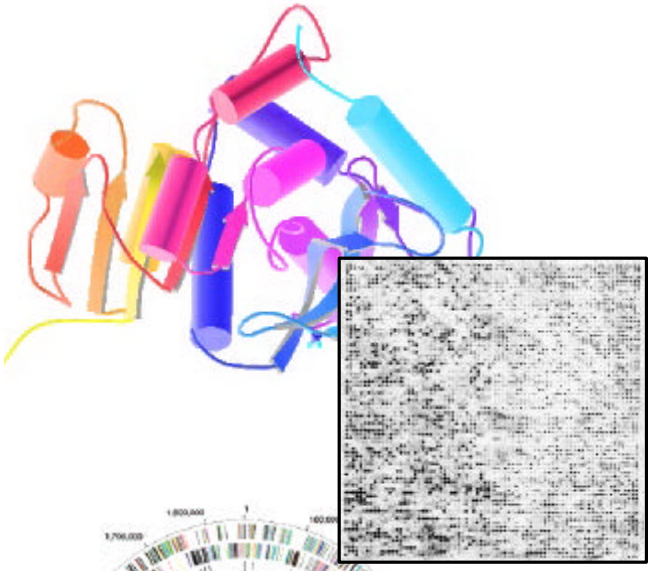
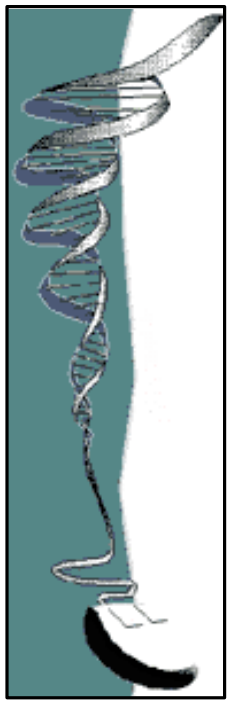
BIOINFORMATICS
CS 440 Guest Lecture

