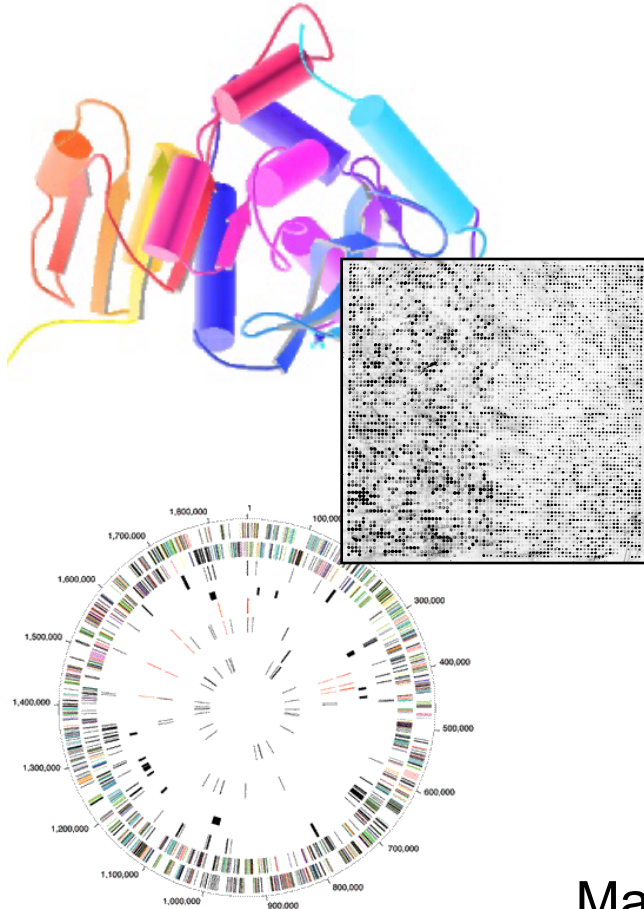


# BIOINFORMATICS

## Introduction



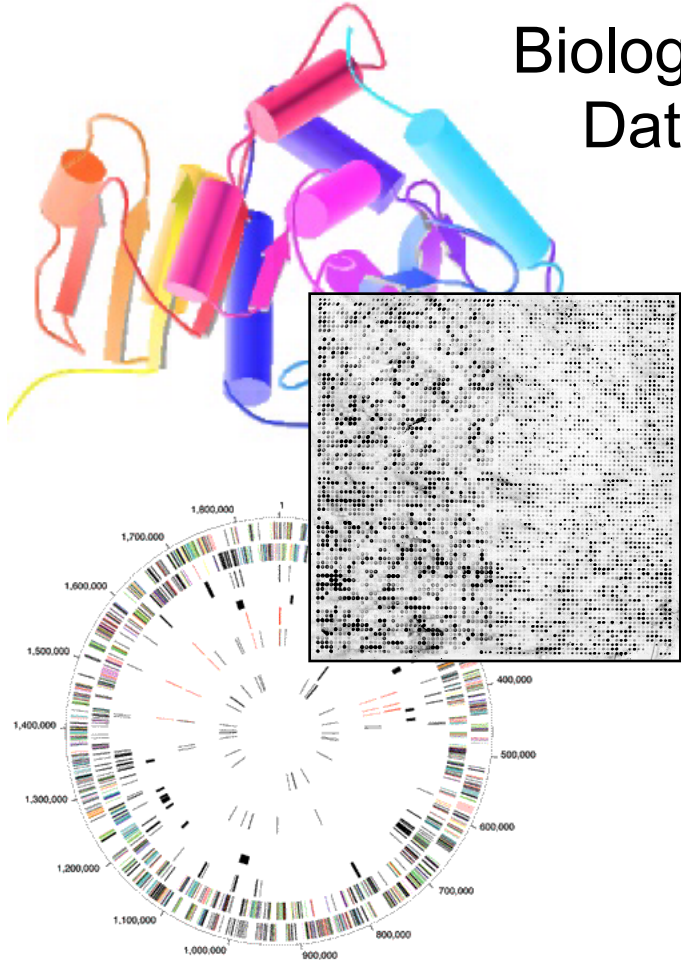
Mark Gerstein, Yale University  
[gersteinlab.org/courses/452](http://gersteinlab.org/courses/452)  
(last edit in spring '11)

# 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

- Central Paradigm for Bioinformatics

Genomic Sequence Information

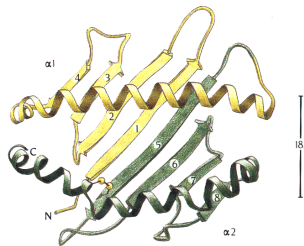
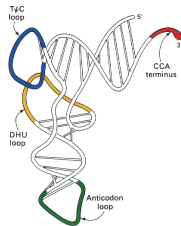
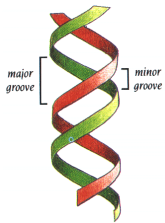
-> mRNA (level)

-> Protein Sequence

-> Protein Structure

-> Protein Function

-> Phenotype



- Genetic material

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

# 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

```
atggcaattaaattggtatcaatggtttggcgtatcgccgtatcgtattccgtgca
gcacaacaccgtgatgacattgaagttgtaggtattaacgacttaatcgacggtgaatac
atggcttataatggtgaaatagattcaactcaccggtcgtttcgacggcactggtgaagt
aaagatggtaacttagtggtaaatggtaaaactatccgtgtaactgcagaacgtgatcca
gcaacttaaaactggggtgcaatcgggttgatcgcctggtgaagcgcactggtttattc
ttaaactgatgaaactgctcgtaaacatatacactgcaggcgcaaaaaaagtgtattaact
ggcccatctaaagatgcaaccctatgttcggttcggtgtaaaacttcaacgcatacgca
ggtcaagatacgtttctaacgcactctgtacaacaaactgtttagctccttagcactg
gttggtcatgaaactttcggtatcaaagatggtttaatgaccactgttcacgcaacgact
gcaactcaaaaaactgtggatgggtccatcagctaaagactggcgcgcccggcgggtgca
tcacaaaacatcattccatcttcaacaggtgcagcgaagcagtaggtaaaagtattacct
gcattaaacggtaaatctaactggatggctttccgtgttccaacgccaacgtatctgtt
gttgatttaacagttaatcttgaaaaaccagcttcttatgatgcaatcaaaacaagcaatc
aaagatgcagcgggaaggtaaaacgttcaatggcgaattaaaaggcgtattaggttacct
gaagatgctgttgttttctactgacttcaacgggttgctttaaacttctgtatttgatgca
gacgctggtatcgcatctaactgattccttccgttaaatggatc . . .
```

```
. . . caaaaatagggttaatatgaatctcgatctccatthttgttcacgtattcaa
caacaagccaaaactcgtacaaatatgaccgcacttcgctataaagaacacggcttgtgg
cgagatatctcttgaaaaactttcaagagcaactcaatcaactttctcgagcattgctt
gctcacaatattgacgtacaagataaaaatcgccatthttgcccataatatggaacgttgg
gttggtcatgaaactttcggtatcaaagatggtttaatgaccactgttcacgcaacgact
acaatcgttgacattgcgaccttacaattcgagcaatcacagtgccattttacgcaacc
aatacagcccagcaagcagaatthtccctaaatcacgccgatgtaaaaattctcttctcgc
ggcgtacaagagcaatacgcatacaacattggaaattgctcatcattgtccaaaattacaa
aaaattgtagcaatgaaatccaccattcaattacaacaagatcctcttcttgcacttgg
```

# 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
- >12 M known protein sequences  
(uniprot, <http://www.ebi.ac.uk/uniprot/TrEMBLstats/>, 2011)

```
d1dhfa_ LNCIVAVSQNMGIGKNGDLPWPPLRNEFRYFQRMTTTTSSVEGKQ-NLVIMGKKTWFSI
d8dfr_  LNSIVAVCQNMGIGKDGNLPPWRNEFYKQFQRMSTSTSHVEGKQ-NAVIMGKKTWFSI
d4dfra_ ISLIAALAVDRVIGMENAMPWN-LPADLAWFKRNTL-----NKPVIMGRHTWESI
d3dfr_  TAFLWAQDRDGLIGKDGHLPHW-LPDDLHYFRAQTV-----GKIMVVGRRTYESF
```

```
d1dhfa_ LNCIVAVSQNMGIGKNGDLPWPPLRNEFRYFQRMTTTTSSVEGKQ-NLVIMGKKTWFSI
d8dfr_  LNSIVAVCQNMGIGKDGNLPPWRNEFYKQFQRMSTSTSHVEGKQ-NAVIMGKKTWFSI
d4dfra_ ISLIAALAVDRVIGMENAMPW-NLPADLAWFKRNTLD-----KPVIMGRHTWESI
d3dfr_  TAFLWAQDRNGLIGKDGHLPHW-HLPDDLHYFRAQTVG-----KIMVVGRRTYESF
```

```
d1dhfa_ VPEKNRPLKGRINLVLSRELKEPPQGAHFLSRSLDDALKLTEQPELANKVDMVWIVGGSSVYKEAMNHP
d8dfr_  VPEKNRPLKDRINIVLSRELKEAPKGAHYLSKSLDDALALLDSPELKSQVDMVWIVGGTAVYKAAMEKP
d4dfra_ ---G-RPLPGRKNIILS-SQPGTDDRVTWVKSVDIAAIAACGDVPE-----EIMVIGGGRVYEQFLPKA
d3dfr_  ---PKRPLPERTNVVLT HQEDYQAQGA-VVVHDVAAVFAYAKQHLDQ----ELVIAGGAQIFTAFKDDV
```

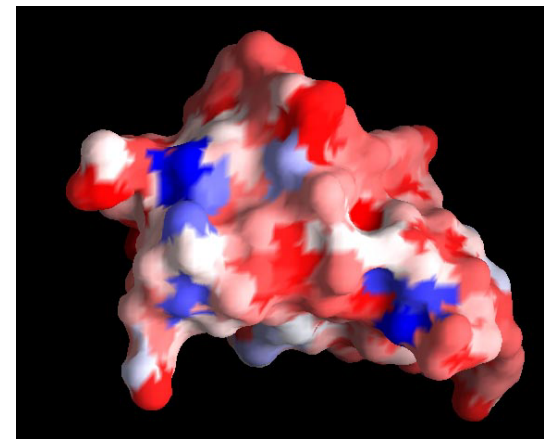
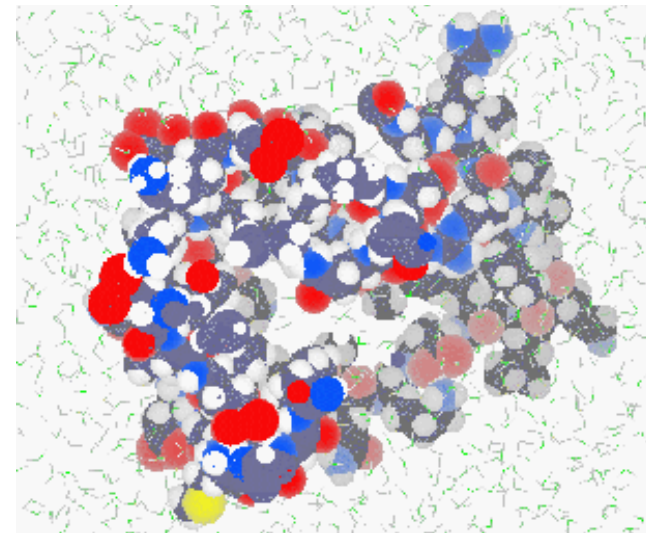
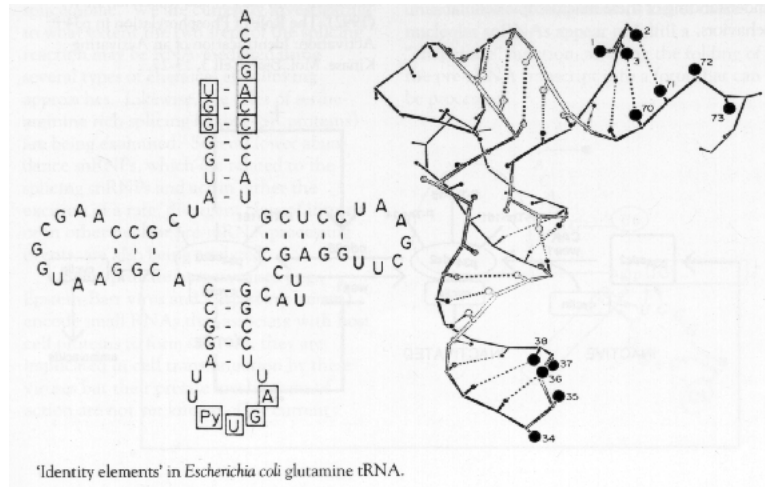
```
d1dhfa_ -PEKNRPLKGRINLVLSRELKEPPQGAHFLSRSLDDALKLTEQPELANKVDMVWIVGGSSVYKEAMNHP
d8dfr_  -PEKNRPLKDRINIVLSRELKEAPKGAHYLSKSLDDALALLDSPELKSQVDMVWIVGGTAVYKAAMEKP
d4dfra_ -G---RPLPGRKNIILSSSQPGTDDRVTWVKSVDIAAIAACGDVPE-----IMVIGGGRVYEQFLPKA
d3dfr_  -P---KRPLPERTNVVLT HQEDYQAQGA-VVVHDVAAVFAYAKQHLD----QELVIAGGAQIFTAFKDDV
```

