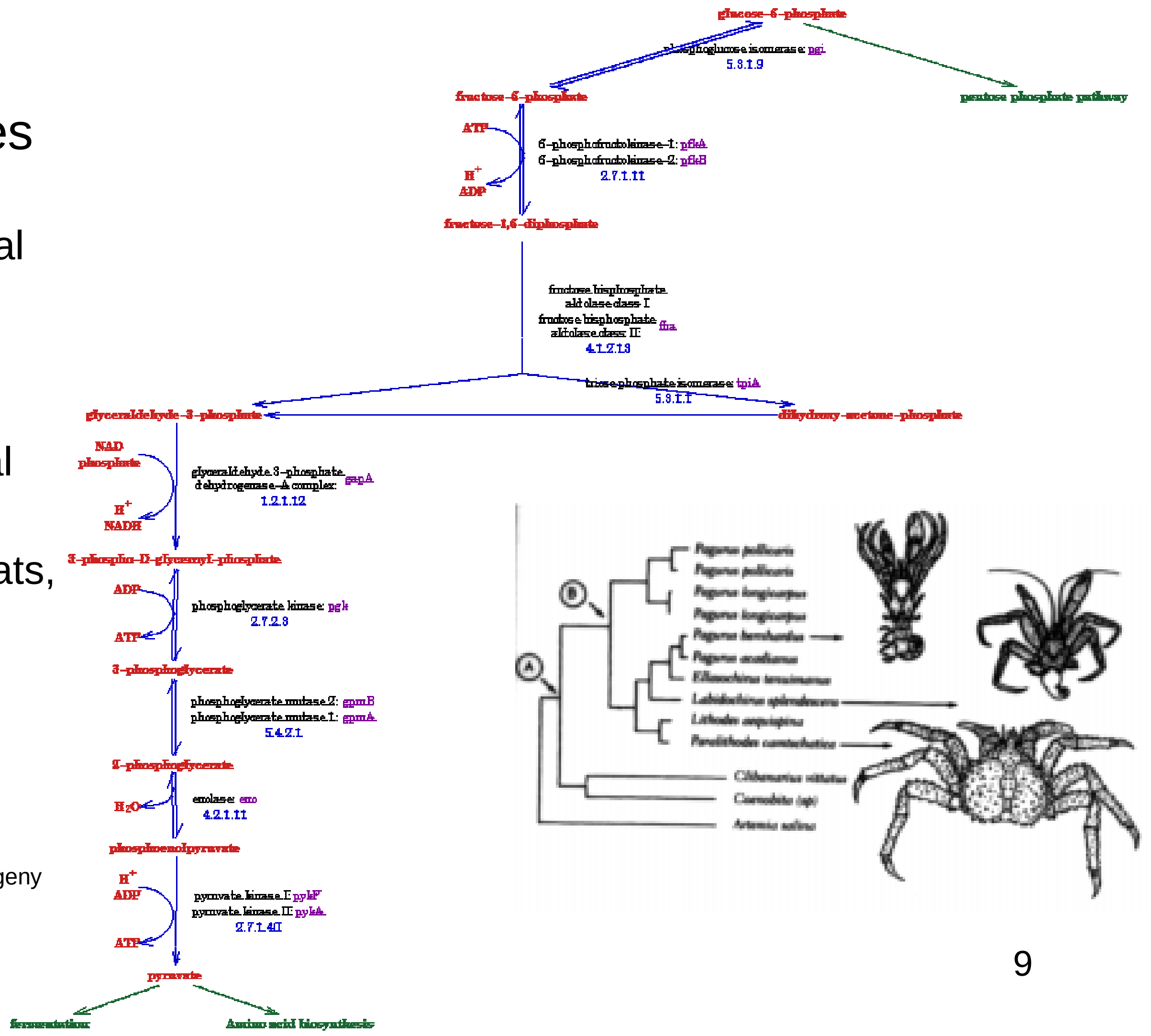
  1998: 14 completed genomes!

  1998, worm: 75 of 100 Mb done
  with 13 K genes so far

  2003, human: 3 Gb & 100 K genes...

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....