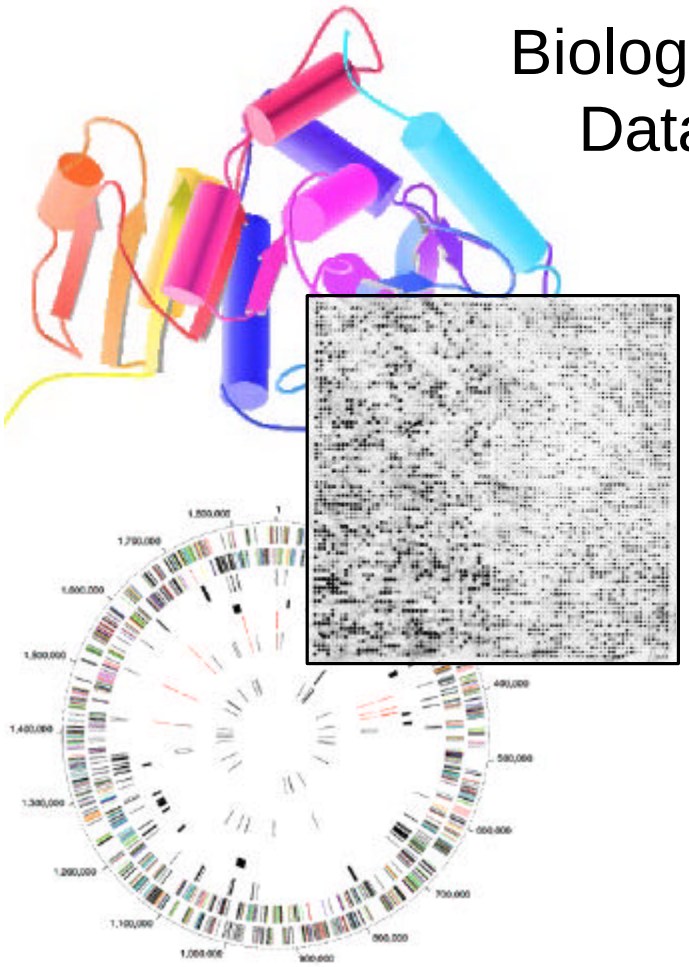




Mark Gerstein, Yale University

bioinfo.mbb.yale.edu/mbb452a

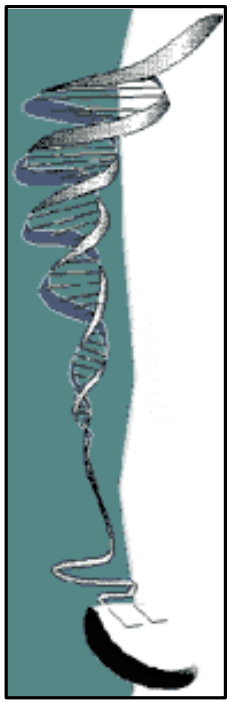
Bioinformatics



Biological Data

+

Computer Calculations



Bioinformatics

- A Very Broad Overview: What is Bioinformatics?
  - ◊ Types of Information, Organizing Principles, Informatics Techniques, Real-world Applications
- Example Calculation 1: Datamining Genome Information
  - ◊ Representing expression data and other features in high-dimensional space; Discriminants
  - ◊ Simple Bayesian analysis
- Example Calculation 2: Aligning Text Strings
  - ◊ Simple dynamic programming
  - ◊ Adding in gaps and other complexities

What is Bioinformatics?

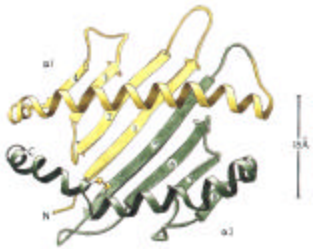
- *(Molecular)* **Bio** - **`informatics`**
- One idea for a definition?
  Bioinformatics is conceptualizing **biology in terms of molecules** (in the sense of physical-chemistry) and then applying **"informatics" techniques** (derived from disciplines such as applied math, CS, and statistics) to understand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



•Genetic material

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

- Central Paradigm for Bioinformatics

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

- Large Amounts of Information
  - ◊ Standardized
  - ◊ Statistical

•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

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

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

Molecular Biology Information: Protein Sequence

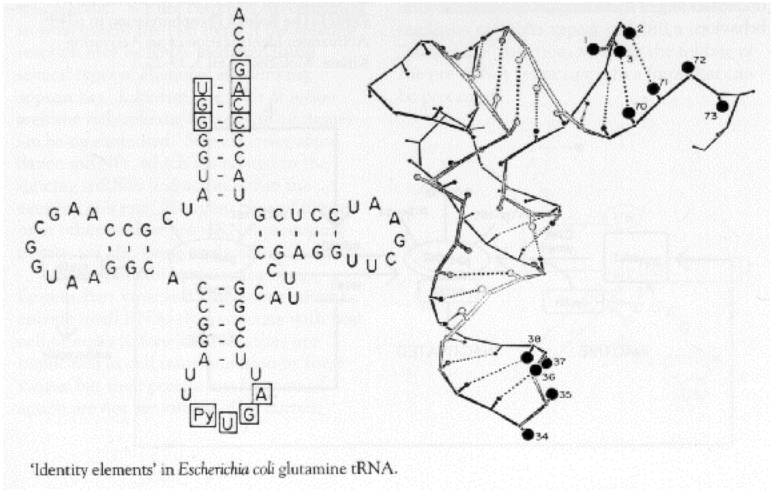
- 20 letter alphabet
  - ◊ ACDEFGHIKLMNPQRSTVWY but not BJOUXZ
- Strings of ~300 aa in an average protein (in bacteria), ~200 aa in a domain
- ~200 K known protein sequences

```
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
```

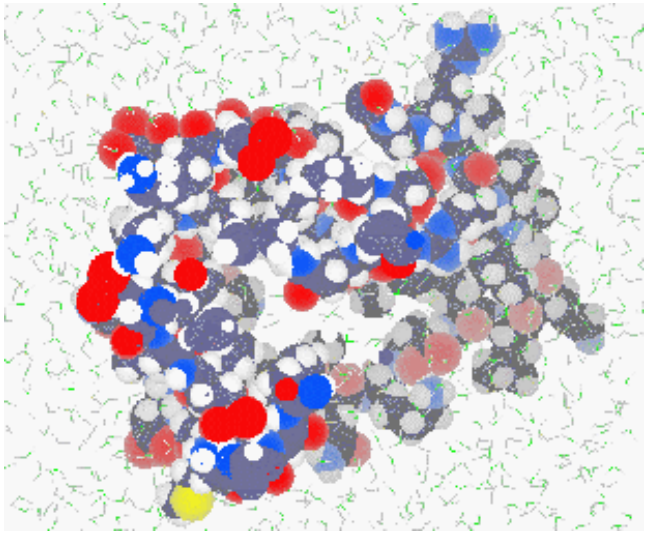
Molecular Biology Information: Macromolecular Structure

