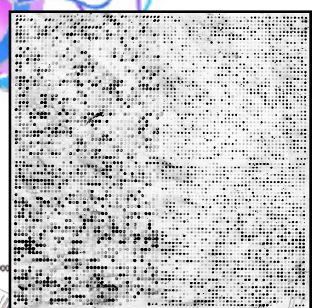
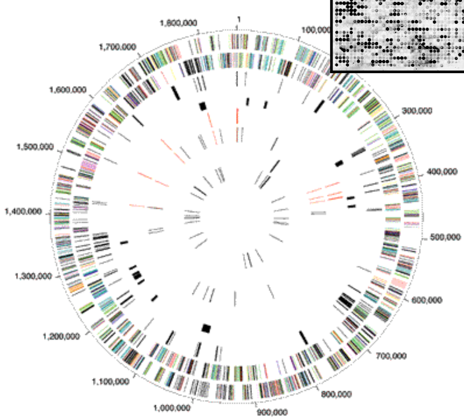
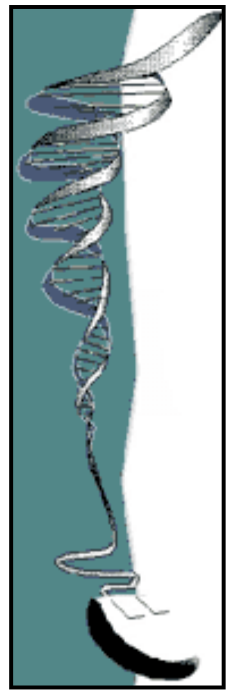
BIOINFORMATICS
Introduction





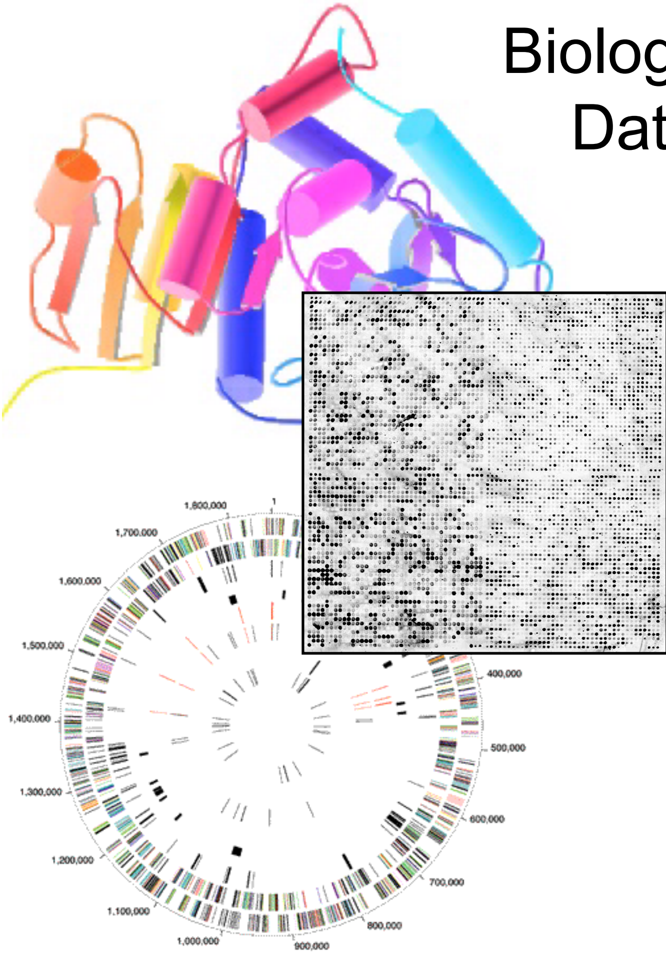


Mark Gerstein, Yale University

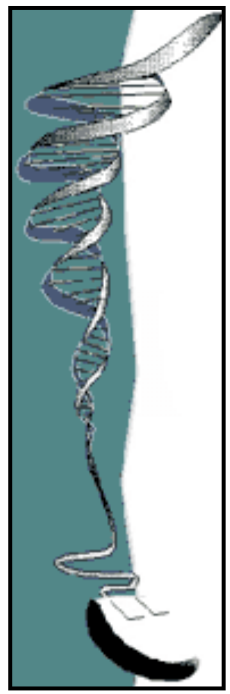
gersteinlab.org/courses/452

(last edit in spring '09, complete "in-class" changes included)

Bioinformatics



Biological Data + Computer Calculations



What is Bioinformatics?

Cor

- *(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**.

What is the Information? Molecular Biology as an Information Science

- Central Dogma of Molecular Biology

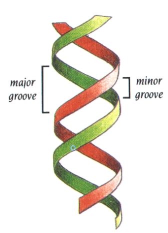
```
DNA
 -> RNA
  -> Protein
   -> Phenotype
    -> DNA
```

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

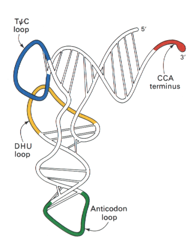
- Central Paradigm for Bioinformatics

```
Genomic Sequence Information
 -> mRNA (level)
  -> Protein Sequence
   -> Protein Structure
    -> Protein Function
     -> Phenotype
```

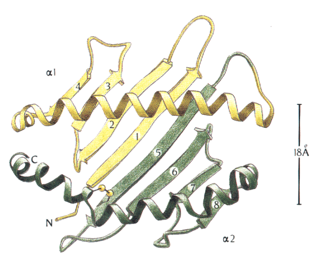
- Large Amounts of Information
  - ◊ Standardized
  - ◊ Statistical



•Genetic material



•Information transfer (mRNA)
•Protein synthesis (tRNA/mRNA)
•Some catalytic activity



•Most cellular functions are performed or facilitated by proteins.
•Primary biocatalyst
•Cofactor transport/storage
•Mechanical motion/support
•Immune protection
•Control of growth/differentiation

(idea from D Brutlag, Stanford, graphics from S Strobel)

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
  - ◊ ~3 Gb Human

```
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

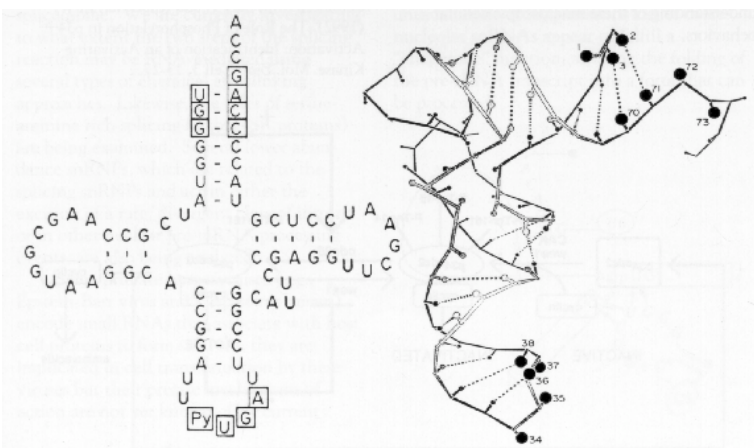
- 20 letter alphabet
  - ◊ ACDEFGHIKLMNPQRSTVWY but not BJOUXZ
- Strings of ~300 aa in an average protein (in bacteria), ~200 aa in a domain
- >1M known protein sequences (uniprot)

```
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
```

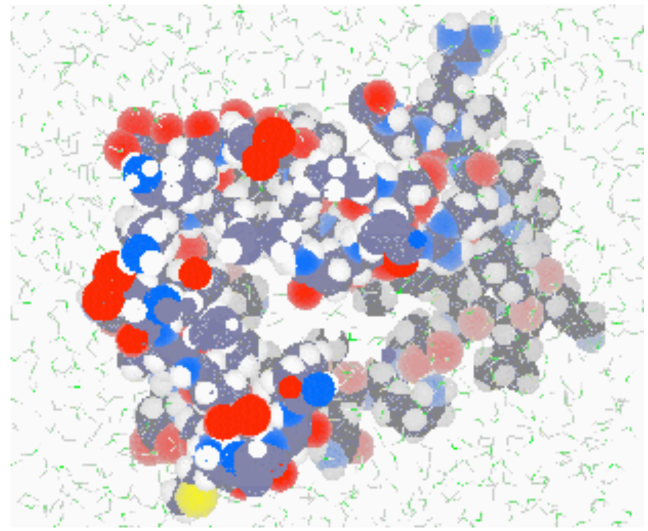

Molecular Biology Information: Macromolecular Structure

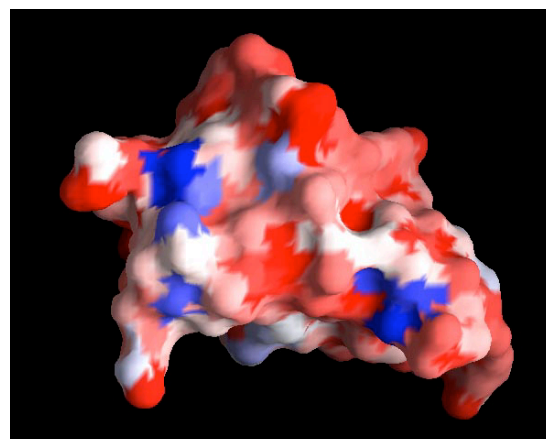
- DNA/RNA/Protein
  - ◊ Almost all protein

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



'Identity elements' in *Escherichia coli* glutamine tRNA.

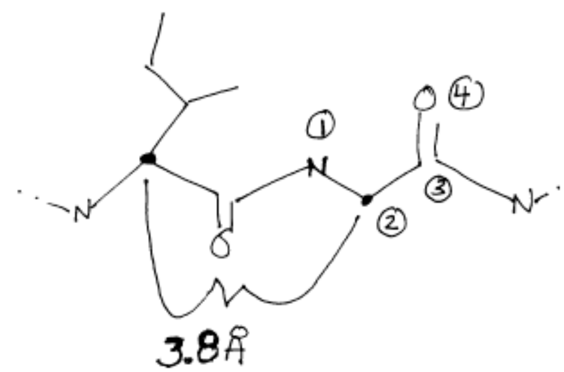




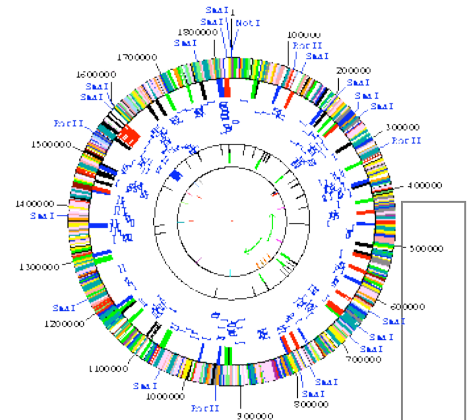
Molecular Biology Information: Protein Structure Details

- 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
  - ◊ >40K 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
```