(Pathway drawing from P Karp's EcoCyc, Phylogeny from S J Gould, Dinosaur in a Haystack)

Molecular Biology Information: Redundancy and Multiplicity

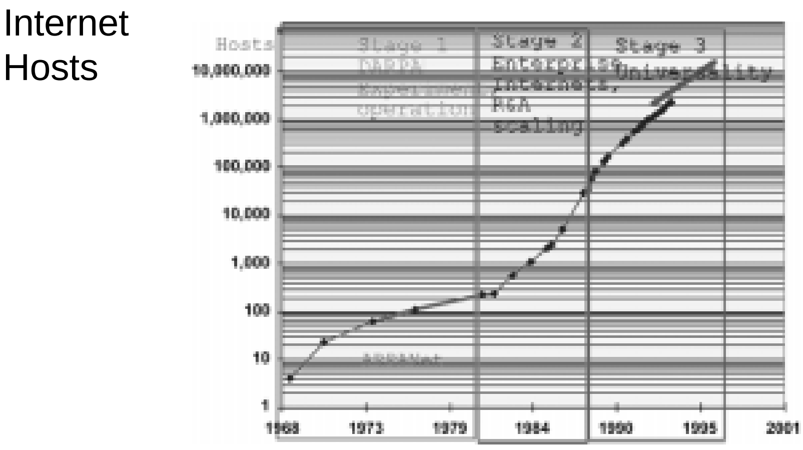
- Different Sequences Have the Same Structure
- Organism has many similar genes
- Single Gene May Have Multiple Functions
- Genomic Sequence Redundancy due to the Genetic Code

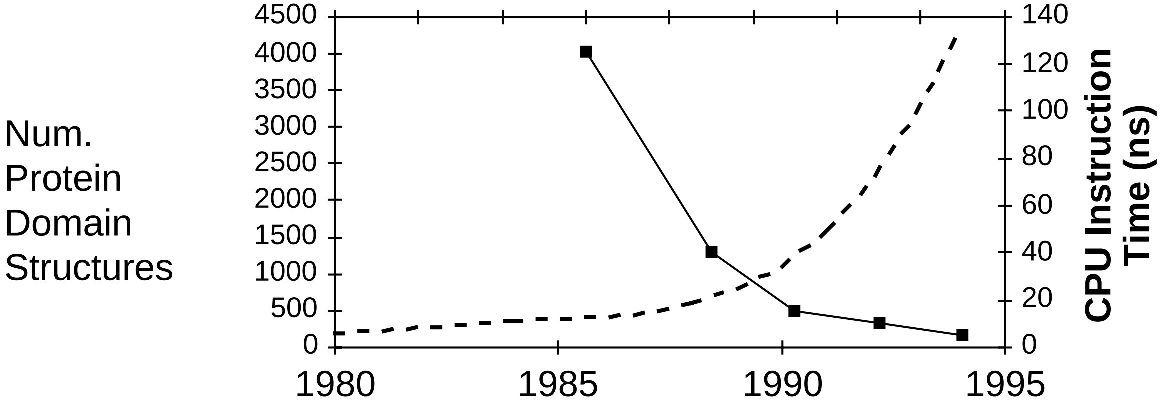
(idea from D Brutlag, Stanford)

Explonential Growth of Data Matched by Development of Computer Technology

- CPU vs Disk & Net
  - ◊ As important as the increase in computer speed has been, the ability to store large amounts of information on computers is even more crucial
- Driving Force in Bioinformatics

(Internet picture adapted from D Brutlag, Stanford)





New Paradigm for Scientific Computing

- Because of increase in data and improvement in computers, new calculations become possible
- But Bioinformatics has a new style of calculation...
  - ◊ Two Paradigms

- Physics
  - ◊ Prediction based on physical principles
  - ◊ Exact Determination of Rocket Trajectory
  - ◊ Supercomputer, CPU
- Biology
  - ◊ Classifying information and discovering unexpected relationships
  - ◊ globin ~ colicin~ plastocyanin~ repressor
  - ◊ networks, “federated” database

General Types of "Informatics" in Bioinformatics

- Databases
  - ◊ Building, Querying
  - ◊ Object DB
- Text String Comparison
  - ◊ Text Search
  - ◊ 1D Alignment
  - ◊ Significance Statistics
  - ◊ Alta Vista, grep
- Finding Patterns
  - ◊ AI / Machine Learning
  - ◊ Clustering
- Geometry
  - ◊ Robotics
  - ◊ Graphics (Surfaces, Volumes)
  - ◊ Comparison and 3D Matching (Visision, recognition)
- Physical Simulation
  - ◊ Newtonian Mechanics
  - ◊ Electrostatics
  - ◊ Numerical Algorithms
  - ◊ Simulation