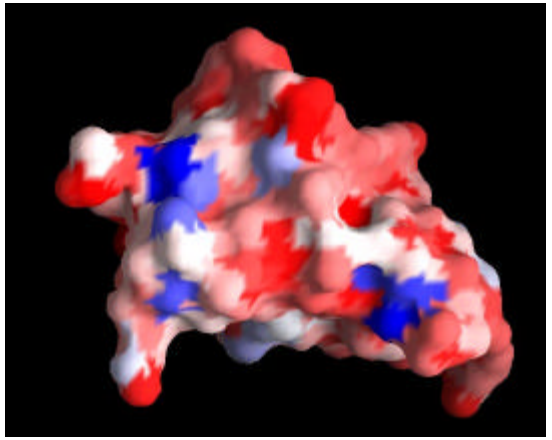
- DNA/RNA/Protein
  - ◊ Almost all protein

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



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

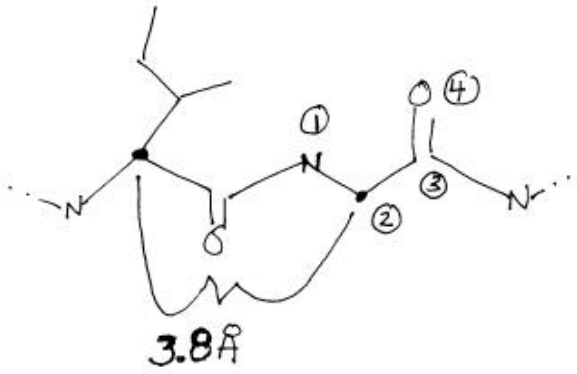




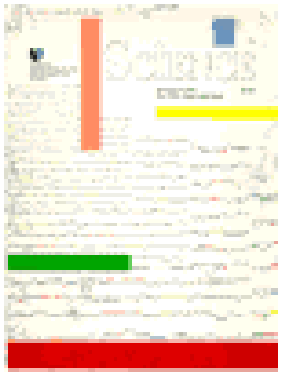
Molecular Biology Information: Protein Structure Details