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, worm: ~100Mb with 19 K genes

  1999: >30 completed genomes!

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

Genome sequence 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)

1995

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

1997

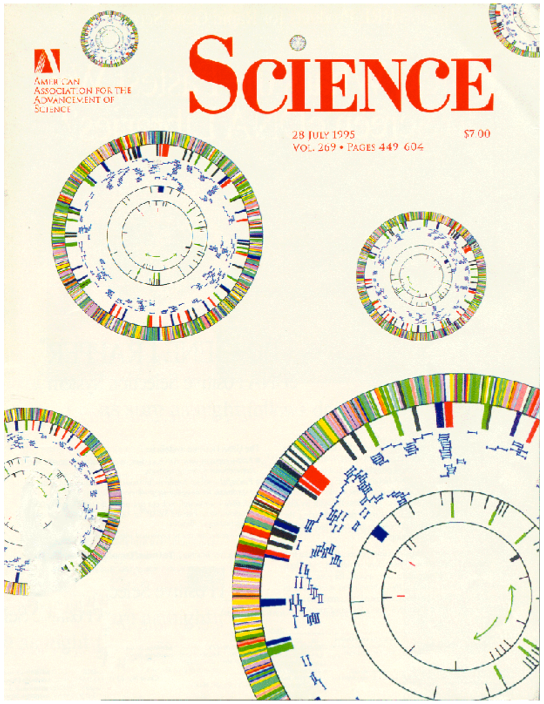
Eukaryote,
13 Mb,
~6K genes
[Nature **387**: 1]

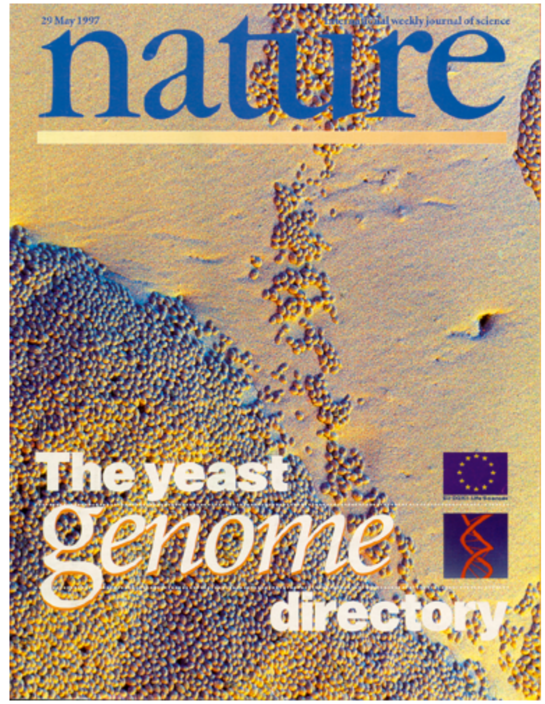
1998

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

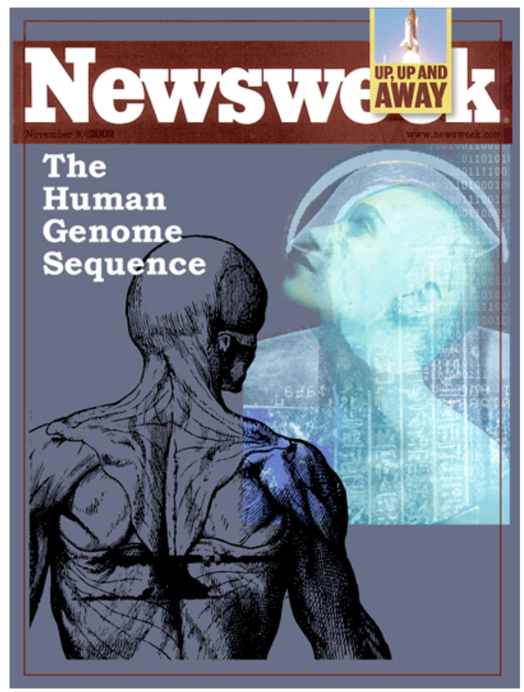
2000?

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





<u>Genomes highlight the</u> <u>**Finiteness**</u> <u>of the “Parts” in Biology</u>



real thing, Apr ‘00

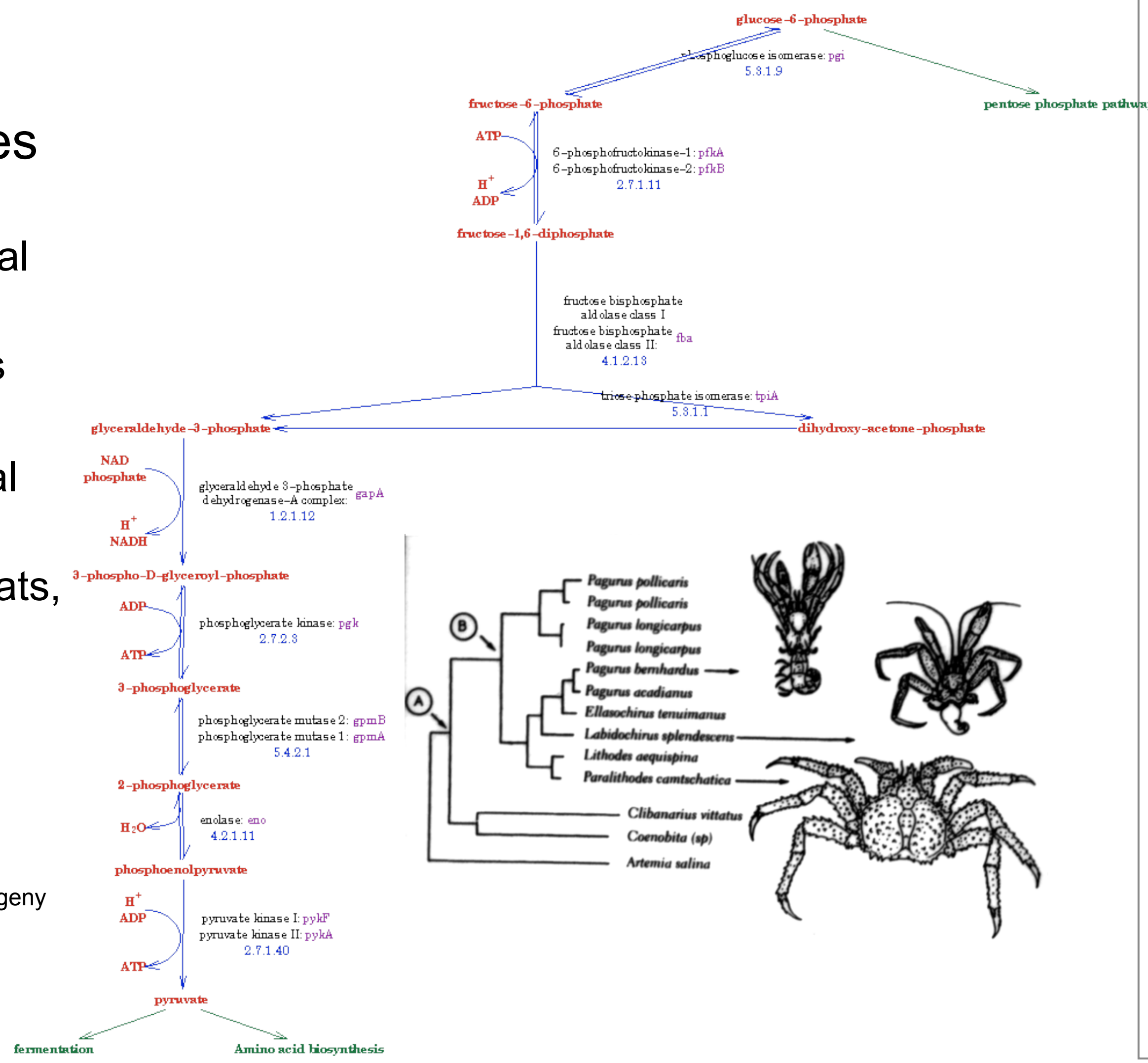
‘98 spoof

Other Types of Data

- Gene Expression
  - ◊ Early experiments yeast
    - Complexity at 10 time points,
      6000 x 10 = 60K floats
  - ◊ Now tiling array technology
    - 50 M data points to tile the human genome at ~50 bp res.
  - ◊ Can only sequence genome once but can do an infinite variety of array experiments
- Phenotype Experiments
  - ◊ Davis - KOs
  - ◊ Snyder - transposons
- Protein Interactions
  - ◊ For yeast: 6000 x 6000 / 2 ~ 18M possible interactions
  - ◊ maybe 30K real