# 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)

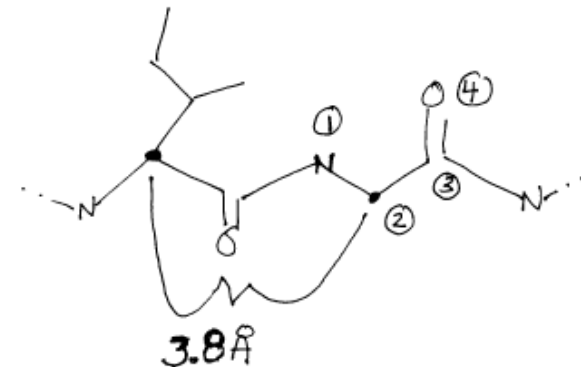


# Molecular Biology Information: Protein Structure Details

- Statistics on Number of XYZ triplets
  - ◇ 200 residues/domain → 200 CA atoms, separated by 3.8 Å
  - ◇ Avg. Residue is Leu: 4 backbone atoms + 4 sidechain atoms, 150 cubic Å
    - => ~1500 xyz triplets (=8x200) per protein domain
  - ◇ >110K Domains, ~1200 folds (scop 1.75)

ATOM 67	1	C	ACE	0	9.401	30.166	60.595	1.00	49.88	1GKY
ATOM 68	2	O	ACE	0	10.432	30.832	60.722	1.00	50.35	1GKY
ATOM 69	3	CH3	ACE	0	8.876	29.767	59.226	1.00	50.04	1GKY
ATOM 70	4	N	SER	1	8.753	29.755	61.685	1.00	49.13	1GKY
ATOM 71	5	CA	SER	1	9.242	30.200	62.974	1.00	46.62	1GKY
ATOM 72	6	C	SER	1	10.453	29.500	63.579	1.00	41.99	1GKY
ATOM 73	7	O	SER	1	10.593	29.607	64.814	1.00	43.24	1GKY
ATOM 74	8	CB	SER	1	8.052	30.189	63.974	1.00	53.00	1GKY
ATOM 75	9	OG	SER	1	7.294	31.409	63.930	1.00	57.79	1GKY
ATOM 76	10	N	ARG	2	11.360	28.819	62.827	1.00	36.48	1GKY
ATOM 77	11	CA	ARG	2	12.548	28.316	63.532	1.00	30.20	1GKY
ATOM 78	12	C	ARG	2	13.502	29.501	63.500	1.00	25.54	1GKY

• • •





# 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>)

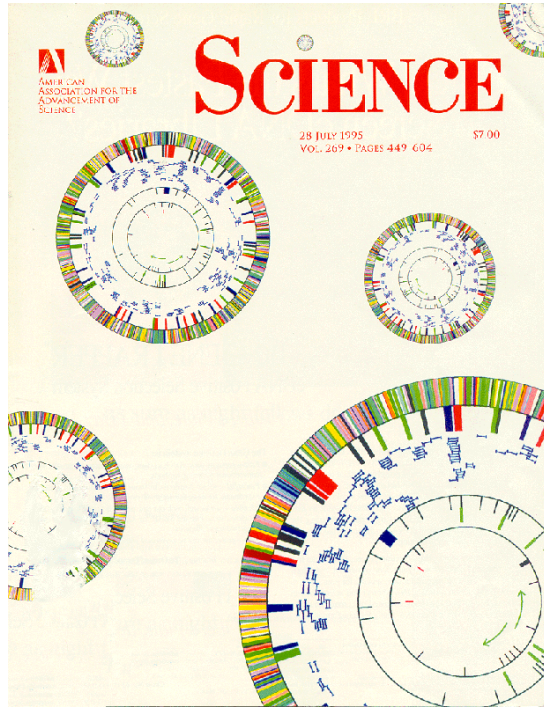
- Timeline

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!  
 2000, draft human  
 2003, human: 3 Gb & 100 K genes...  
 2010, 1000 human genomes!



**1995**

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



**1997**

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



A  
Bioinformatics  
prediction that  
came true!

**1998**

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



**2000?**

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

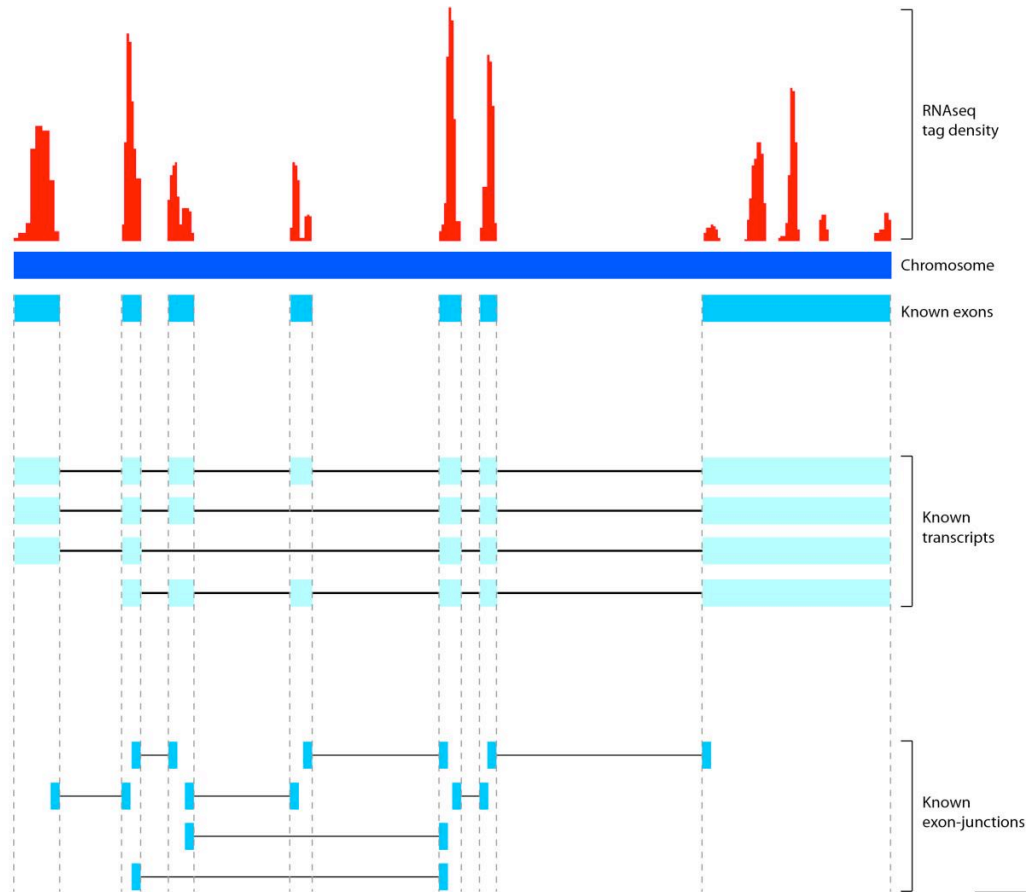


real thing, Apr '00



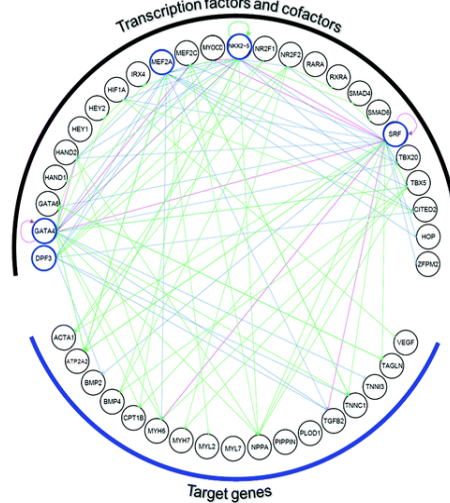
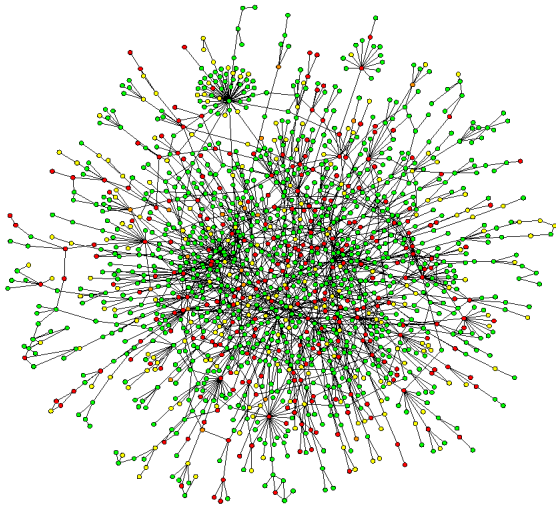
'98 spoof

# Gene Expression Data: On & Off



- Early experiments yeast
  - ◇ Complexity at 10 time points,  
 $6000 \times 10 = 60\text{K}$  floats
- Then tiling array technology
  - ◇ 50 M data points to tile the human genome at  $\sim 50$  bp res.
- Now Next-Gen Sequencing (RNAseq)
  - ◇ 10M+ reads on the human genome, counts
- Can only sequence genome once but can do an infinite variety of expression experiments