Course Format

- New Field, Indisciplinary
  - ◊ No Universal Canon
  - ◊ No Universal Background
- Discussion, NOT Lecture Format
  - ◊ Theoretical Background, Ideas
  - ◊ Interactive
- Demos?
- Class Participation
  - ◊ Come to class! Ask questions
  - ◊ 1-2 Short Assignments Related to this
- Final Project
  - ◊ Critically Review an area
  - ◊ Critically Analyze Data
  - ◊ Propose a New Approach
  - ◊ Implement a New Approach
    - Computer Coding
  - ◊ Summarize an area in detail
    - Computer Implementation?

Background

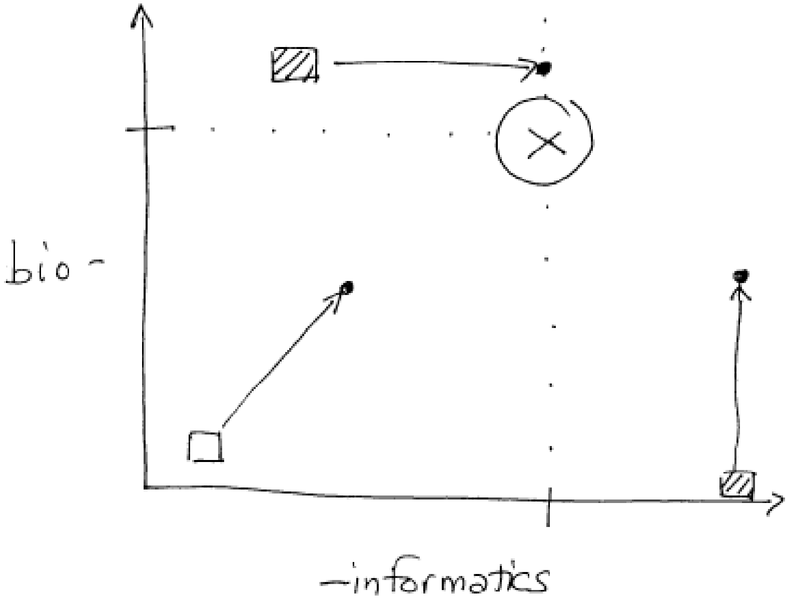
	Math	Biology
Need to Know Today	Calculation of Standard Deviation, a Bell-shaped Distribution (of test scores), a 3D vector	DNA, RNA, alpha-helix, the cell nucleus, ATP
What You'll Learn	Force is the Derivative (grad) of Energy, Rotation Matrices (3D), a P-value of .01 and an Extreme Value Distribution	Proteins are tightly packed, sequence homology twilight zone, protein families
Not Necessary at all	Poisson-Boltzman Equation, Design a Hashing Function, Write a Recursive Descent Parser	What GroEL does, a worm is a metazoa, E. coli is gram negative, what chemokines are

Computer Ability

- Course Website
  - ◊ http://bioinfo.mbb.yale.edu/course
  - ◊ http://bioinfo.mbb.yale.edu/course/survey.txt
  - ◊ Mark.Gerstein@yale.edu
  - ◊ Some Lectures at http://bioinfo.mbb.yale.edu/course/classes
- Will need to be able to read and write web pages
  - ◊ Read now, HTML “coding” will be explained later
  - ◊ Read PDF files via acrobat reader
    - http://www.adobe.com/prodindex/acrobat/readstep.html
  - ◊ Put Final Project on Course Website -- Project Gallery
- Programming in C or Perl **Optional**

Survey (Assignment 0!)