Molecular Biology Information: Other Integrative Data

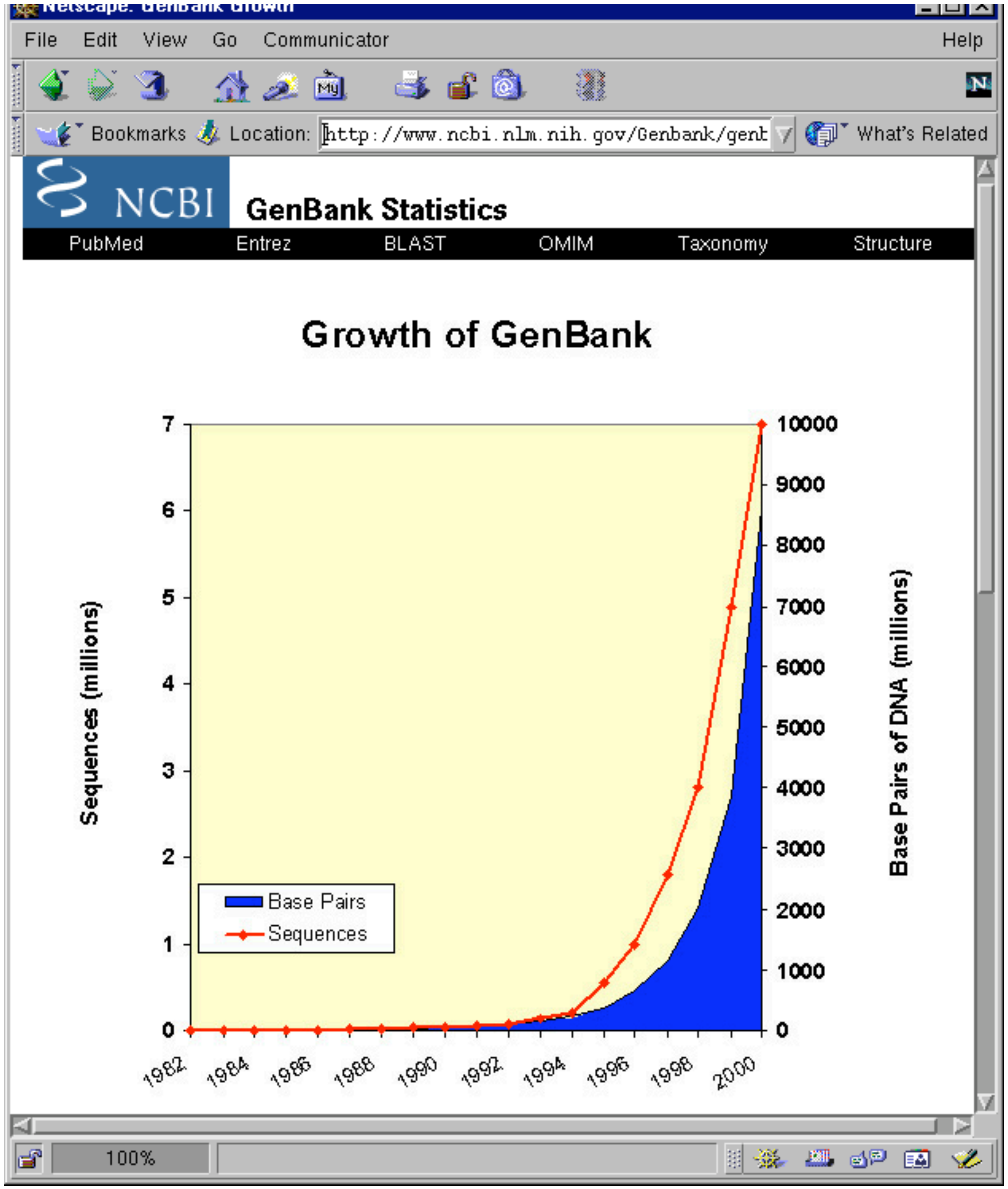
- Information to understand genomes
  - ◊ Metabolic Pathways (glycolysis), traditional biochemistry
  - ◊ Regulatory Networks
  - ◊ Whole Organisms Phylogeny, traditional zoology
  - ◊ Environments, Habitats, ecology
  - ◊ The Literature (MEDLINE)
- The Future....



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

What is Bioinformatics?

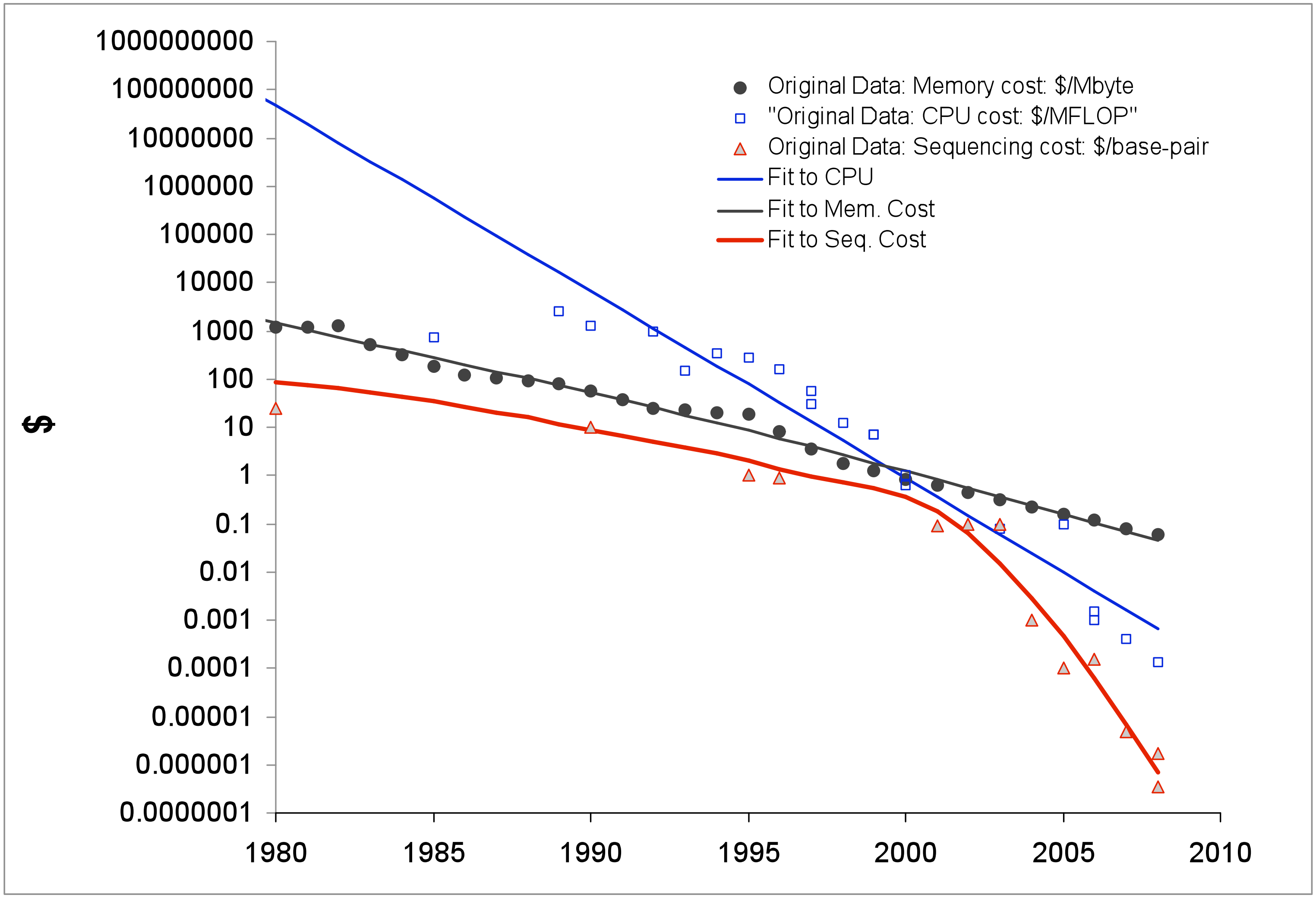
- *(Molecular)* **Bio** - **`informatics`**
- One idea for a definition?
  Bioinformatics is conceptualizing **biology in terms of molecules** (in the sense of physical-chemistry) and then applying **"informatics" techniques** (derived from disciplines such as applied math, CS, and statistics) to understand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