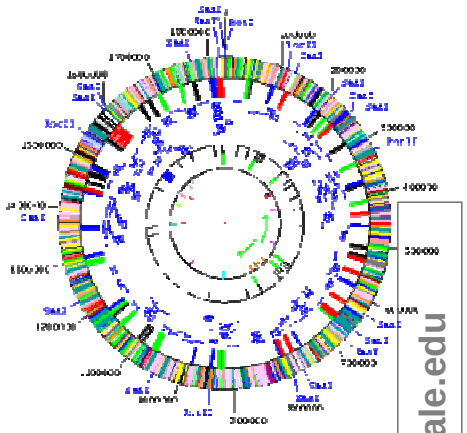
- Statistics on Number of XYZ triplets
  - ◊ 200 residues/domain -> 200 CA atoms, separated by 3.8 A
  - ◊ Avg. Residue is Leu: 4 backbone atoms + 4 sidechain atoms, 150 cubic A
    - => ~1500 xyz triplets (=8x200) per protein domain
  - ◊ 10 K known domain, ~300 folds

```
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, worm: ~100Mb with 19 K genes

  1999: >30 completed genomes!

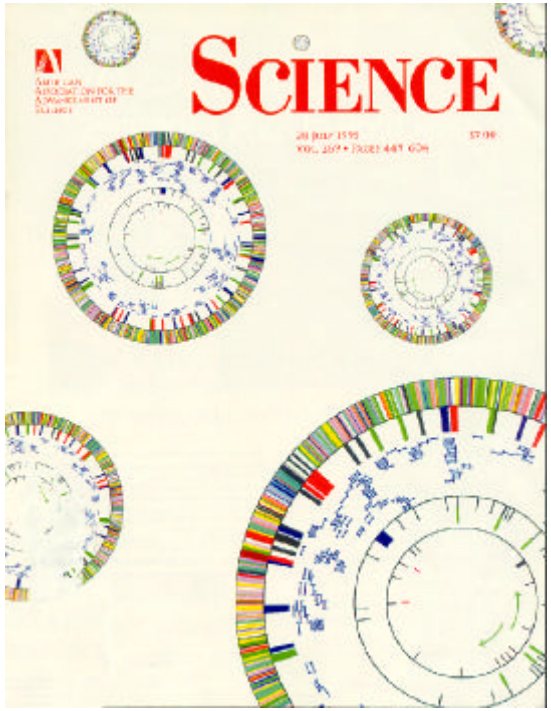
  2003, human: 3 Gb & 100 K genes...

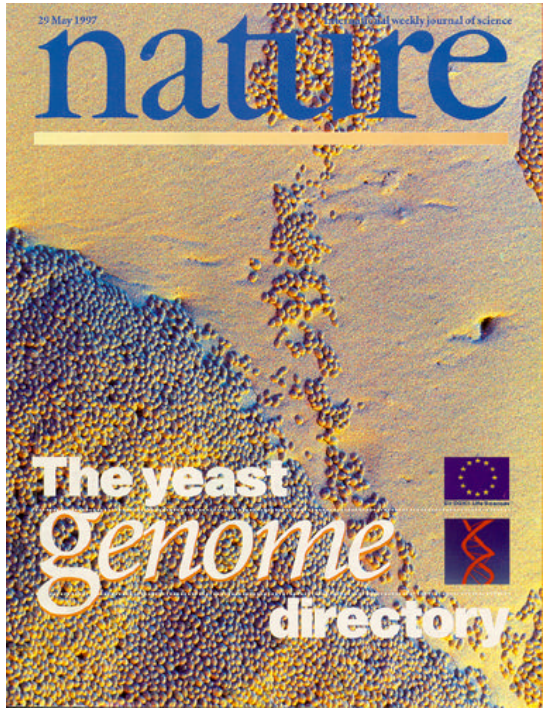
Genome sequence now accumulate so quickly that, in less than a week, a single laboratory can produce more bits of data than Shakespeare managed in a lifetime, although the latter make better reading.

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

1995

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





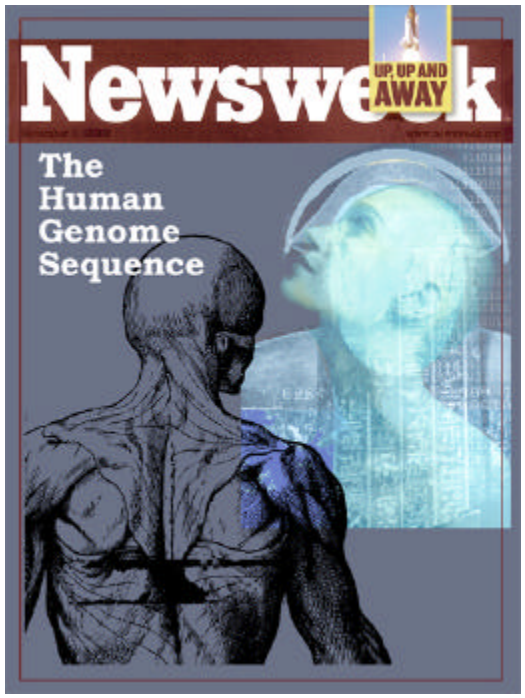
1997

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

Genomes highlight the **Finiteness** of the "Parts" in Biology

1998

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



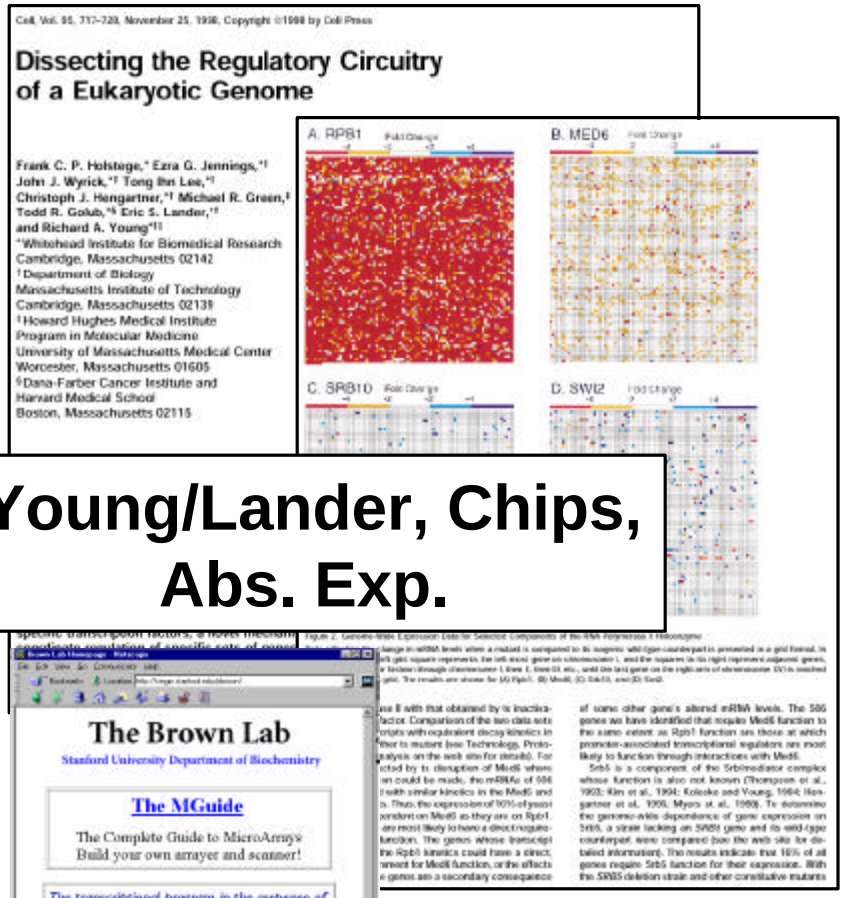
real thing, Apr '00

2000?

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

'98 spoof