# Molecular Networks: Connectivity



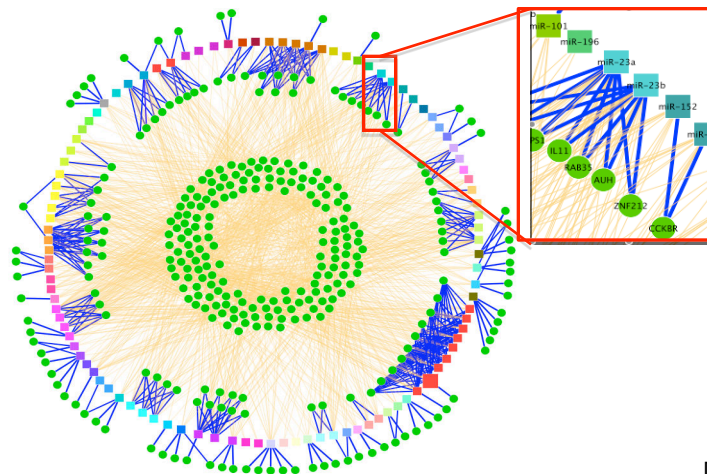
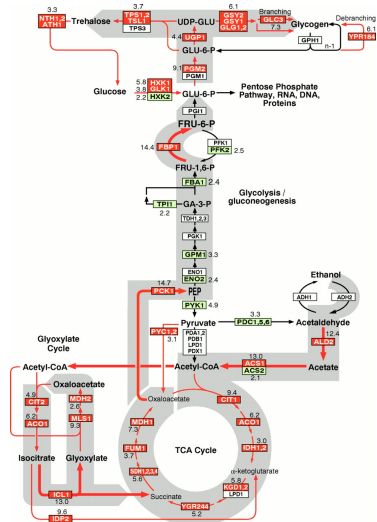
Regulatory Networks  
Get readouts of where proteins bind to DNA : Chip-chip then chip-seq

Protein Interaction Networks

For yeast: 6000 x 6000 / 2 ~ 18M possible interactions (maybe ~30K real)

Protein-protein Interaction networks

TF-target-gene Regulatory networks



Metabolic pathway networks

miRNA-target networks

[Toenjes, et al, *Mol. BioSyst.* (2008); Jeong et al, *Nature* (2001); [Horak, et al, *Genes & Development*, 16:3017-3033; DeRisi, Iyer, and Brown, *Science*, 278:680-686]

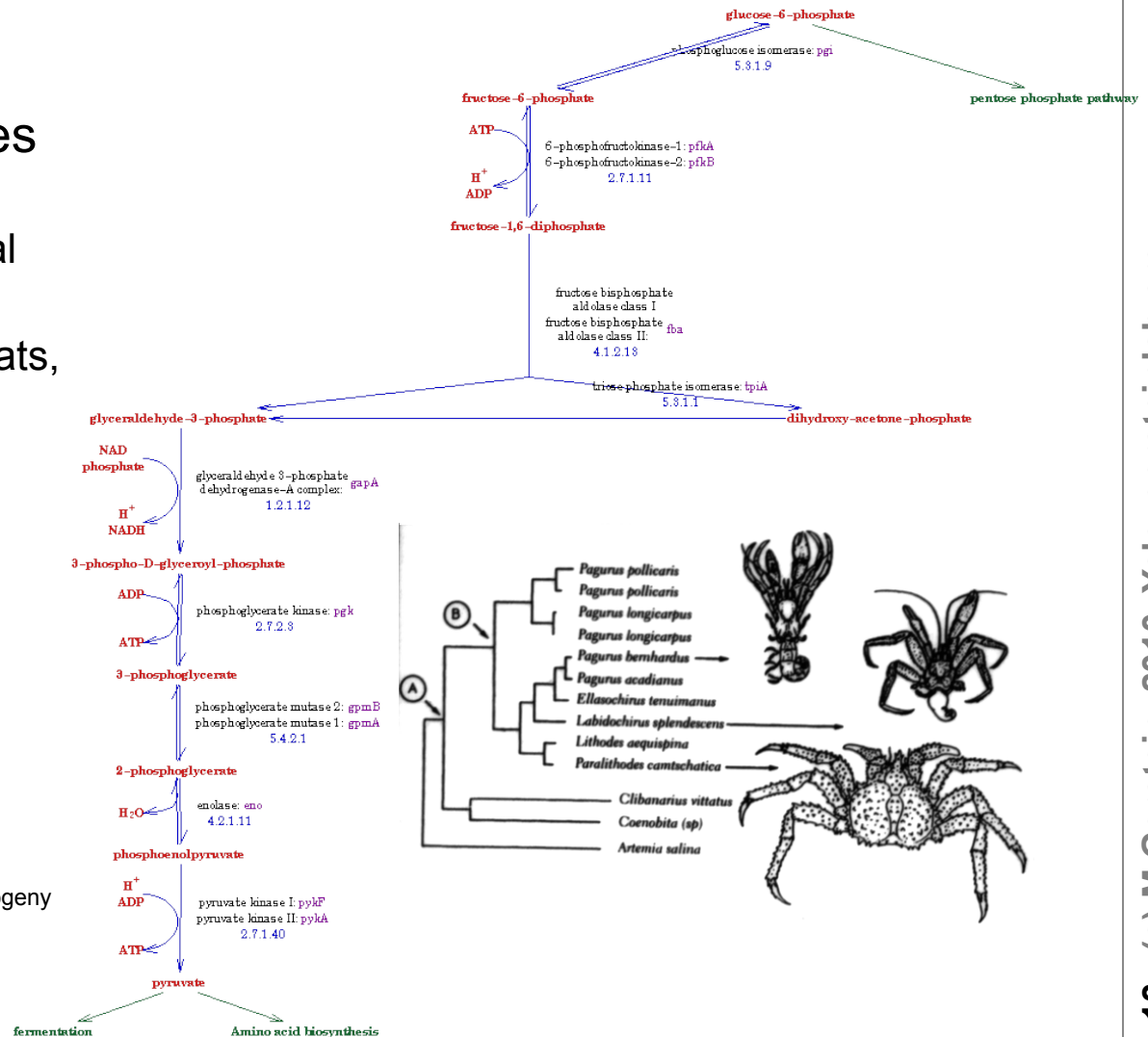
# Molecular Biology Information: Other Integrative Data

- Information to understand genomes

- ◊ Whole Organisms  
Phylogeny, traditional zoology
- ◊ Environments, Habitats, ecology
- ◊ Phenotype Experiments  
(large-scale KOs, transposons)
- ◊ 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**

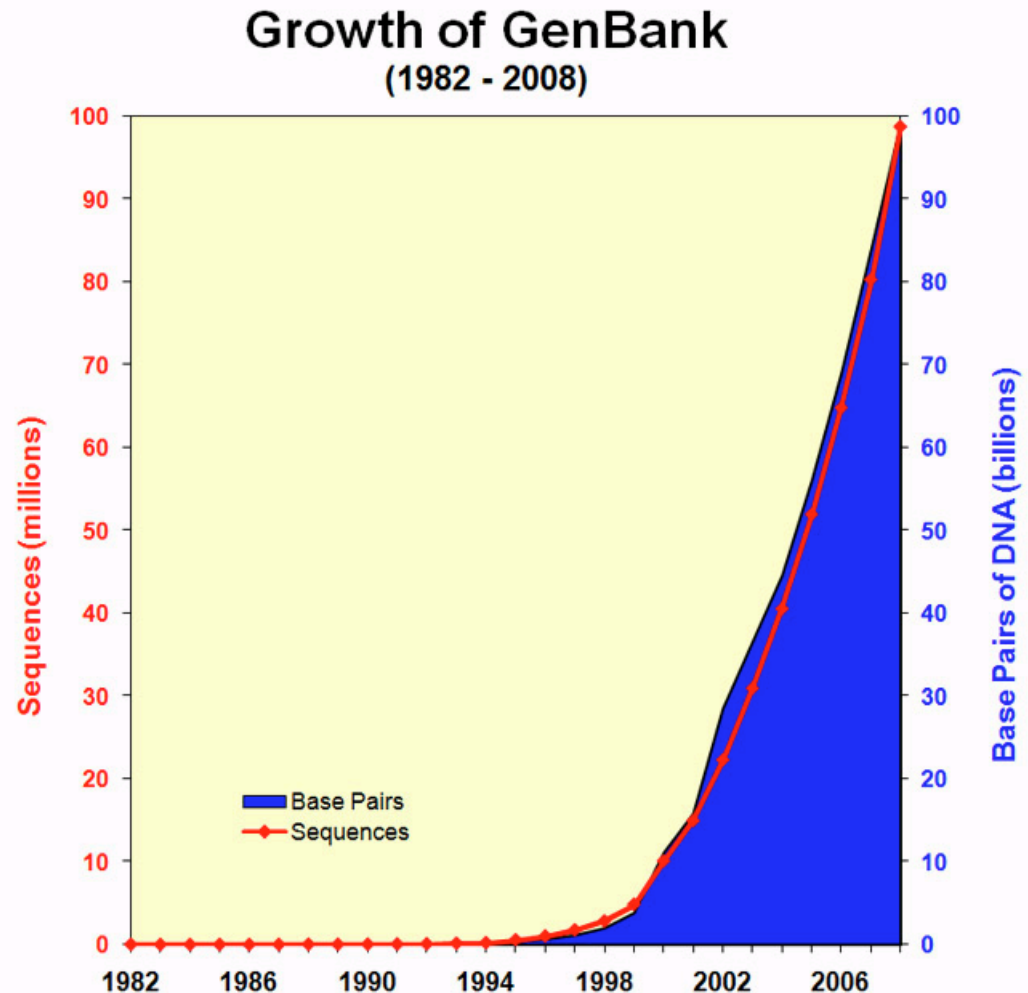
- One idea for a definition?

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.**

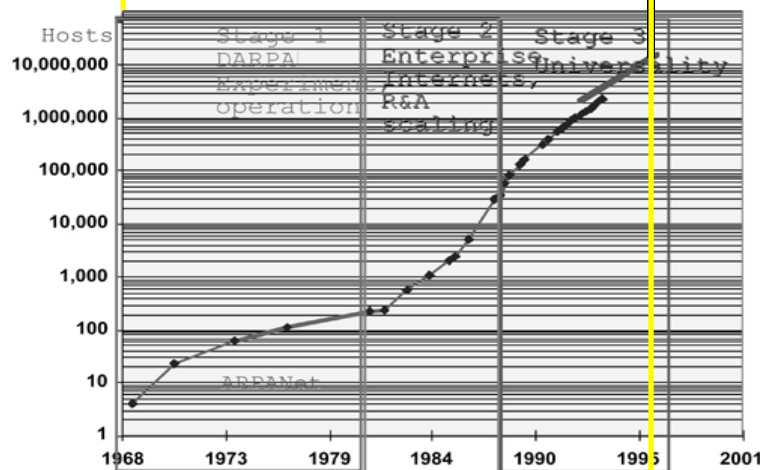
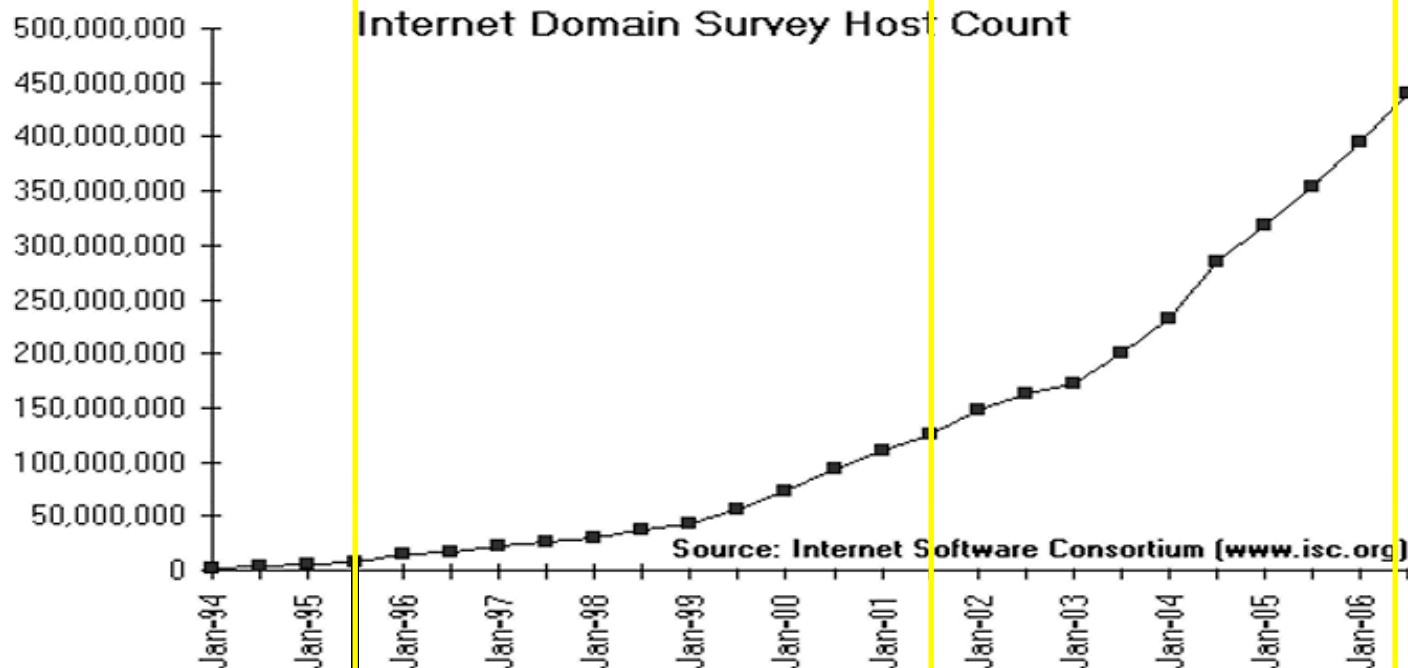
# Large-scale Information: Exponential Growth of Data Matched by Development of Computer Technology