Large-scale Information: GenBank Growth

GenBank Data		
Year	**Base Pairs**	**Sequences**
1982	680338	606
1983	2274029	2427
1984	3368765	4175
1985	5204420	5700
1986	9615371	9978
1987	15514776	14584
1988	23800000	20579
1989	34762585	28791
1990	49179285	39533
1991	71947426	55627
1992	101008486	78608
1993	157152442	143492
1994	217102462	215273
1995	384939485	555694
1996	651972984	1021211
1997	1160300687	1765847
1998	2008761784	2837897
1999	3841163011	4864570
2000	8604221980	7077491

Plummeting Cost of Sequencing



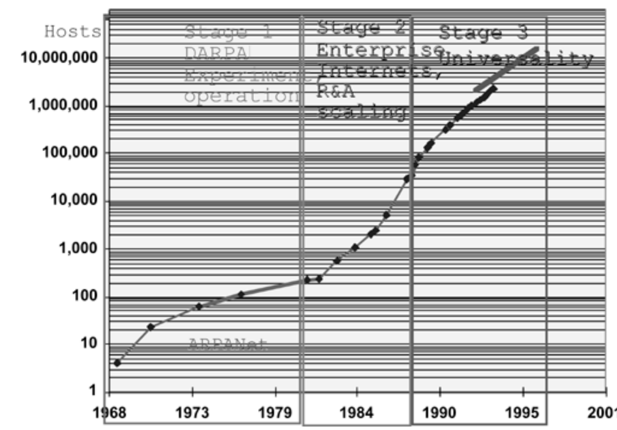
[Greenbaum et al., Am. J. Bioethics ('08)]

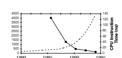
Large-scale Information: Explonential Growth of Data Matched by Development of Computer Technology

- 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)