Gene Expression Datasets: the Transcriptosome


Dissecting the Regulatory Circuitry of a Eukaryotic Genome

Frank C. P. Holstege,* Ezra G. Jennings,*† John J. Wyrick,*† Tong Ihn Lee,*† Christoph J. Hengartner,*† Michael R. Green,‡ Todd R. Golub,*§ Eric S. Lander,*† and Richard A. Young*†‖

*Whitehead Institute for Biomedical Research, Cambridge, Massachusetts 02142
†Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139
‡Howard Hughes Medical Institute, Program in Molecular Medicine, University of Massachusetts Medical Center, Worcester, Massachusetts 01605
§Dana-Farber Cancer Institute and Harvard Medical School, Boston, Massachusetts 02115

Young/Lander, Chips, Abs. Exp.

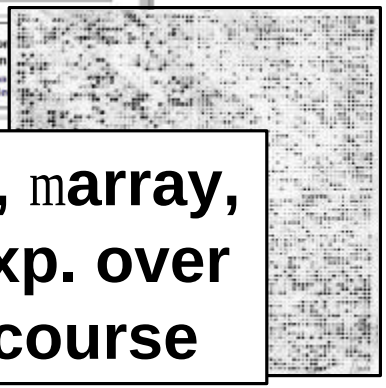
The Brown Lab

Stanford University Department of Biochemistry

The MGuide

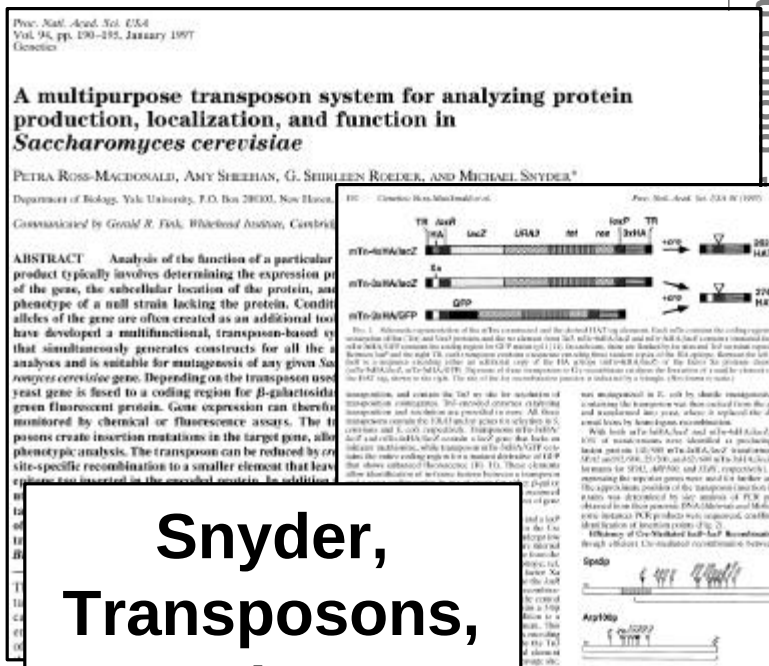
The Complete Guide to MicroArrays
Build your own arrayer and scanner!

The transcriptional program in the response of human fibroblasts to serum



Brown, marray, Rel. Exp. over Timecourse

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


Proc. Natl. Acad. Sci. USA
Vol. 94, pp. 190–195, January 1997
Genetics

A multipurpose transposon system for analyzing protein production, localization, and function in *Saccharomyces cerevisiae*