- 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
  - ◇ Comparison with **Moore's Law**
- A Driving Force in Bioinformatics



# Internet Hosts

(adapted from D Brutlag, Stanford & <http://navigators.com/stats.html>)

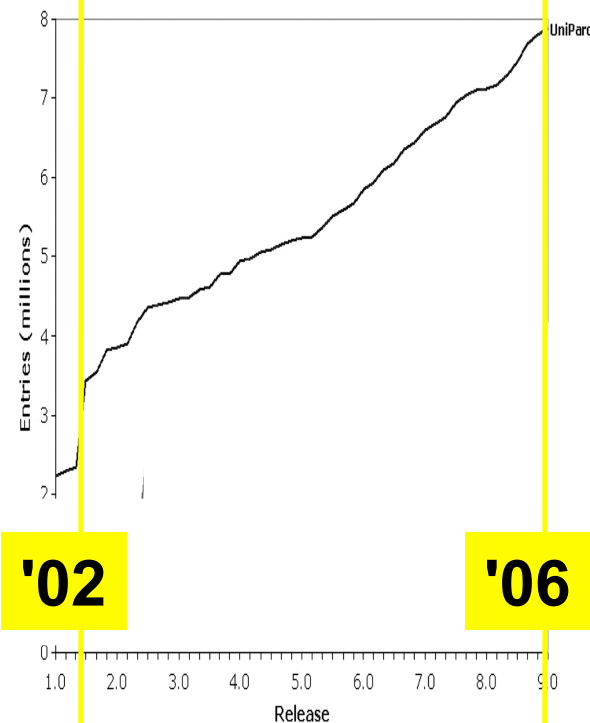


'68

'95

# Proteins

Suzek, B. E. et al.  
 Bioinformatics 2007  
 23:1282-1288; doi:  
 10.1093/bioinformatics/  
 btm098