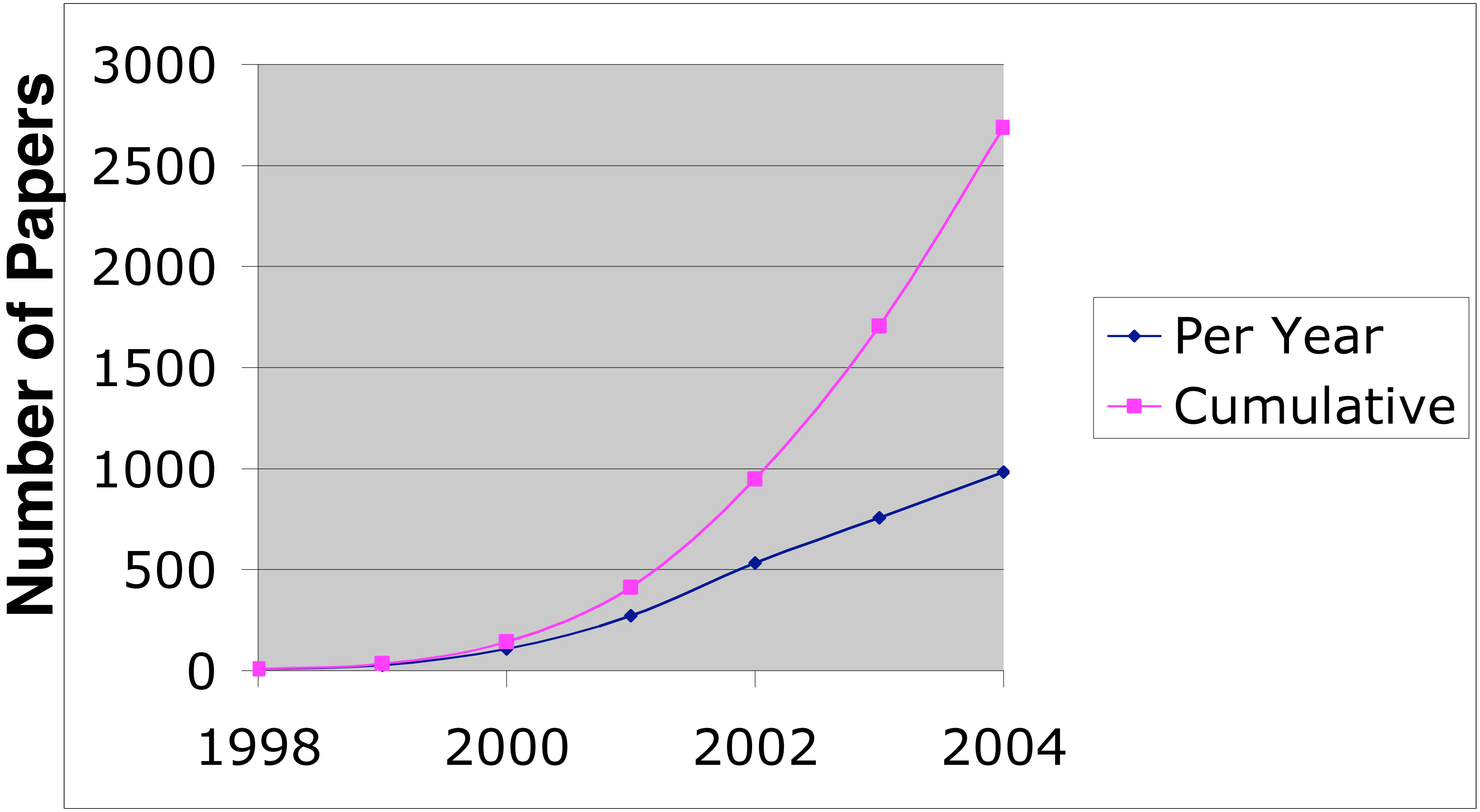
Internet Hosts



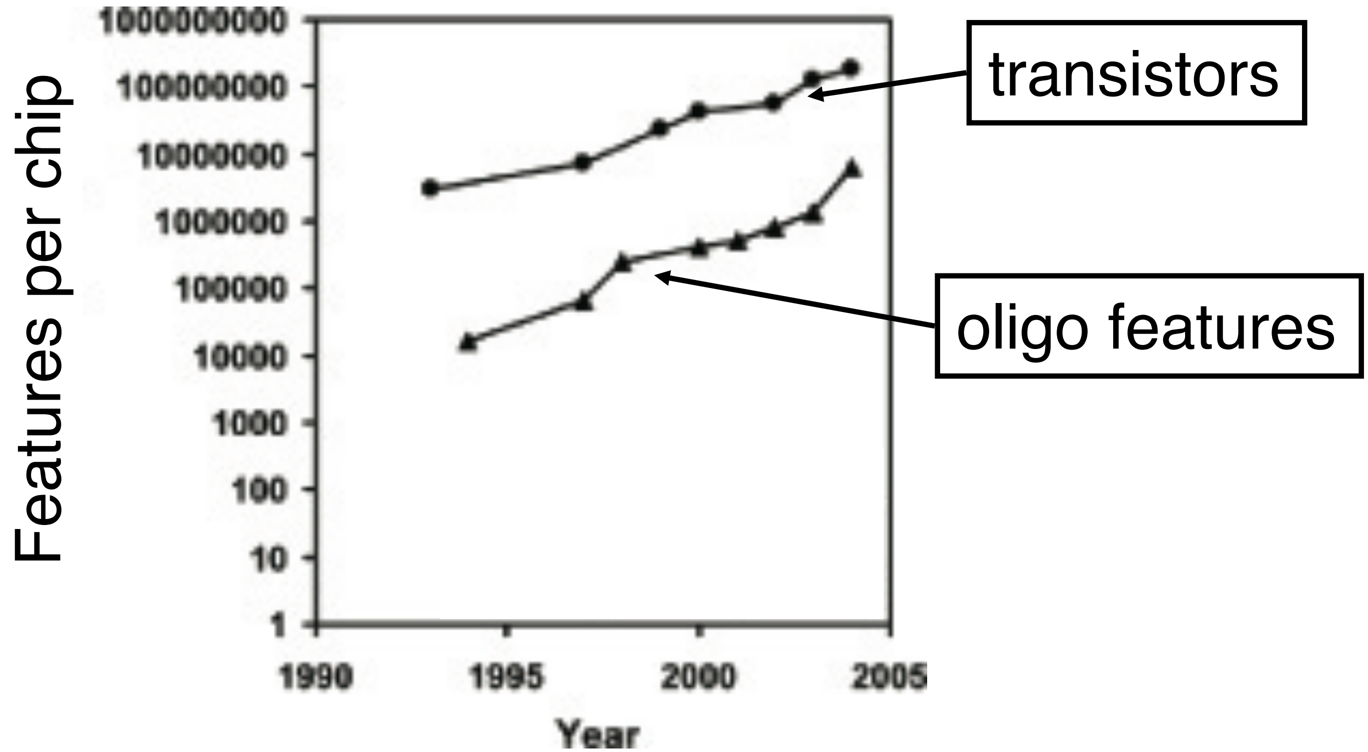


Num. Protein Domain Structures

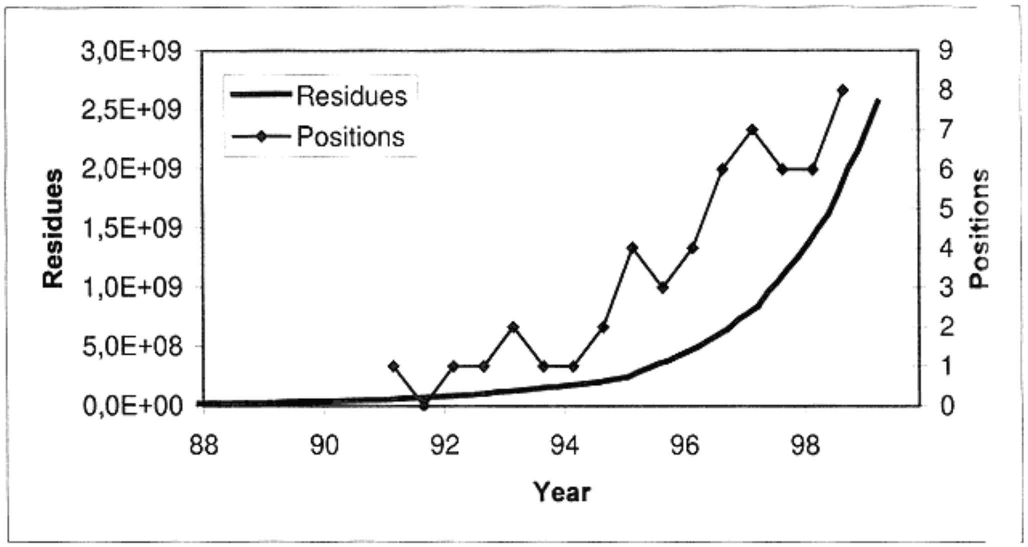
PubMed publications with title "microarray"



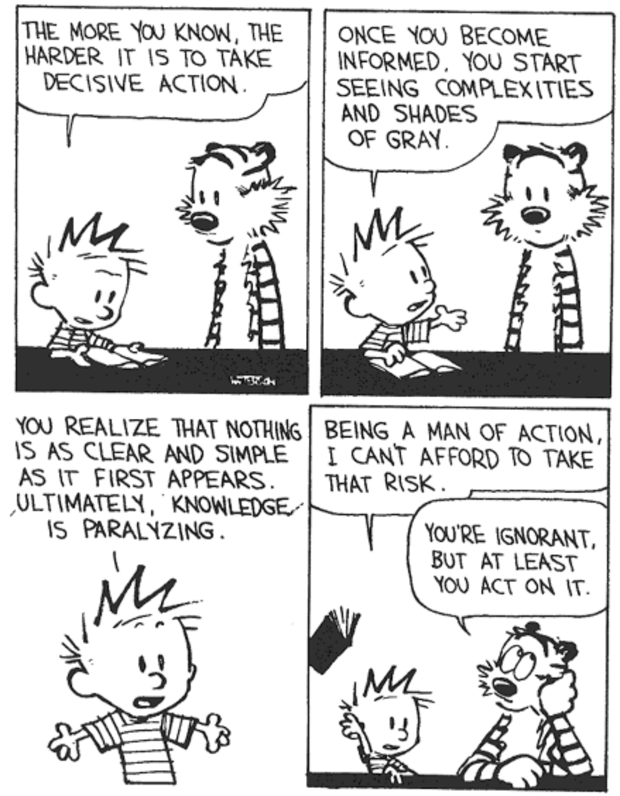
Features per Slide



Bioinformatics is born!



Growth in number of residues in Genbank, a central database for sequence data, compared to the request for people with competence in bioinformatics. The request for scientists is estimated from the number of relevant positions advertised in the first number of Nature in March and September of each year.



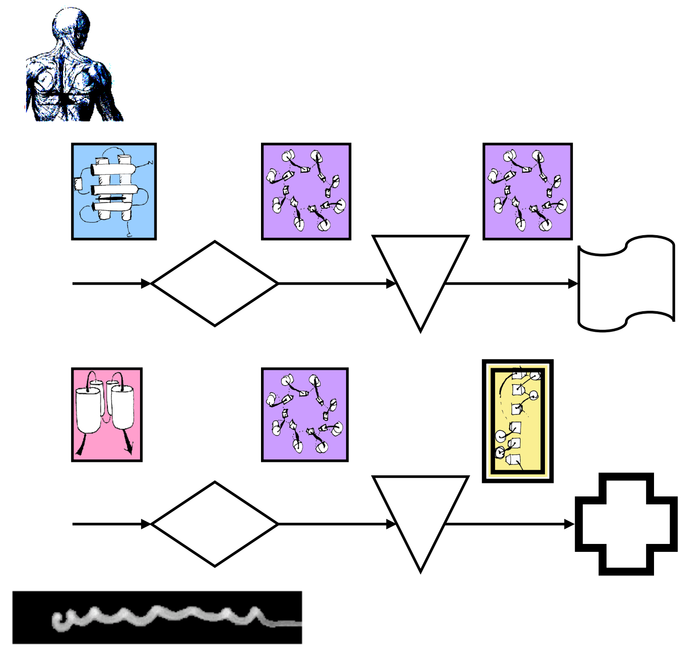
B. Watterson, "*There's treasure everywhere*", Andrews and McMeel, 1996.

(courtesy of Finn Drablos)

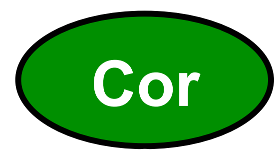
What is Bioinformatics?

- *(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**.

Organizing
Molecular Biology Information: Redundancy and Multiplicity

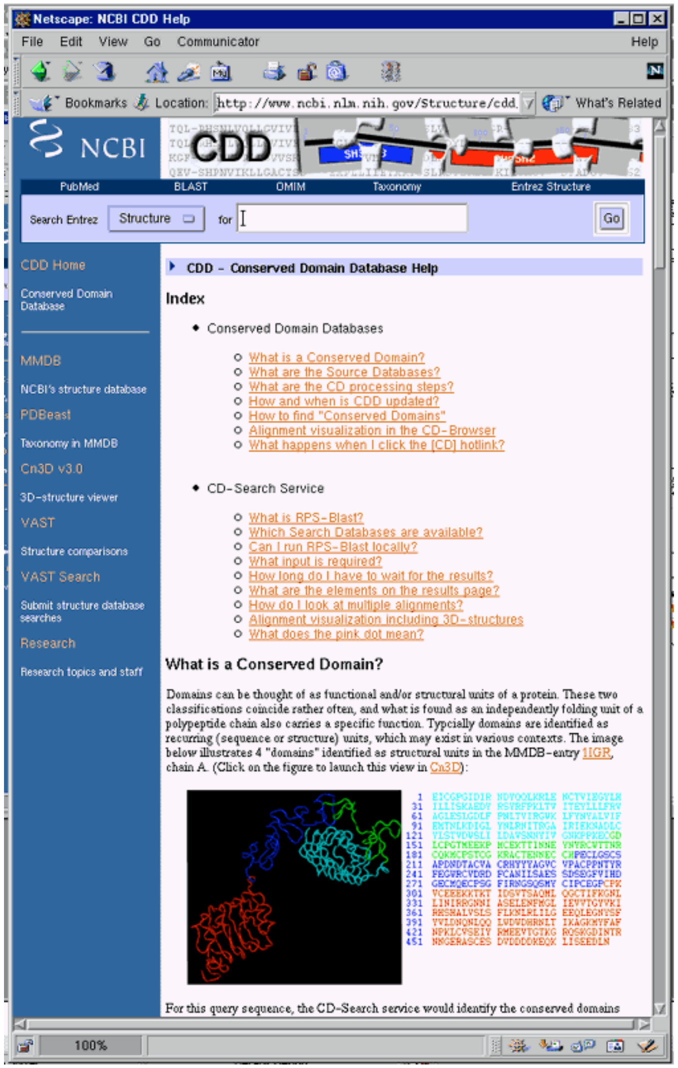


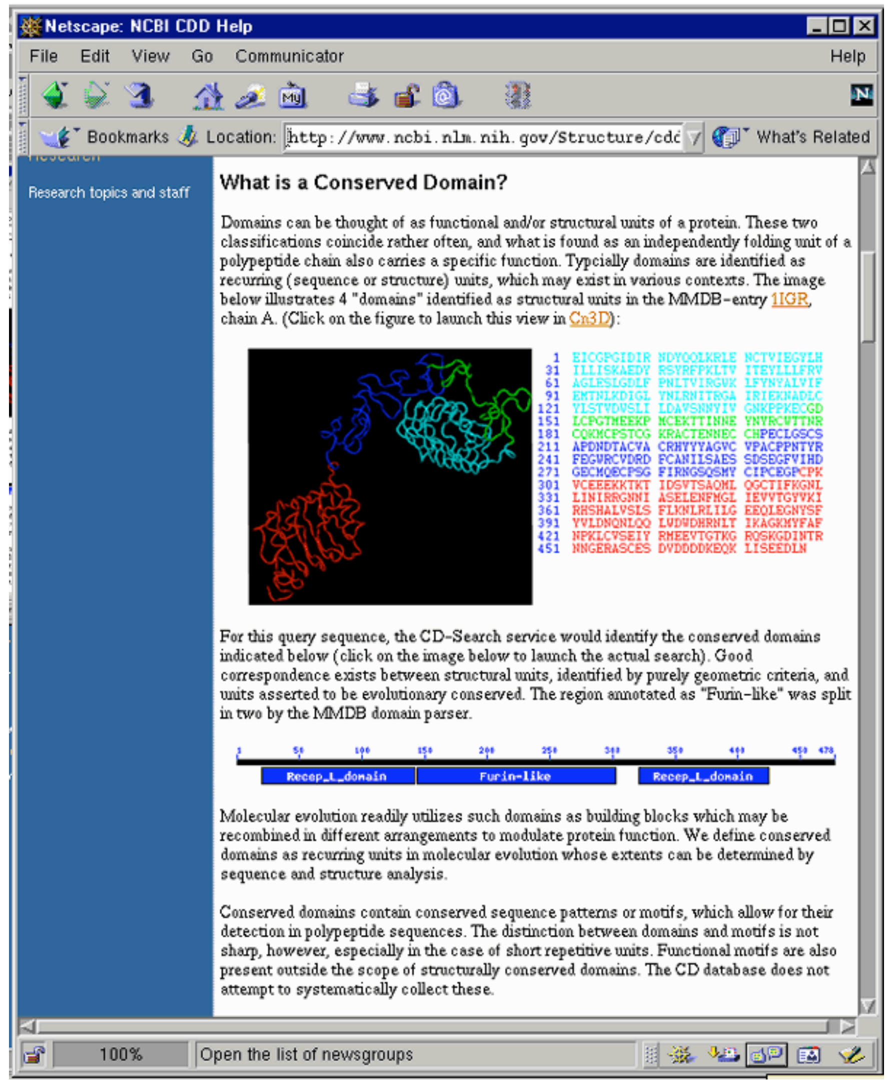
- Different Sequences Have the Same Structure
- Organism has many similar genes
- Single Gene May Have Multiple Functions
- Genes are grouped into Pathway & Networks
- Genomic Sequence Redundancy due to the Genetic Code
- **How do we find the similarities?.....**