PETRA ROSS-MACDONALD, AMY SHEEHAN, G. SHIRLEEN ROEDER, AND MICHAEL SNYDER*

Snyder, Transposons, Protein Exp.

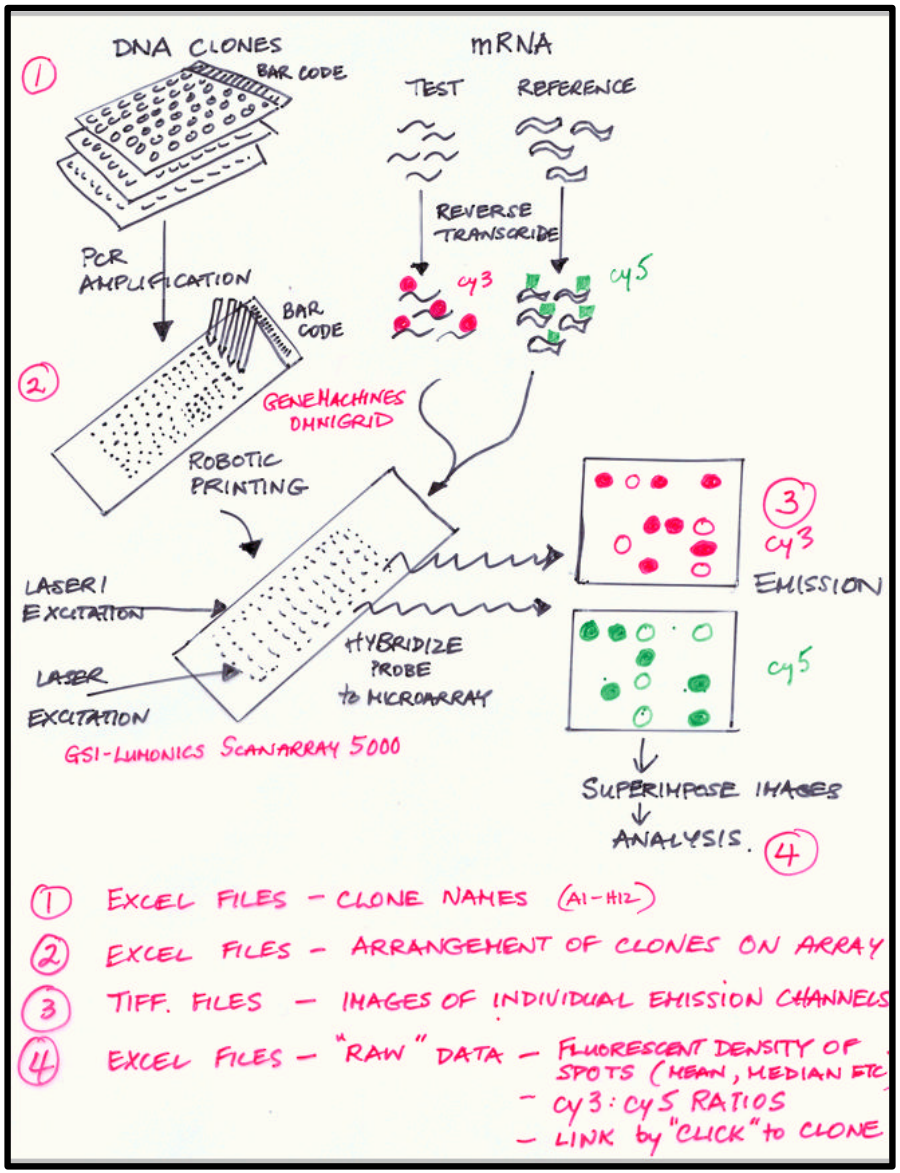
Array Data

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

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

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

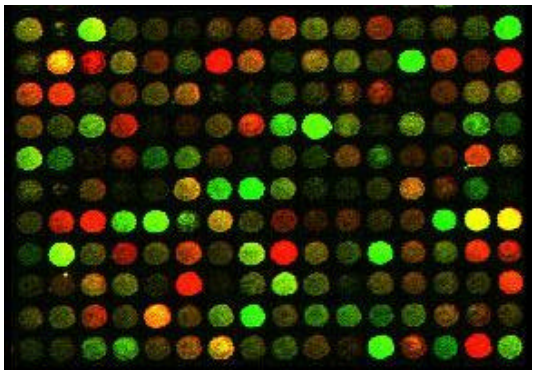
telling signal from background

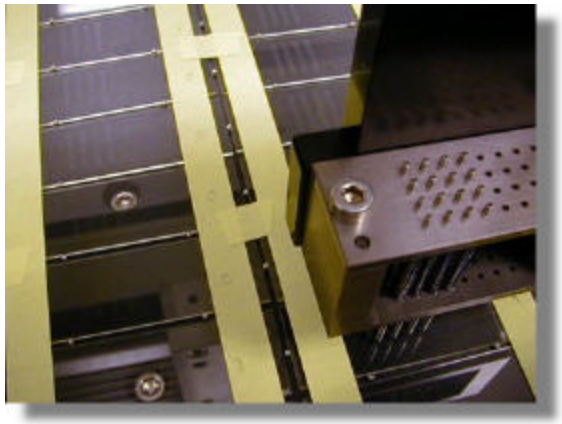


(courtesy of J Hager)

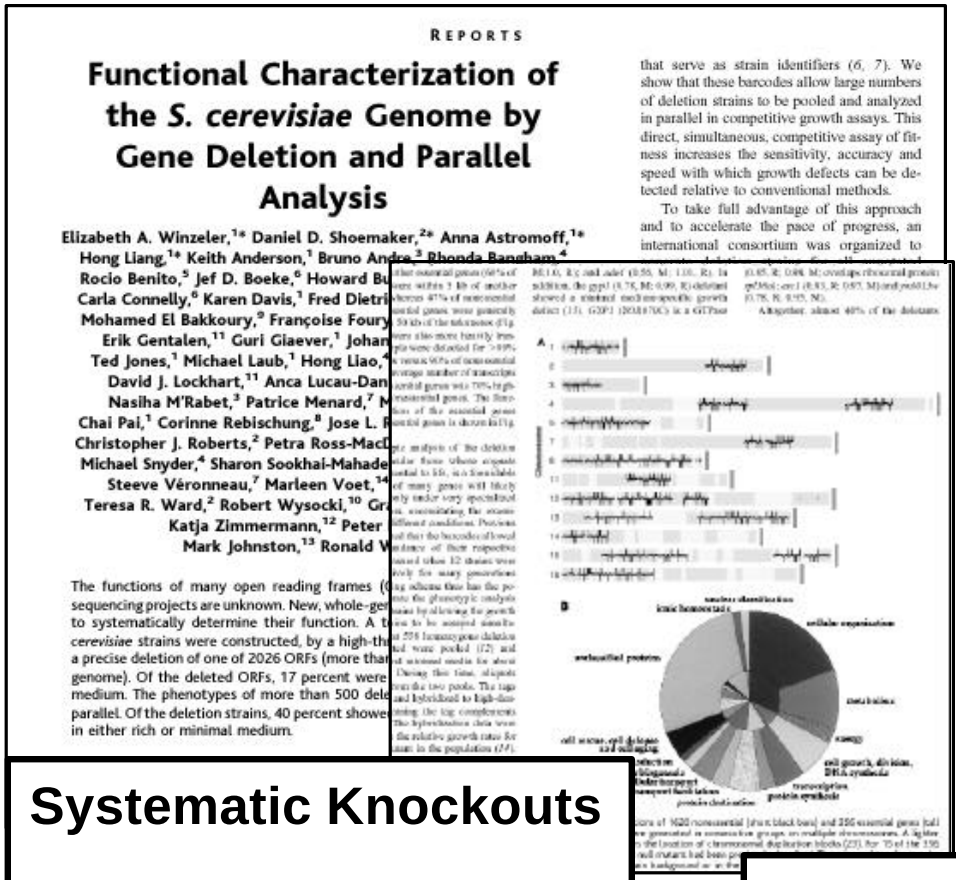
microarrays

- Affymetrix
  - Oligos
    - Don't have to know sequence
- Glass slides
  - ◊ Pat brown





Other Whole-Genome Experiments



REPORTS

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

Elizabeth A. Winzeler,[1]* Daniel D. Shoemaker,[2]* Anna Astromoff,[1]* Hong Liang,[1]* Keith Anderson,[1] Bruno Andre,[3] Rhonda Bangham,[4] Rocio Benito,[5] Jef D. Boeke,[6] Howard Bu... Carla Connelly,[6] Karen Davis,[1] Fred Dietri... Mohamed El Bakkoury,[9] Françoise Foury... Erik Gentalen,[11] Guri Giaever,[1] Johar... Ted Jones,[1] Michael Laub,[1] Hong Liao,... David J. Lockhart,[11] Anca Lucau-Dan... Nasiha M'Rabet,[3] Patrice Menard,[7] M... Chai Pai,[1] Corinne Rebischung,[8] Jose