- http://bioinfo.mbb.yale.edu/course/survey.txt



Time Change

- Longer Format, Starting Next Monday

- Preferred:
  Monday and Wednesday 9:30-10:45,
  NOT Friday

- Also possible:
  Monday and Wednesday 9:05-10:20,
  NOT Friday

Limits

Specific Course Topics -- Sequences

- Sequence Alignment
  - ◊ non-exact string matching
  - ◊ How to align two strings optimally
  - ◊ via Dynamic Programming
  - ◊ Local vs Global Alignment
  - ◊ Hashing to increase speed (BLAST)
  - ◊ 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
  - ◊ HMMs, Profiles
- 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
- Structure “Prediction”
  - ◊ Secondary Structure Prediction, Propensities
  - ◊ TM-helix finding
  - ◊ The wall, why tertiary structure is so hard?
    - Fold Recognition
    - Threading

Course 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
- Structural Alignment
  - ◊ Aligning sequences on the basis of 3D structure.
  - ◊ DP does not converge, unlike sequences, what to do?
  - ◊ Other Approaches: Distance Matrices, Hashing
- Molecular Simulation
  - ◊ Geometry -> Energy -> Forces
  - ◊ Basic interactions, potential energy functions
  - ◊ How structure changes over time?
    - How to measure the change in a vector (gradient)
  - ◊ Molecular Dynamics & MC
  - ◊ Energy Minimization

Course Topics -- Databases

- Relational Database Concepts
  - ◊ 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)
- 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
- Genome Comparisons
  - ◊ Ortholog Families, pathways
  - ◊ Large-scale censuses
  - ◊ Frequent Words Analysis
  - ◊ Genome Annotation

Are They or Aren’t They Bioinformatics? (#1)

- Digital Libraries
  - ◊ Automated Bibliographic Search and Textual Comparison
  - ◊ Knowledge bases for biological literature
- Motif Discovery Using Gibb's Sampling
- Methods for Structure Determination
  - ◊ Computational Crystallography
    - Refinement
  - ◊ NMR Structure Determination
    - Distance Geometry
- Metabolic Pathway Simulation
- The DNA Computer

Are They or Aren't They Bioinformatics? (#1, Answers)

- **(YES)** Digital Libraries
  - ◊ Automated Bibliographic Search and Textual Comparison
  - ◊ Knowledge bases for biological literature
- **(YES)** Motif Discovery Using Gibb's Sampling
- **(NO)** Methods for Structure Determination
  - ◊ Computational Crystallography
    - Refinement
  - ◊ NMR Structure Determination
    - **(YES)** Distance Geometry
- **(YES)** Metabolic Pathway Simulation
- **(NO)** The DNA Computer

Are They or Aren’t They Bioinformatics? (#2)

- Gene identification by sequence inspection
  - ◊ Prediction of splice sites
- DNA methods in forensics
- Modeling of Populations of Organisms
  - ◊ Ecological Modeling
- Genomic Sequencing Methods
  - ◊ Assembling Contigs
  - ◊ Physical and genetic mapping
- Linkage Analysis
  - ◊ Linking specific genes to various traits

Are They or Aren’t They Bioinformatics? (#2, Answers)

- **(YES)** Gene identification by sequence inspection
  - ◊ Prediction of splice sites
- **(YES)** DNA methods in forensics
- **(NO)** Modeling of Populations of Organisms
  - ◊ Ecological Modeling
- **(NO?)** Genomic Sequencing Methods
  - ◊ Assembling Contigs
  - ◊ Physical and genetic mapping
- **(YES)** Linkage Analysis
  - ◊ Linking specific genes to various traits

Are They or Aren't They Bioinformatics? (#3)

- RNA structure prediction
  Identification in sequences
- Radiological Image Processing
  - ◊ Computational Representations for Human Anatomy (visible human)
- Artificial Life Simulations
  - ◊ Artificial Immunology / Computer Security
  - ◊ Genetic Algorithms in molecular biology
- Homology modeling
- Determination of Phylogenies Based on Non-molecular Organism Characteristics
- Computerized Diagnosis based on Genetic Analysis (Pedigrees)

Are They or Aren't They Bioinformatics? (#3, Answers)

- **(YES)** RNA structure prediction Identification in sequences
- **(NO)** Radiological Image Processing
  - ◊ Computational Representations for Human Anatomy (visible human)
- **(NO)** Artificial Life Simulations
  - ◊ Artificial Immunology / Computer Security
  - ◊ Genetic Algorithms in molecular biology
- **(YES)** Homology modeling
- **(NO)** Determination of Phylogenies Based on Non-molecular Organism Characteristics
- **(NO)** Computerized Diagnosis based on Genetic Analysis (Pedigrees)