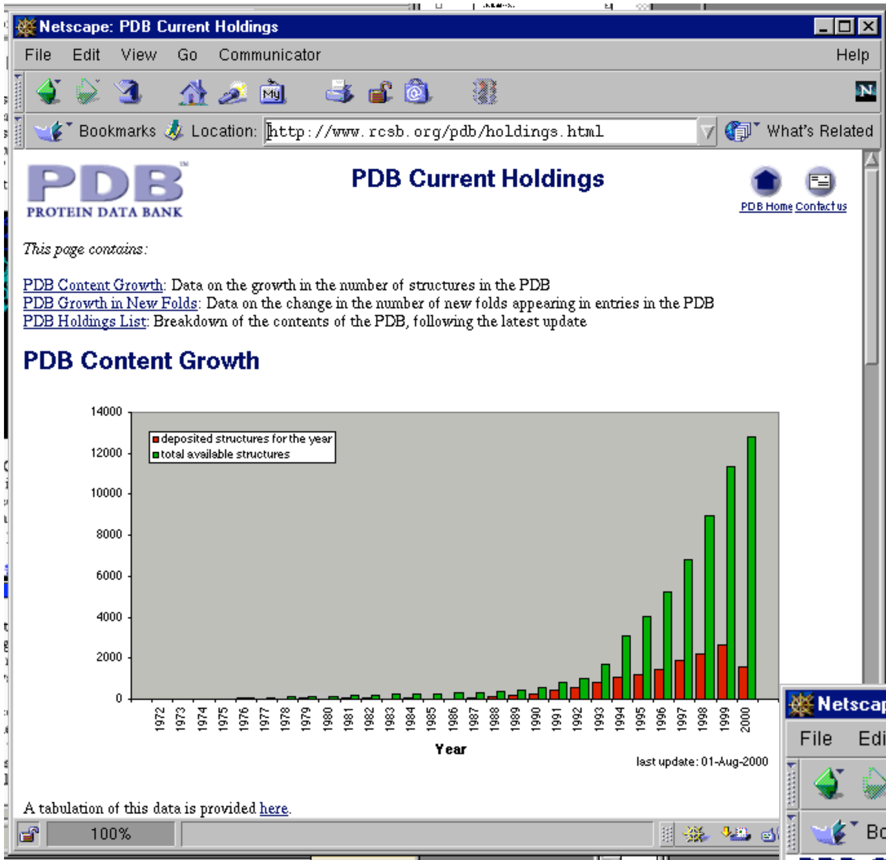
Integrative Genomics -
genes ↔ structures ↔
functions ↔ **pathways** ↔
expression levels ↔
regulatory systems ↔ ….

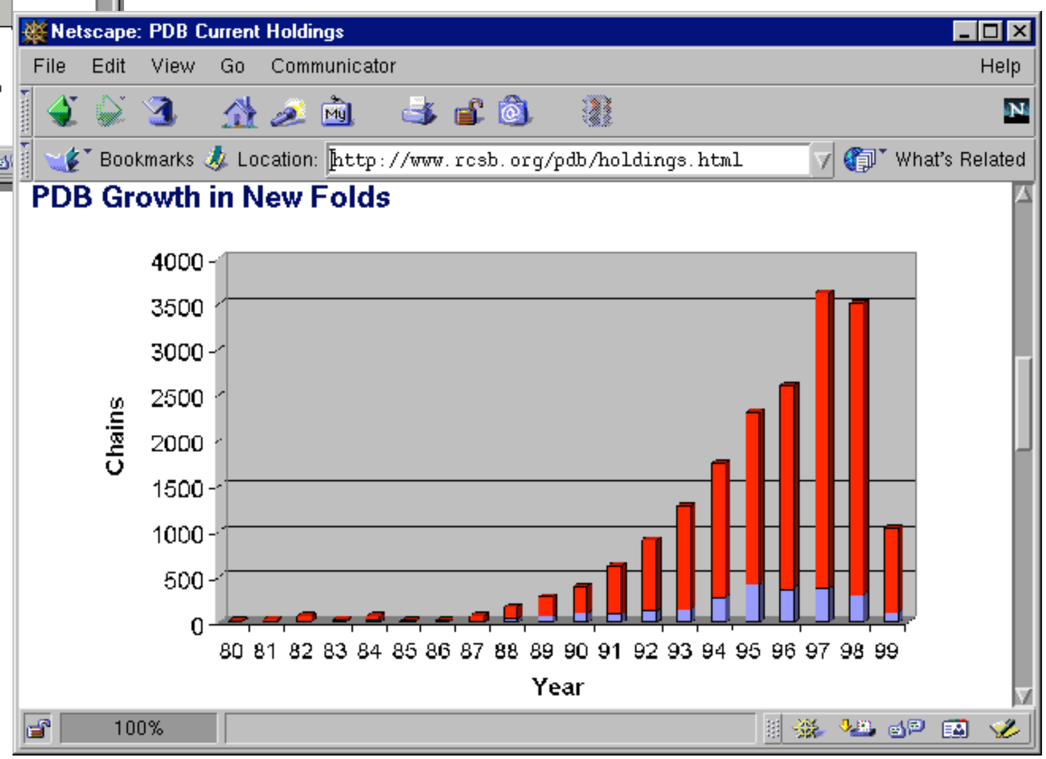
Molecular Parts = Conserved Domains, Folds, &c





Vast Growth in (Structural) Data... but number of Fundamentally New (Fold) Parts Not Increasing that Fast





Total in Databank

New Submissions

New Folds

What is Bioinformatics?

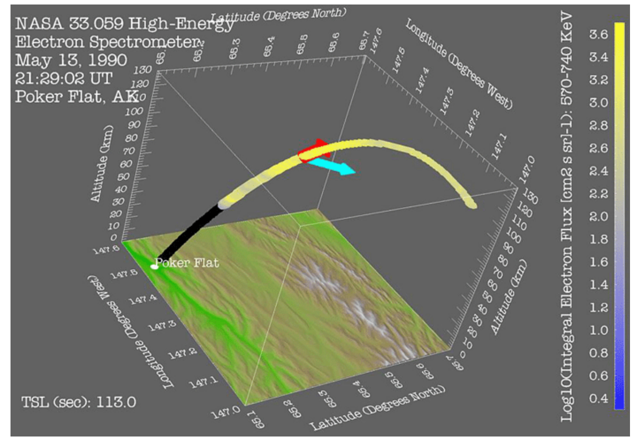
- *(Molecular)* **Bio** - **`informatics`**
- One idea for a definition?
  Bioinformatics is conceptualizing **biology in terms of molecules** (in the sense of physical-chemistry) and then applying **"informatics" techniques** (derived from disciplines such as applied math, CS, and statistics) to understand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

General Types of "Informatics" techniques in Bioinformatics

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

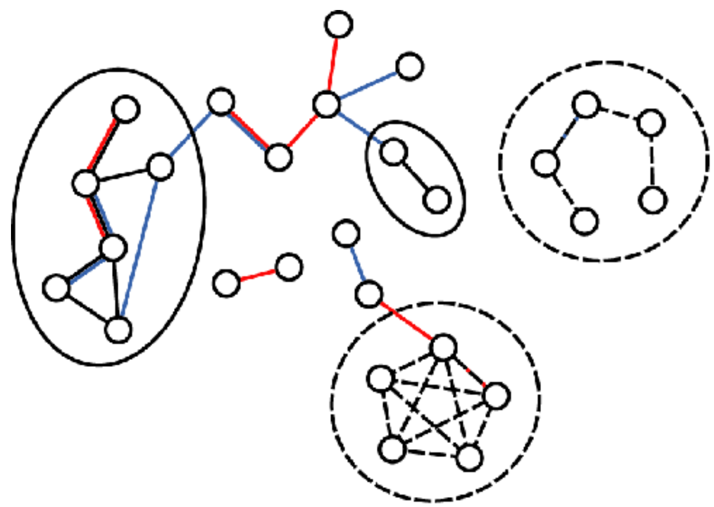
Bioinformatics as New Paradigm for Scientific Computing

- Physics
  - ◊ Prediction based on physical principles
  - ◊ EX: Exact Determination of Rocket Trajectory
  - ◊ Emphasizes: Supercomputer, CPU



Cor

- Biology
  - ◊ Classifying information and discovering unexpected relationships
  - ◊ EX: Gene Expression Network
  - ◊ Emphasizes: networks, “federated” database



Statistical Analysis vs. Classical Physics

Bioinformatics, Genomic Surveys

Vs.