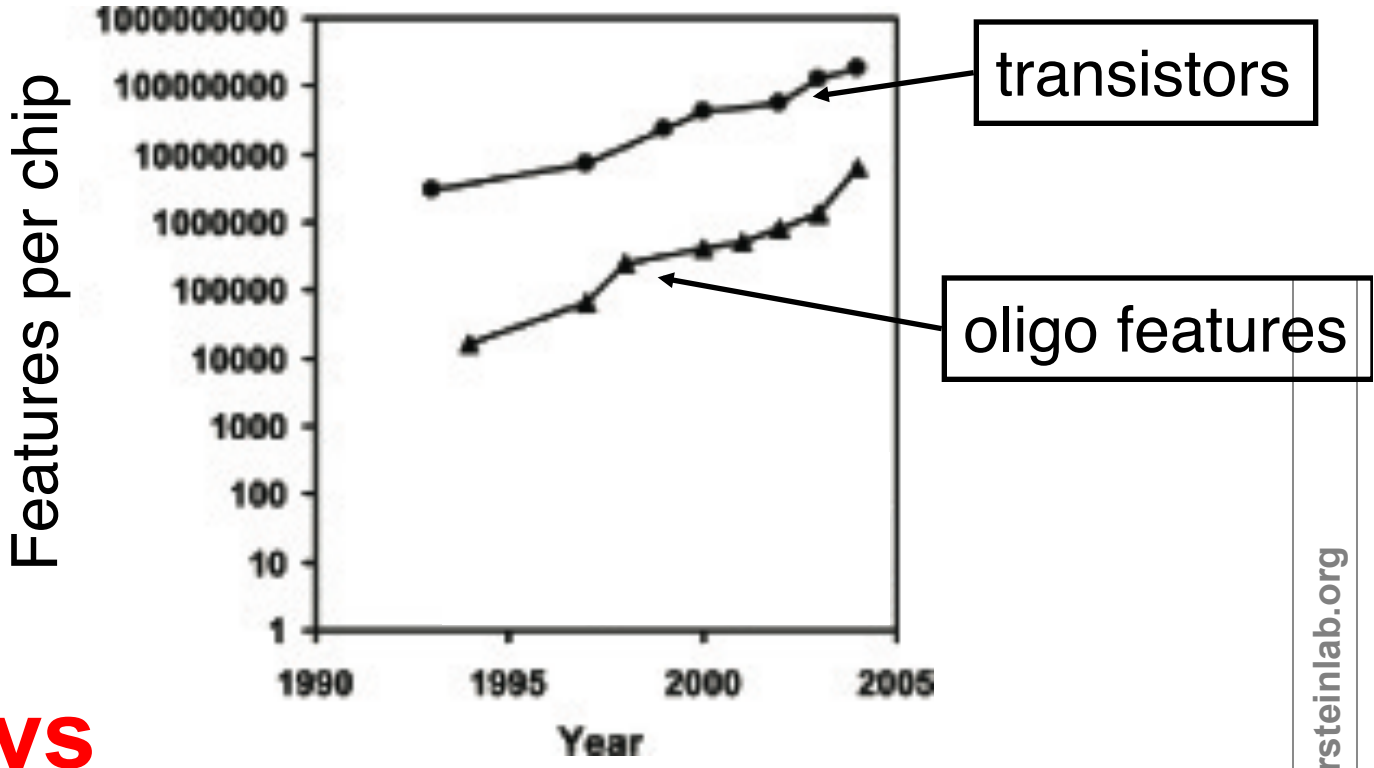


'02

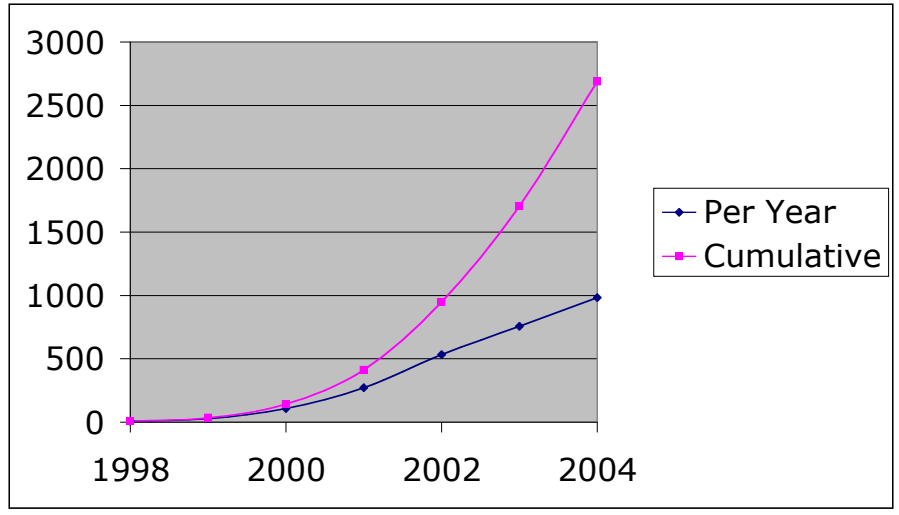
'06



Features per Slide

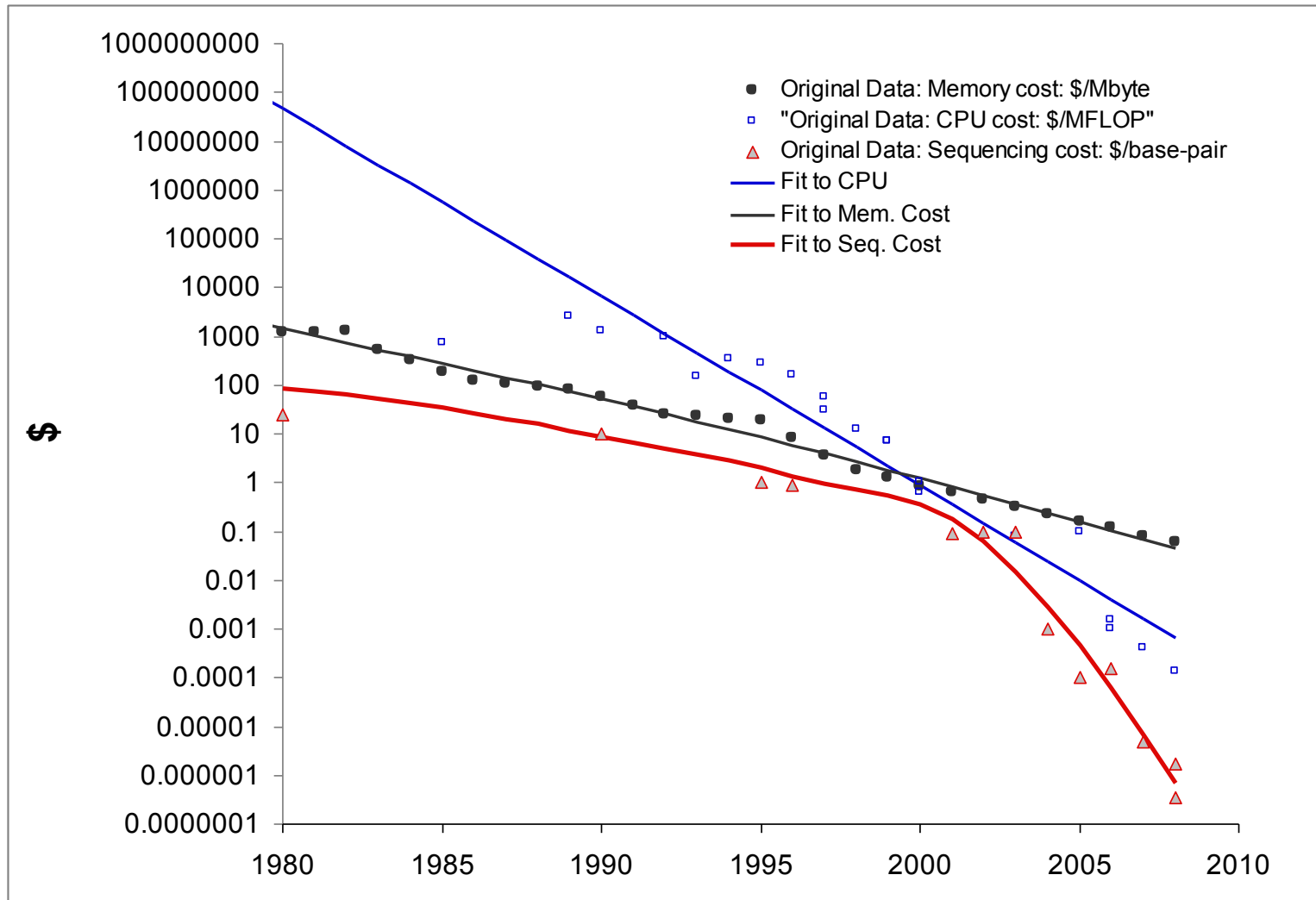


# Microarrays



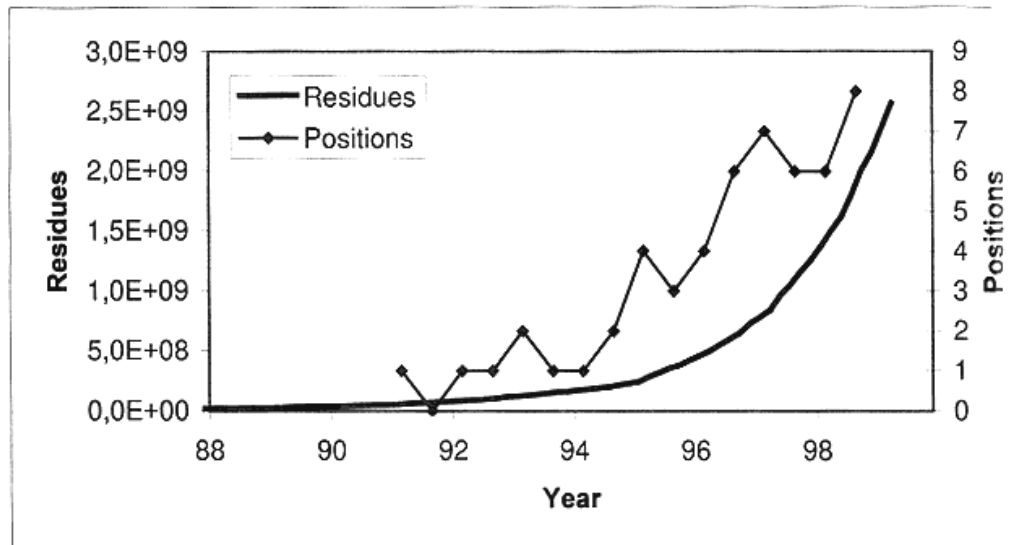
PubMed  
publications with  
title "microarray"

# Plummeting Cost of Sequencing

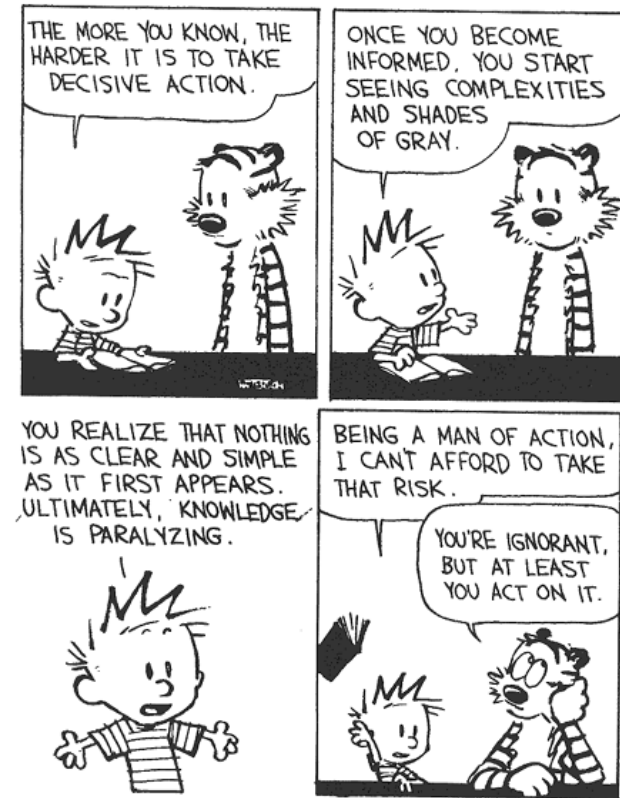


[Greenbaum et al., Am. J. Bioethics ('08)]

# Jobs: 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**

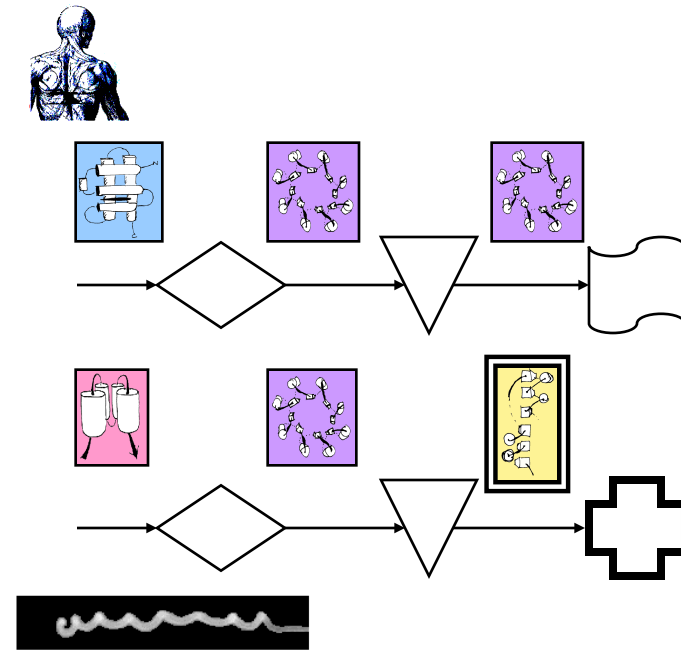
- One idea for a definition?

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

Netscape: NCBI CDD Help

File Edit View Go Communicator Help

Location: <http://www.ncbi.nlm.nih.gov/Structure/cdd/>

NCBI CDD

PubMed BLAST OMIM Taxonomy Entrez Structure

Search Entrez Structure for [ ] Go

CDD Home

CDD - Conserved Domain Database Help

Index

- Conserved Domain Databases
  - What is a Conserved Domain?
  - What are the Source Databases?
  - What are the CD processing steps?
  - How and when is CDD updated?
  - How to find "Conserved Domains"
  - Alignment visualization in the CD-Browser
  - What happens when I click the [CD] hotlink?
- CD-Search Service
  - What is RPS-Blast?
  - Which Search Databases are available?
  - Can I run RPS-Blast locally?
  - What input is required?
  - How long do I have to wait for the results?
  - What are the elements on the results page?
  - How do I look at multiple alignments?
  - Alignment visualization including 3D-structures
  - What does the pink dot mean?

What is a Conserved Domain?

Domains can be thought of as functional and/or structural units of a protein. These two classifications coincide rather often, and what is found as an independently folding unit of a polypeptide chain also carries a specific function. Typically domains are identified as recurring (sequence or structure) units, which may exist in various contexts. The image below illustrates 4 "domains" identified as structural units in the MMDB-entry [1IGR](#), chain A. (Click on the figure to launch this view in [Cn3D](#)):

For this query sequence, the CD-Search service would identify the conserved domains

Netscape: NCBI CDD Help

File Edit View Go Communicator Help

Location: <http://www.ncbi.nlm.nih.gov/Structure/cdd/>

Research topics and staff

### What is a Conserved Domain?

Domains can be thought of as functional and/or structural units of a protein. These two classifications coincide rather often, and what is found as an independently folding unit of a polypeptide chain also carries a specific function. Typically domains are identified as recurring (sequence or structure) units, which may exist in various contexts. The image below illustrates 4 "domains" identified as structural units in the MMDB-entry [1IGR](#), chain A. (Click on the figure to launch this view in [Cn3D](#)):

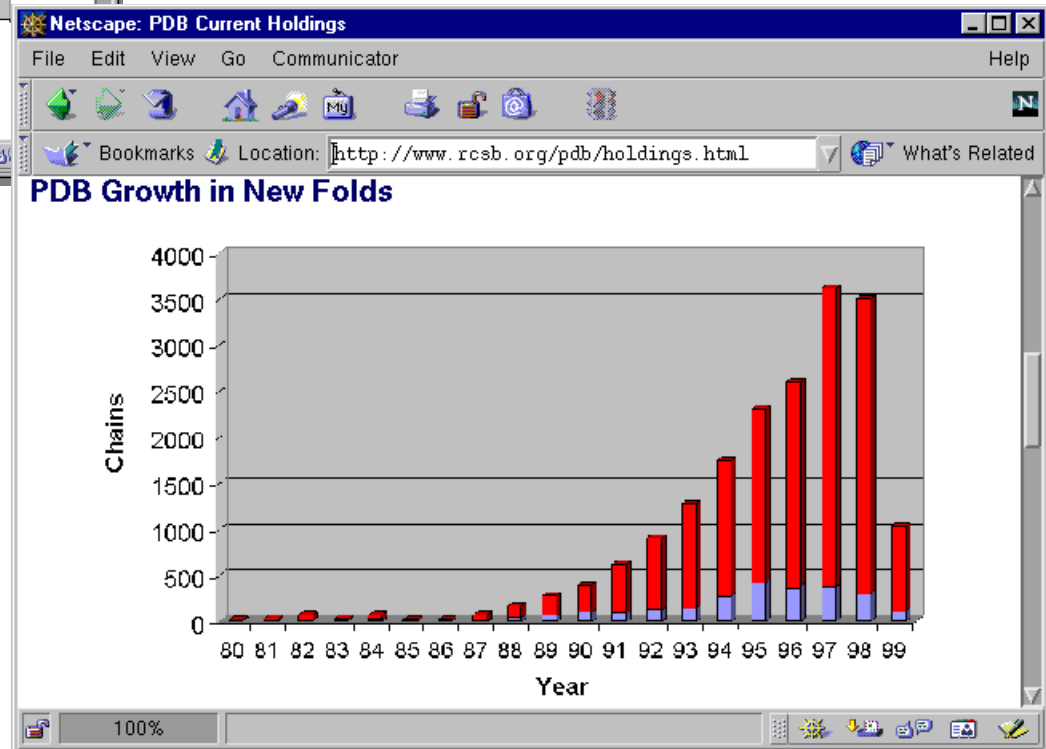
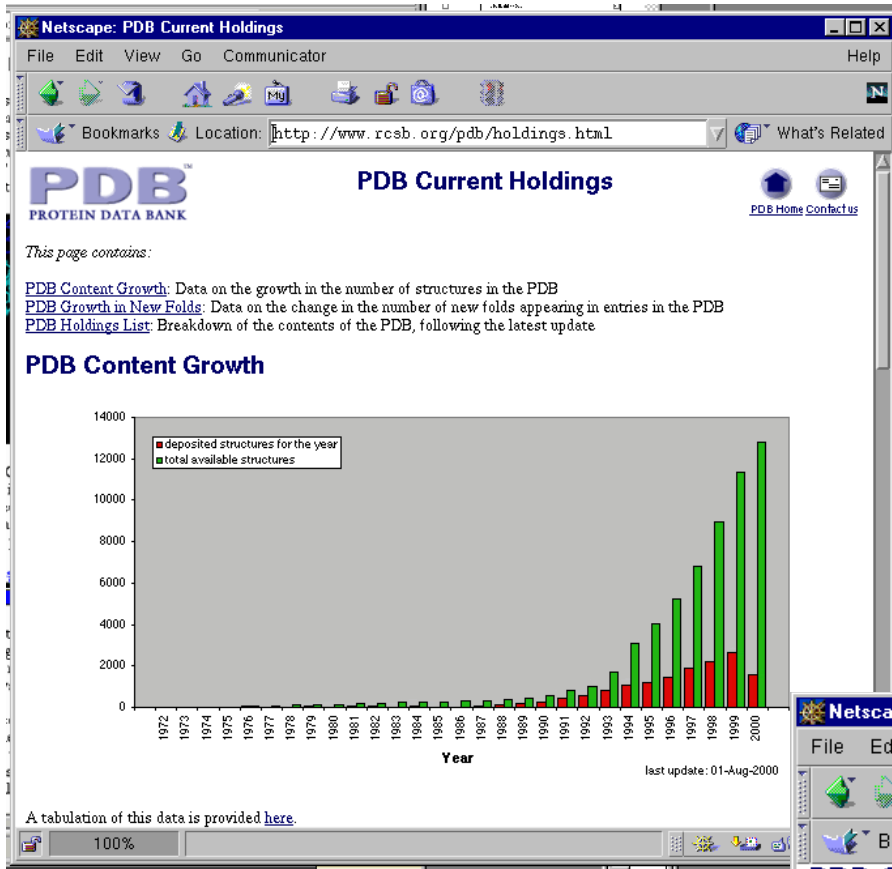
```
1 EICGPGIDIR NDYQOLKRL NCTVIEGYLH
31 ILLISKAEDY RSYRFPKLTV ITEVLLIFV
61 AGLRSIGDLF PHLTVIRGKQ IFTYVALVIF
91 BMTLKRDIQL YLNRNITRGA IRREKHADEL
121 YLSTVDVSLI LDVSNWVIV GSKPPKQGD
151 LCPGTMEKRP KCEKTTINNE YNRCVTTNR
181 CQKCPSTCG KRACTENNEC CHPECLGSCS
211 AFQNDTACVA CRWYTAGQC VPACFPNTFR
241 FEGVRCVDRD FCANILSAES SSBEGFVIHD
271 GECMOCPSG FIRNGSQSHY CIPCEGPCPK
301 VCEREKTKT IDSVTSAQML OQCTIFKGNL
331 LINIRRGNNI ASELENFMGL IEVYTGPKKI
361 RRSNALVSLG FLYKRLILG EKOLEGQVSE
391 YVLDNQNLQO LVDNDRRLIT IKAGKMYFAP
421 NPKLCVSEIY RMEEVYTKG ROSKGDINTR
451 NNGERASCES DVDDCKEOK LISEEDLN
```