L. ... Christopher J. Roberts,[2] Petra Ross-Mac... Michael Snyder,[4] Sharon Sookhai-Mahade... Steeve Véronneau,[7] Marleen Voet,[14] Teresa R. Ward,[2] Robert Wysocki,[10] Gr... Katja Zimmermann,[12] Peter ... Mark Johnston,[13] Ronald W...

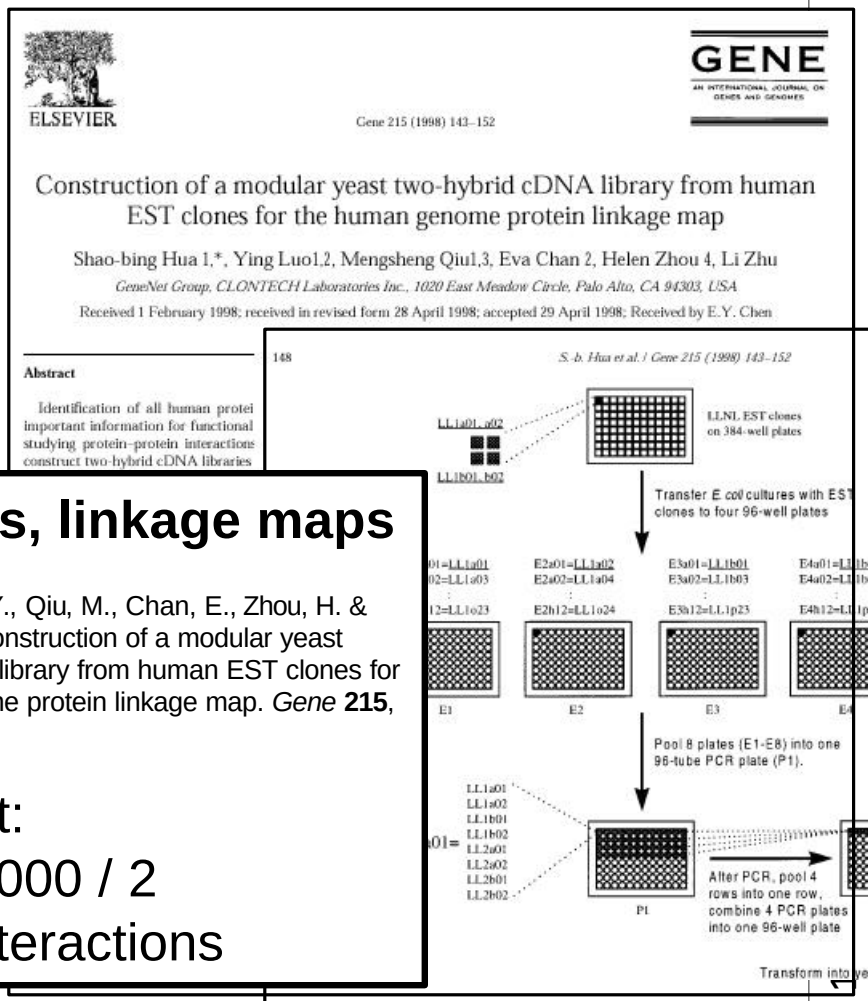
The functions of many open reading frames (... sequencing projects are unknown. New, whole-ge... to systematically determine their function. A t... *cerevisiae* strains were constructed, by a high-th... a precise deletion of one of 2026 ORFs (more than ... genome). Of the deleted ORFs, 17 percent were ... medium. The phenotypes of more than 500 dele... parallel. Of the deletion strains, 40 percent showe... in either rich or minimal medium.

that serve as strain identifiers (*6*, *7*). We show that these barcodes allow large numbers of deletion strains to be pooled and analyzed in parallel in competitive growth assays. This direct, simultaneous, competitive assay of fitness increases the sensitivity, accuracy and speed with which growth defects can be detected relative to conventional methods.

To take full advantage of this approach and to accelerate the pace of progress, an international consortium was organized to

Systematic Knockouts