Chemical Understanding, Mechanism, Molecular Biology

How Does Prediction Fit into the Definition?

Bioinformatics Topics -- Genome Sequence

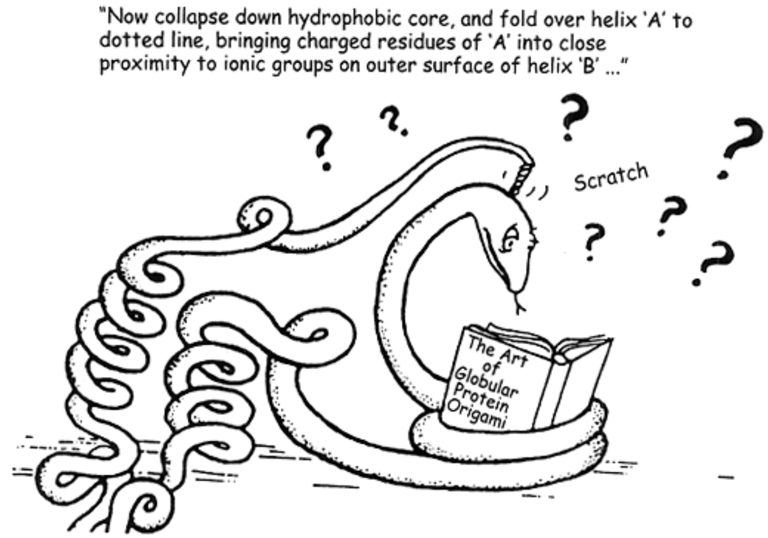
- Finding Genes in Genomic DNA
  - ◊ introns
  - ◊ exons
  - ◊ promotors
- Characterizing Repeats in Genomic DNA
  - ◊ Statistics
  - ◊ Patterns
- Duplications in the Genome
  - ◊ Large scale genomic alignment
- Whole-Genome Comparisons
- Finding Structural RNAs

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

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 v RNA



Reproduced in U. Tollemar, "*Protein Engineering i USA*", Sveriges Tekniska Attachéer, 1988

- Tertiary Structure Prediction
  - ◊ Fold Recognition
  - ◊ Threading
  - ◊ Ab initio
  - ◊ (Quaternary structure prediction)
- Direct Function Prediction
  - ◊ Active site identification
- Relation of Sequence Similarity to Structural Similarity

Topics -- Structures

- Structure Comparison
  - ◊ 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
  - ◊ Hinge prediction
  - ◊ 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
- Docking and Drug Design as Surface Matching

Topics – DBs/ Surveys

- 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
- DB interoperation
- What are the Units ?
  - ◊ What are the units of biological information for organization?
    - 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

Mining

- Information integration and fusion
  - ◊ Dealing with heterogeneous data
- Dimensionality Reduction (PCA etc)

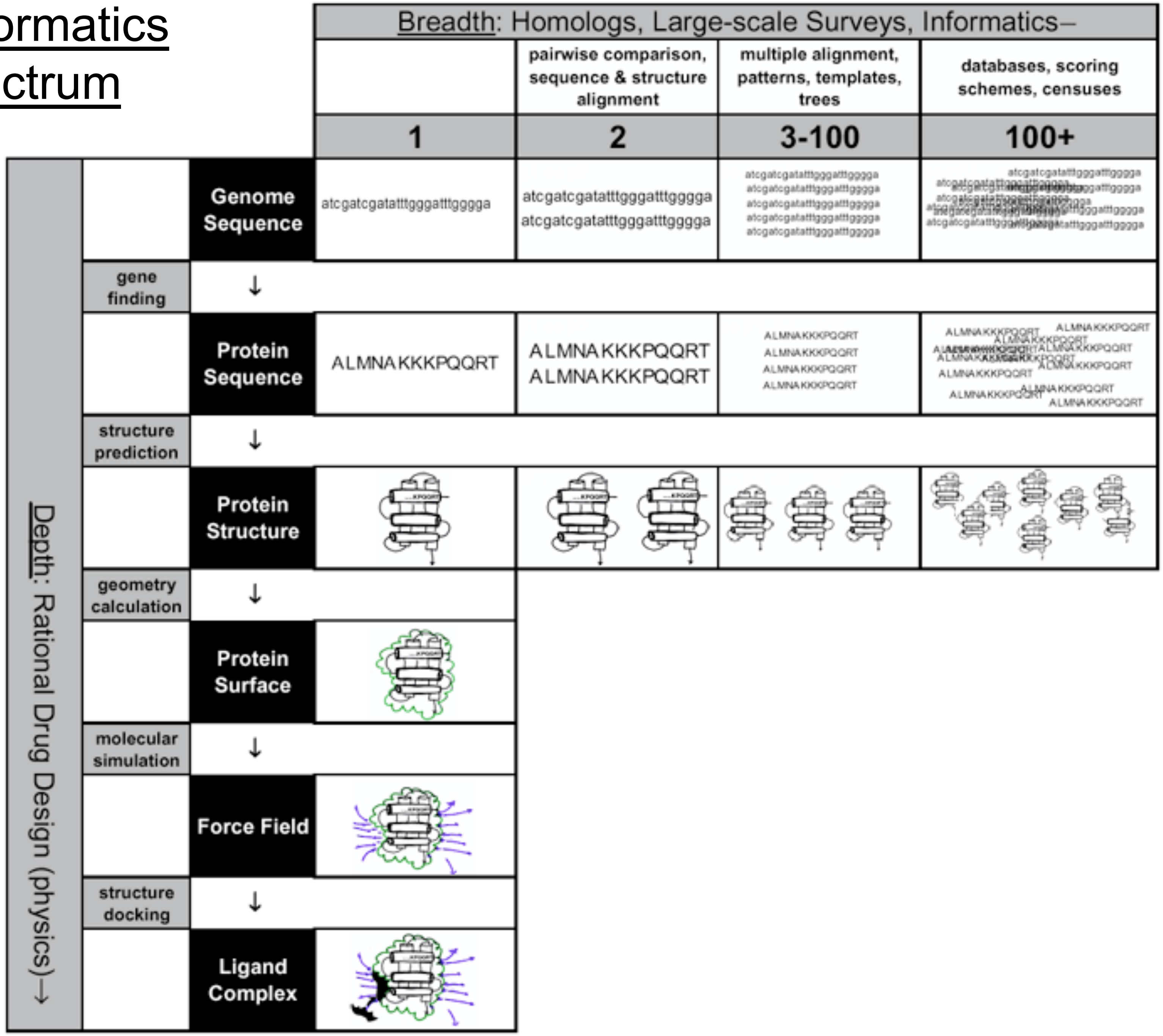
Topics – (Func) Genomics

- 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
  - ◊ Identification of interacting proteins
- Networks
  - ◊ Global structure and local motifs
- Structural Genomics
  - ◊ Folds in Genomes, shared & common folds
  - ◊ Bulk Structure Prediction
- Genome Trees

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
- Simplifications
  - ◊ Poisson-Boltzman Equation
  - ◊ Lattice Models and Simplification

Bioinformatics Spectrum



What is Bioinformatics?

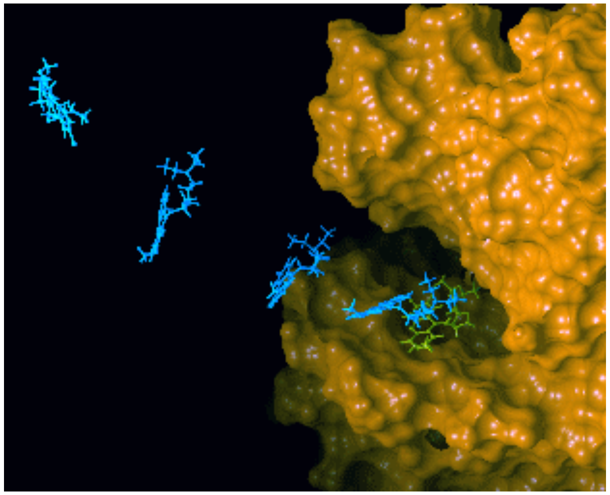
- *(Molecular)* **Bio** - **`informatics`**
- One idea for a definition?
  Bioinformatics is conceptualizing **biology in terms of molecules** (in the sense of physical-chemistry) and then applying **"informatics" techniques** (derived from disciplines such as applied math, CS, and statistics) to understand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