For this query sequence, the CD-Search service would identify the conserved domains indicated below (click on the image below to launch the actual search). Good correspondence exists between structural units, identified by purely geometric criteria, and units asserted to be evolutionary conserved. The region annotated as "Furin-like" was split in two by the MMDB domain parser.

Molecular evolution readily utilizes such domains as building blocks which may be recombined in different arrangements to modulate protein function. We define conserved domains as recurring units in molecular evolution whose extents can be determined by sequence and structure analysis.

Conserved domains contain conserved sequence patterns or motifs, which allow for their detection in polypeptide sequences. The distinction between domains and motifs is not sharp, however, especially in the case of short repetitive units. Functional motifs are also present outside the scope of structurally conserved domains. The CD database does not attempt to systematically collect these.

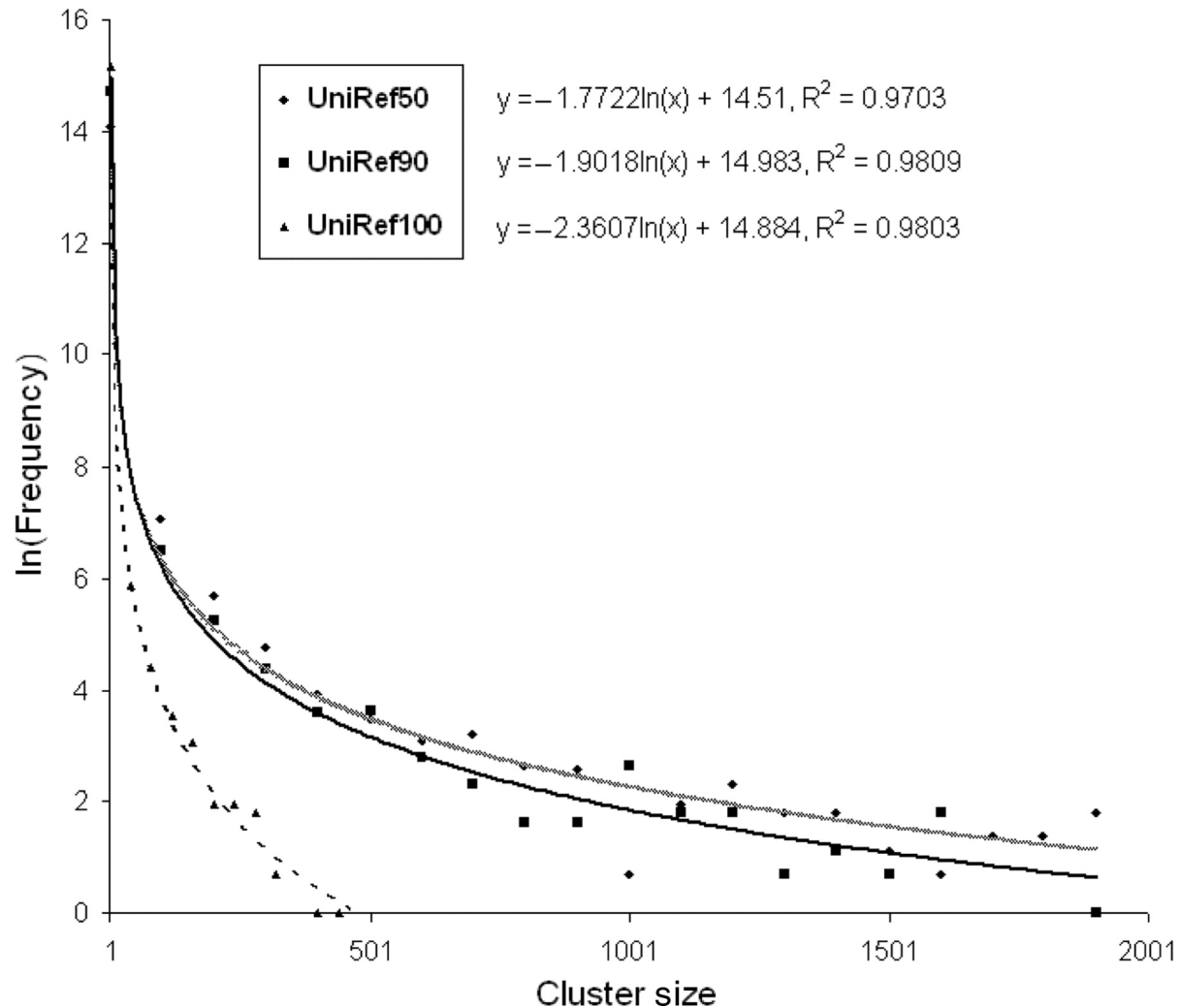
Vast Growth in (Structural)  
Data...  
but number of  
Fundamentally New (Fold)  
Parts Not Increasing that  
Fast



Total in Databank  
 New Submissions  
 New Folds



# Power-law Size to Protein Families



Suzek, B. E. et al. *Bioinformatics* 2007 23:1282-1288; doi:10.1093/bioinformatics/btm098; See also Luscombe et al., 2002, *JMB*.



# 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.**

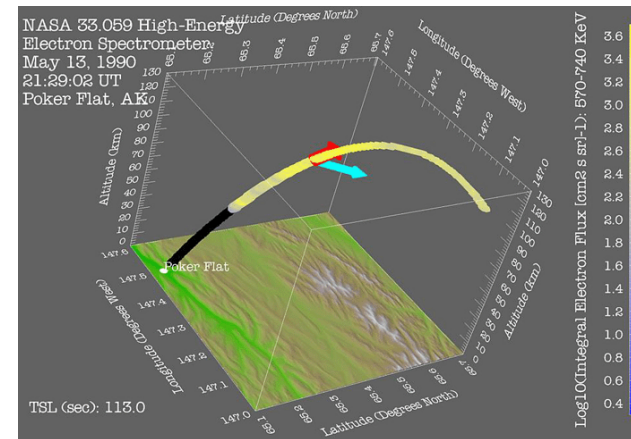
# General Types of “Informatics” techniques in Bioinformatics

- Databases
  - ◇ Building, Querying
  - ◇ Representing Complex data
- Data mining
  - ◇ Machine Learning techniques
  - ◇ Clustering & Tree construction
  - ◇ Text String Comparison
  - ◇ Significance Statistics
- Network Analysis
  - ◇ Topology
  - ◇ Connectivity predictions
- Structure Analysis & Geometry
  - ◇ Graphics (Surfaces, Volumes)
  - ◇ Comparison and 3D Matching (Vision, recognition)
- Physical Simulation
  - ◇ Newtonian Mechanics
  - ◇ Electrostatics
  - ◇ Numerical Algorithms
  - ◇ Simulation
  - ◇ Modelling Chemical Reactions & Cellular Processes

# Data Mining as New Paradigm for Scientific Computing

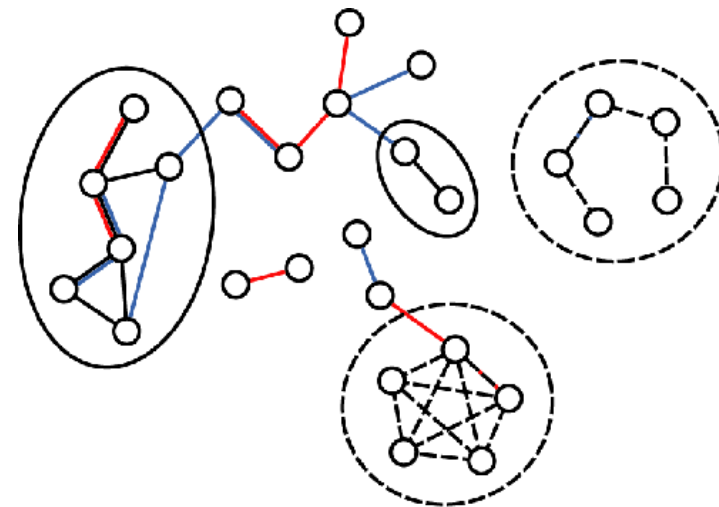
- Physics

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



- Biology

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





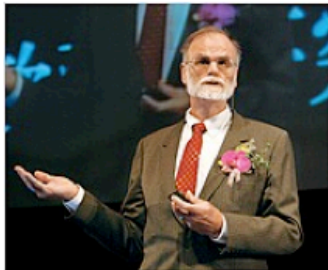
BOOKS ON SCIENCE

## A Deluge of Data Shapes a New Era in Computing

By JOHN MARKOFF  
Published: December 14, 2009

In a speech given just a few weeks before he was lost at sea off the California coast in January 2007, Jim Gray, a database software pioneer and a [Microsoft](#) researcher, sketched out an argument that computing was fundamentally transforming the practice of science.