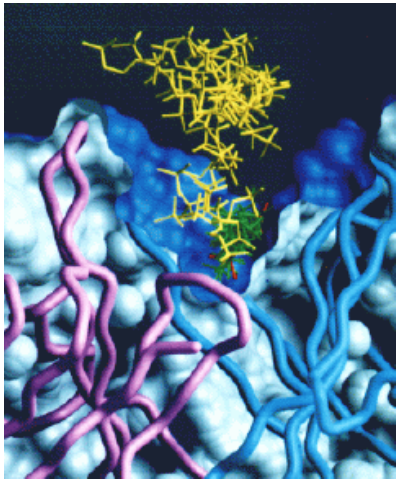
Major Application I: Designing Drugs

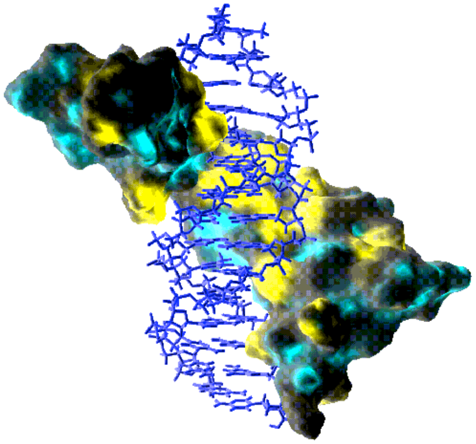
Cor

- Understanding How Structures Bind Other Molecules (Function)
- Designing Inhibitors
- Docking, Structure Modeling

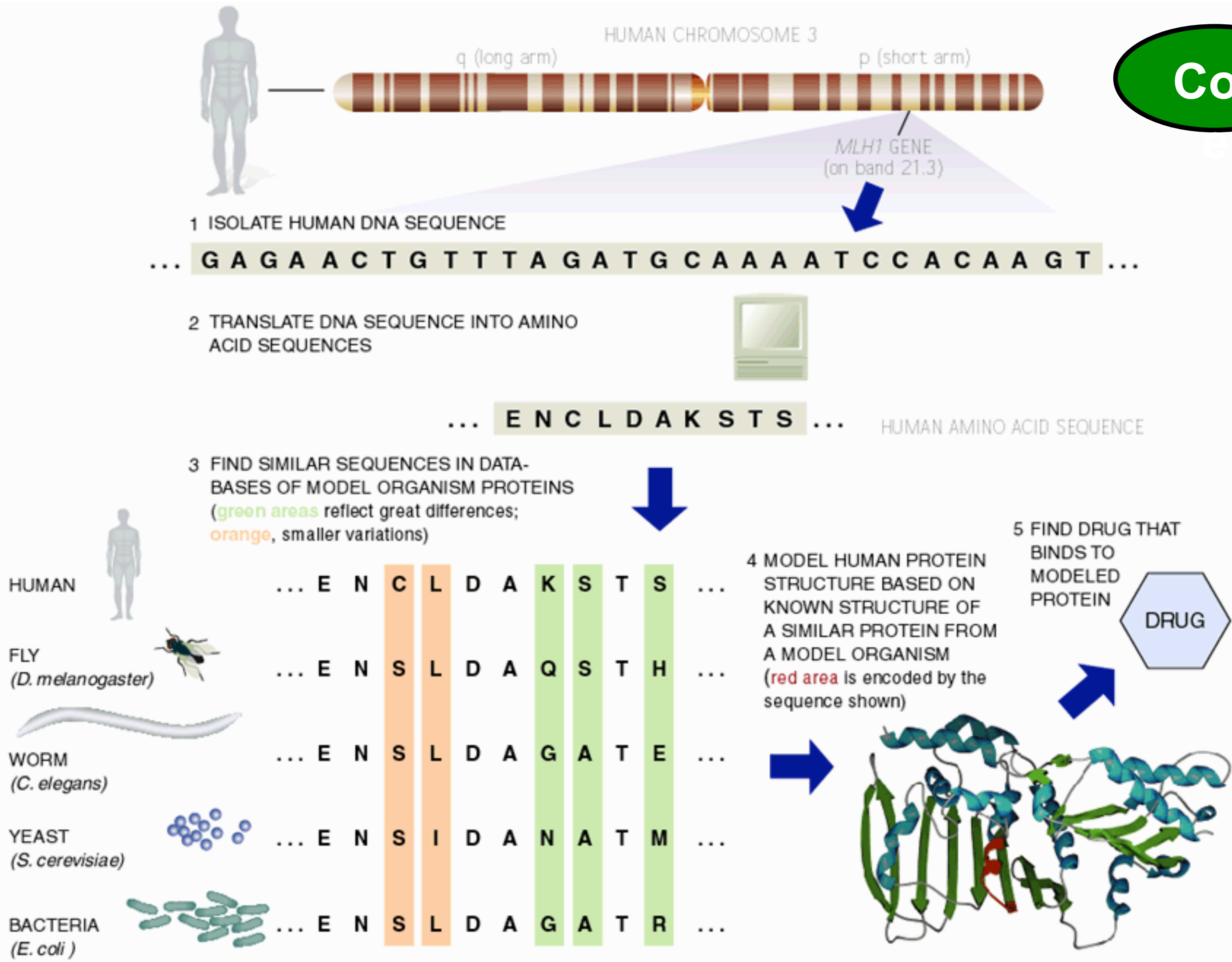
(From left to right, figures adapted from Olsen Group Docking Page at Scripps, Dyson NMR Group Web page at Scripps, and from Computational Chemistry Page at Cornell Theory Center).







Major Application II: Finding Homologs



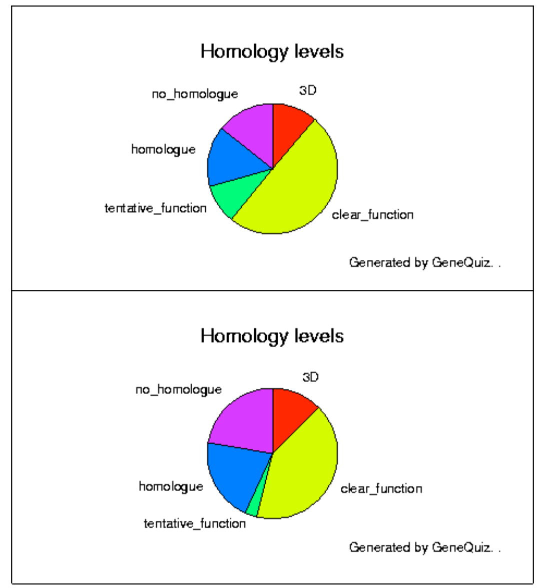
Cor

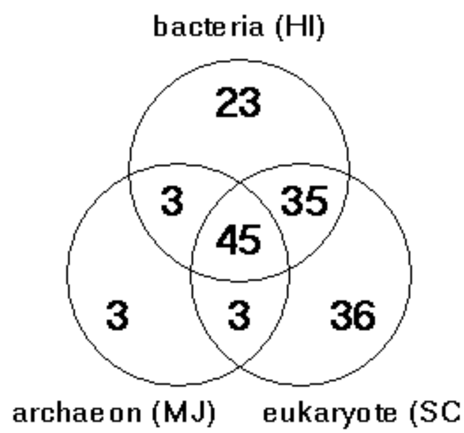
Cor

Major Application I|I: Overall Genome Characterization

- Overall Occurrence of a Certain Feature in the Genome
  - ◊ e.g. how many kinases in Yeast
- Compare Organisms and Tissues
  - ◊ Expression levels in Cancerous vs Normal Tissues
- Databases, Statistics

(Clock figures, yeast v. Synechocystis,
adapted from GeneQuiz Web Page, Sander Group, EBI)





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**.

Defining the Boundaries of the Field

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

- Digital Libraries
  - ◊ Automated Bibliographic Search of the biological literature 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
  - ◊ **(NO?)** 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)

Further Thoughts in 2005 on the "Boundary of Bioinformatics"

- Issues that were uncovered
  - ◊ Does topic stand alone?
  - ◊ Is bioinformatics acting as tool?
  - ◊ How does it relate to lab work?
  - ◊ **Prediction**?
- Relationship to other disciplines
  - ◊ Medical informatics
  - ◊ Genomics and Comp. Bioinformatics
  - ◊ Systems biology
- Biological question is important, not the specific technique -- but it has to be computational
  - ◊ Using computers to understand biology vs using biology to inspire computation
- Some new ones (2005)
  - ◊ Disease modeling [are you modeling molecules?]
  - ◊ Enzymology (kinetics and rates?) [is it a simulation or is it interpreting 1 expt.?]
  - ◊ Genetic algs used in gene finding HMMs used in gene finding
    - vs. Genetic algs used in speech recognition HMMs used in speech recognition
  - ◊ Semantic web used for representing biological information

Some Further Boundary Examples in 2006

- Char. drugs and other small molecules (cheminformatics or bioinformatics?) [YES]
- Molecular phenotype discovery – looking for gene expression signatures of cancer [YES]
  - ◊ What if it incluced non-molecular data such as age ?
- Use of whole genome sequences to create phylogenies [YES]
- Integration and organization of biological databases [YES]

Defining the Core of the Field

What is Core Bioinformatics

- Core Stuff
  - ◊ Computing with sequences and structures
  - ◊ protein structure prediction
  - ◊ biological databases and mining them
- New Stuff: Networks and Expression Analysis
- Fairly Speculative: simulating cells