[Enlarge This Image](#)



Microsoft

**THINKER** A collection of essays pays tribute to Jim Gray, a database software engineer who disappeared off the California coast almost three years ago.

### THE FOURTH PARADIGM

Data-Intensive Scientific Discovery. Edited by Tony Hey, Stewart Tansley and Kristin Tolle. Microsoft Research. 252

Dr. Gray called the shift a “fourth paradigm.” The first three paradigms were experimental, theoretical and, more recently, computational science. He explained this paradigm as an evolving era in which an “exaflood” of observational data was threatening to overwhelm scientists. The only way to cope with it, he argued, was a new generation of scientific computing tools to manage, visualize and analyze the data flood.

In essence, computational power created computational science, which produced the overwhelming flow of data, which now requires a computing change. It is a positive feedback loop in which the data stream becomes the data flood and sculpts a new computing landscape.

In computing circles, Dr. Gray’s crusade was described as, “It’s the data, stupid.” It was a point of view that caused

- SIGN IN TO RECOMMEND
- TWITTER
- E-MAIL
- SEND TO PHONE
- PRINT
- REPRINTS
- SHARE

# NY Times 14-Dec-09 article on Jim Gray's 4<sup>th</sup> Paradigm

Statistical  
Analysis  
vs.  
Classical  
Physics

Bioinformatics,  
Genomic Surveys

Vs.

Chemical  
Understanding,  
Mechanism,  
Molecular Biology

**How Does Prediction Fit into the Definition?**

# **Bioinformatics: Practical Application of Simulation and Data Mining**

## ***Practical Stuff***

<http://www.gersteinlab.org/courses/452>

- Hopefully exact schedule posted
- PDFs and PPTs of lectures
- Previous year's courses (more than 10 years!)
- **cbb752@gersteinlab.org**

# People & Times

- Timing
  - ◇ MW 1.00-2.15 BASS 305
  - ◇ Discussion sect. each week (TBD)
- Instructors
  - ◇ Mark Gerstein (in charge)
  - ◇ Guest Lecturers
    - James Noonan + Jesse Rinehart
    - Kei Cheung
    - Corey S. O'Hern, simulation
    - Steven Kleinstein
- TFs (Pedro + Jia)
- Office Hours for Mark
  - ◇ 15' after this class
  - ◇ 30' after class on Wed.
  - ◇ By appointment
- No Fri. class this week



# Grading

- Quizzes (~4) [1<sup>st</sup> probably on 2 Feb. or earlier]
- No Final Exam
- Discussion Section Participation
  
- Final Project\*
- Homework\*
- \* CBB/CS students will do programming here and MBB will do writing
  
- Prerequisites
  - ◇ MB&B 301b *and* MATH 115a or b,  
***or permission of instructor***

# Two courses

- CBB 752/CS 752
  - ◇ Programming Assignments, Extra computation
  - ◇ Programming final project
  
- MBB 452/752
  - ◇ No programming
  - ◇ Written Final Project

# Course Catalog Description

**Techniques** in  
**data mining & simulation**  
**applied** to bioinformatics,  
the computational analysis of **gene sequences, macromolecular structures, and functional genomics data on a large scale.**

**(Some topics include: )** Sequence alignment, comparative genomics and phylogenetics, biological databases, geometric analysis of protein structure, molecular-dynamics simulation, biological networks, microarray normalization, and machine-learning approaches to data integration.

**Not the same as Genomics & Bioinformatics** in previous years contains all of the "**Bioinformatics**" and then **more** (!) with less "**Genomics**".  
Previous years slides are a rough guide to the bioinformatics part.

# Bioinformatics Topics -- Genome Sequence

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

- 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

# Bioinformatics

## Topics --

# Protein Sequence

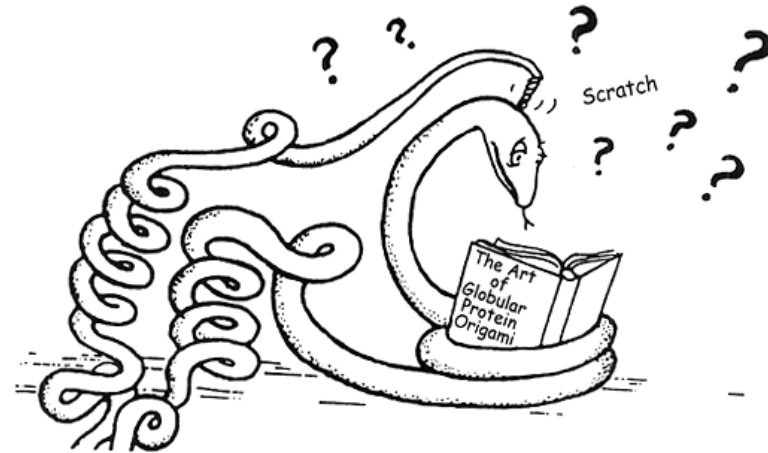
- Scoring schemes and Matching statistics
  - ◇ How to tell if a given alignment or match is statistically significant
  - ◇ A P-value (or an e-value)?
  - ◇ Score Distributions (extreme val. dist.)
  - ◇ Low Complexity Sequences
- Evolutionary Issues
  - ◇ Rates of mutation and change

# Bioinformatics

## Topics -- Sequence / Structure

- Secondary Structure “Prediction”
  - ◇ via Propensities
  - ◇ TM-helix finding
  - ◇ Assessing Secondary Structure Prediction
- Structure Prediction: Protein v RNA

“Now collapse down hydrophobic core, and fold over helix 'A' to dotted line, bringing charged residues of 'A' into close proximity to ionic groups on outer surface of helix 'B' ...”



Reproduced in U. Tollemar, “Protein Engineering i USA”, Sveriges Tekniska Attach er, 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

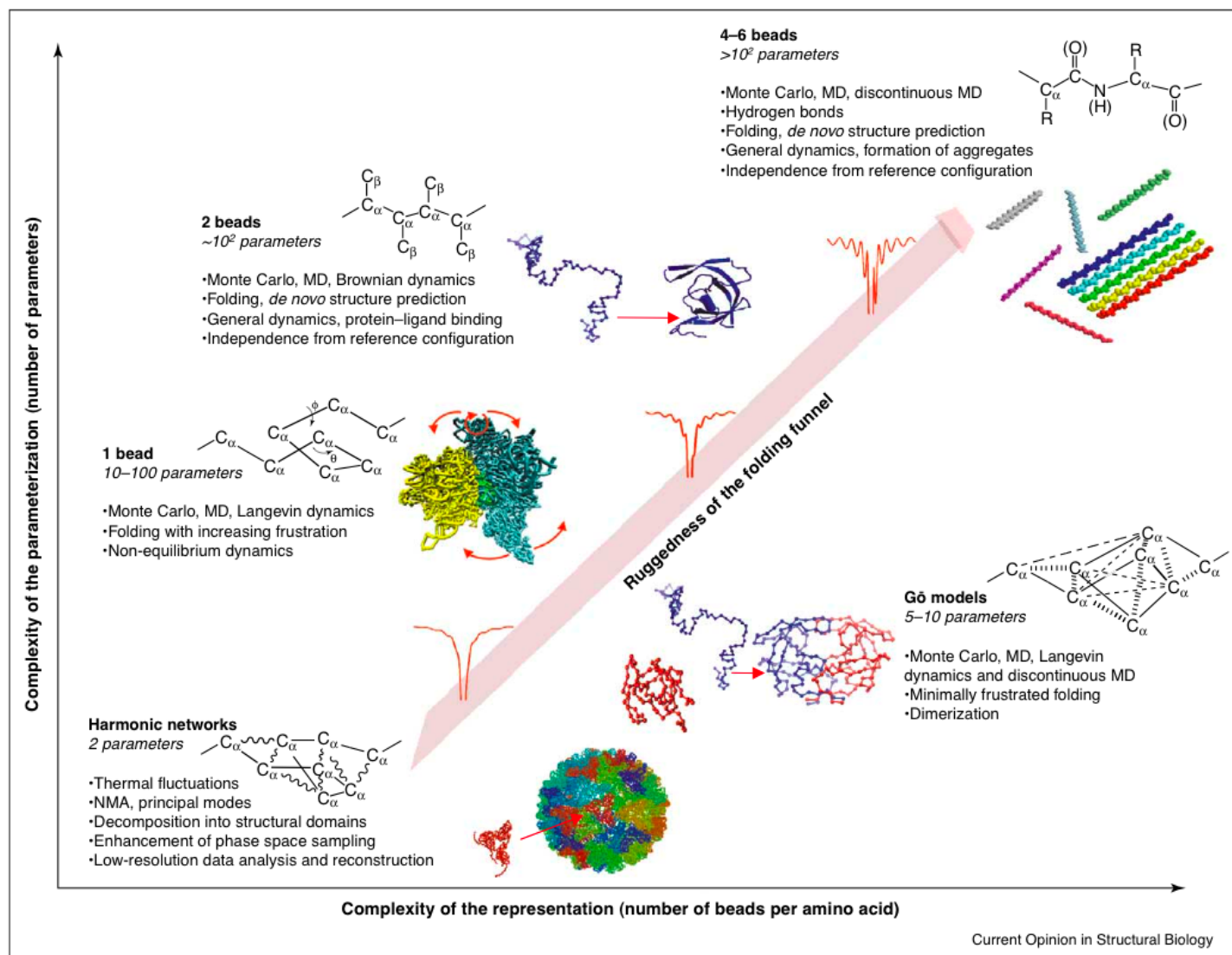
- Structure Comparison
  - ◇ Basic Protein Geometry and Least-Squares Fitting
- Distances, Angles, Axes, Rotations
  - ◇ Calculating a helix axis in 3D via fitting a line
  - ◇ LSQ fit of 2 structures
  - ◇ Molecular Graphics
- Calculation of Volume and Surface
  - ◇ Hinge prediction
  - ◇ Packing Measurement
- 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 -- 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



Figure 1



Pictorial representation of the features of bead models. For each class of model, the following aspects are reported: schematic representation of the model, indicative number of parameters, methods of solution, main characteristics and applications. Sample applications are also illustrated with representative pictures (prepared using crystallographic coordinates from the PDB [codes 1hhp, 1cwp, 1mwr, 486d]) intended to show the size of system that can be studied and the kind of study that can be done. The location of the models in the  $x$ - $y$  plane is intended to qualitatively illustrate their complexity, which increases following the direction of the arrows.

## Journal Articles

1. J. D. Honeycutt and D. Thirumalai, “The nature of folded states of globular proteins,” *Biopolymers* **32** (1992) 695.
2. W. C. Swope and J. W. Pitera, “Describing protein folding kinetics By molecular dynamics simulations. 1. Theory,” *J. Phys. Chem. B* **108** (2004) 6571.
3. D. Bratko, T. Cellmer, J. M. Prausnitz, and H. W. Blanch, “Molecular Simulation of protein aggregation,” *Biotechnology and Bioengineering* **96** (2007) 1.

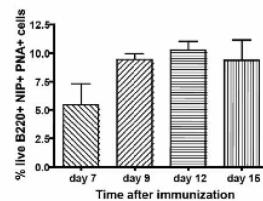
# Cell and Immunology Simulations

## Prof. Steven Kleinstein

- Modeling cell mutation, division and death
- Population dynamics using ODEs
- Viral dynamics and immunological response
- Optimization and matching experimental data

### Germinal Center Population at Steady-State

Solve model for inter-zonal migration rates ( $m_D$  and  $m_L$ )



$$\frac{dS}{dt} = m_L LZ - qS$$

$$\frac{dG2M}{dt} = qS - pM$$

$$\frac{dDZ}{dt} = 2pM - m_D DZ$$

$$\frac{dLZ}{dt} = m_D DZ - (m_L + d)LZ$$

$$m_D = \frac{pq(R+2)}{CRpq - Rp - Rq - p - q}$$

$$m_L = \frac{m_D pq}{Cm_D pq - m_D p - m_D q - pq}$$

Additional experiments to estimate other parameters

# Next Generation Sequencing & Big Data

- Seq. Tech
- Assembly
- RNA-seq
- ChIP-seq
- Metagenomics

# Topics – (Func) Genomics

- Expression Analysis
  - Time Courses clustering
  - Measuring differences
  - Identifying Regulatory Regions
  - Normalization and scoring of arrays
- Function Classification and Orthologs
- Genome Comparisons
  - Large-scale censuses
  - Frequent Words Analysis
  - Genome Annotation
  - Identification of interacting proteins
- Structural Genomics
  - Folds in Genomes, shared & common folds
  - Bulk Structure Prediction










# Topics – DBs/Surveys

- Relational Database Concepts and how they interface with Biological Information
- DB interoperation
- What are the Units ?
  - What are the units of biological information for organization?
- Clustering & Trees
  - Basic clustering
  - Evolutionary implications
- Visualization of Large Amounts of Information

# Mining

- Information integration and fusion
  - Dealing with heterogeneous data
- Dimensionality Reduction (PCA etc)
- Networks
  - Topology Analysis
  - Prediction of linkages
  - Global structure and local motifs

# Bioinformatics Spectrum

		Breadth: Homologs, Large-scale Surveys, Informatics—				
			pairwise comparison, sequence & structure alignment	multiple alignment, patterns, templates, trees	databases, scoring schemes, censuses	
		1	2	3-100	100+	
Depth: Rational Drug Design (physics) →		<b>Genome Sequence</b>	atcgcgatattgggattgggga	atcgcgatattgggattgggga atcgcgatattgggattgggga	atcgcgatattgggattgggga atcgcgatattgggattgggga atcgcgatattgggattgggga atcgcgatattgggattgggga atcgcgatattgggattgggga	
	gene finding	↓				
		<b>Protein Sequence</b>	ALMNAKKKPPQRT	ALMNAKKKPPQRT ALMNAKKKPPQRT	ALMNAKKKPPQRT ALMNAKKKPPQRT ALMNAKKKPPQRT ALMNAKKKPPQRT ALMNAKKKPPQRT	
	structure prediction	↓				
		<b>Protein Structure</b>		 	  	
	geometry calculation	↓				
		<b>Protein Surface</b>				
	molecular simulation	↓				
		<b>Force Field</b>				
	structure docking	↓				
	<b>Ligand Complex</b>					



# What is Bioinformatics?

- *(Molecular)* **Bio - informatics**

- One idea for a definition?

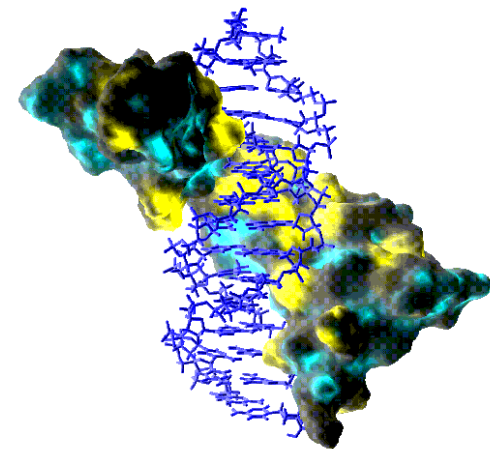
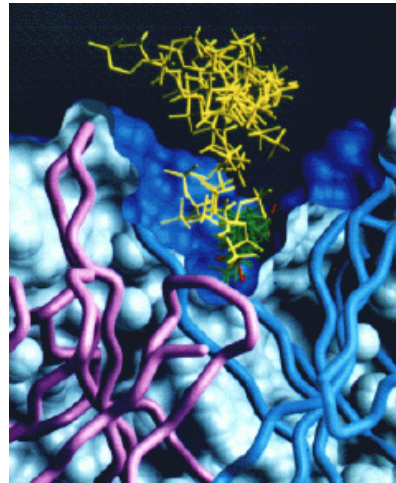
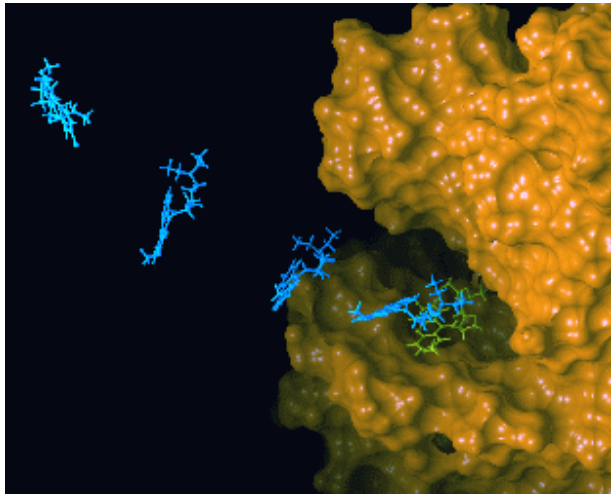
Bioinformatics is conceptualizing **biology in terms of molecules** (in the sense of physical-chemistry) and then applying **“informatics” techniques** (derived from disciplines such as applied math, CS, and statistics) to understand and **organize the information associated** with these molecules, **on a large-scale.**

- Bioinformatics is a practical discipline with many **applications.**

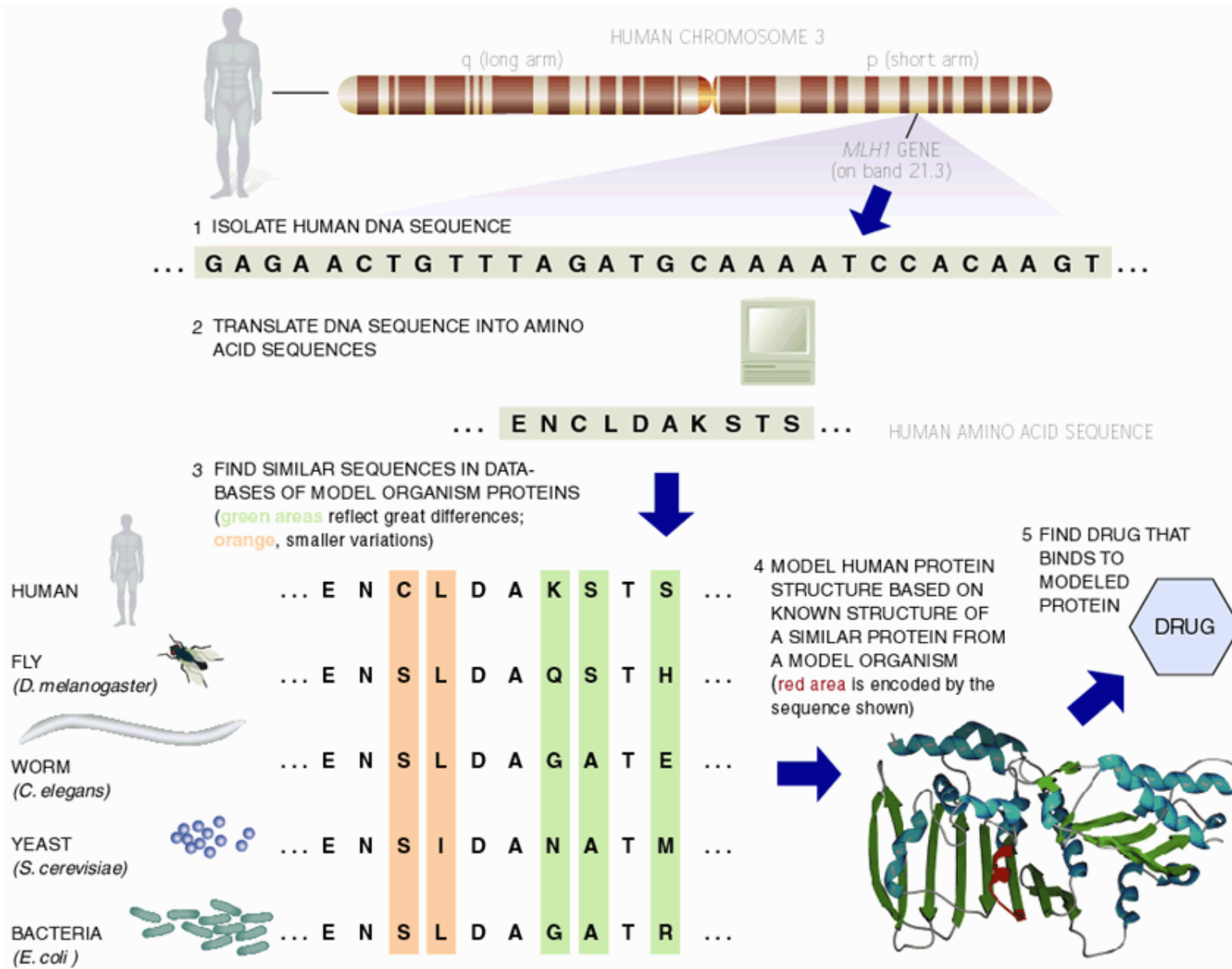
# Major Application I: Designing Drugs

- 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).



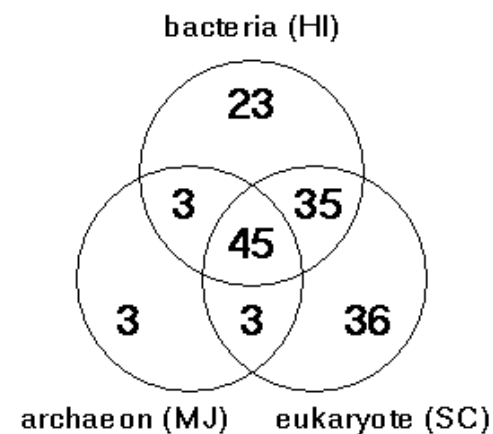
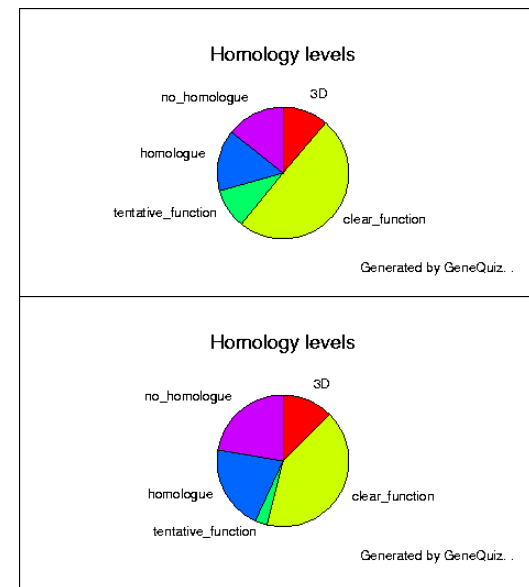
# Major Application II: Finding Homologs



# Major Application III: 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
- Using this for **picking drug targets**

(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)

# Some Further Boundary Examples....

- Char. drugs and other small molecules (cheminformatics or bioinformatics?) [YES]
- Molecular phenotype discovery – looking for gene expression signatures of cancer [YES]
  - ◇ What if it included non-molecular data such as age ?
- Use of whole genome sequences to create phylogenies [YES]
- Integration and organization of biological databases [YES]
- Processing of NextGen sequencing image files [NO]