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



ELSEVIER

GENE

AN INTERNATIONAL JOURNAL ON GENES AND GENOMES

Gene 215 (1998) 143–152

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

Shao-bing Hua 1,*, Ying Luo1,2, Mengsheng Qiu1,3, Eva Chan 2, Helen Zhou 4, Li Zhu

GeneNet Group, CLONTECH Laboratories Inc., 1020 East Meadow Circle, Palo Alto, CA 94303, USA

Received 1 February 1998; received in revised form 28 April 1998; accepted 29 April 1998; Received by E.Y. Chen

Abstract

Identification of all human protei... important information for functional studying protein–protein interactions construct two-hybrid cDNA libraries

148 S.-b. Hua et al. / Gene 215 (1998) 143–152

LL1a01, a02

LL1b01, b02

LLNL EST clones on 384-well plates

Transfer *E. coli* cultures with EST clones to four 96-well plates

E1a01=LL1a01 E2a01=LL1a02 E3a01=LL1b01 E4a01=LL1b02

E1a02=LL1a03 E2a02=LL1a04 E3a02=LL1b03 E4a02=LL1b04

E1h12=LL1o23 E2h12=LL1o24 E3h12=LL1p23 E4h12=LL1p24

E1 E2 E3 E4

Pool 8 plates (E1-E8) into one 96-tube PCR plate (P1).

LL1a01 LL1a02 LL1b01 LL1b02 LL2a01 LL2a02 LL2b01 LL2b02

P1

After PCR, pool 4 rows into one row, combine 4 PCR plates into one 96-well plate

Transform into yeast

2 hybrids, linkage maps

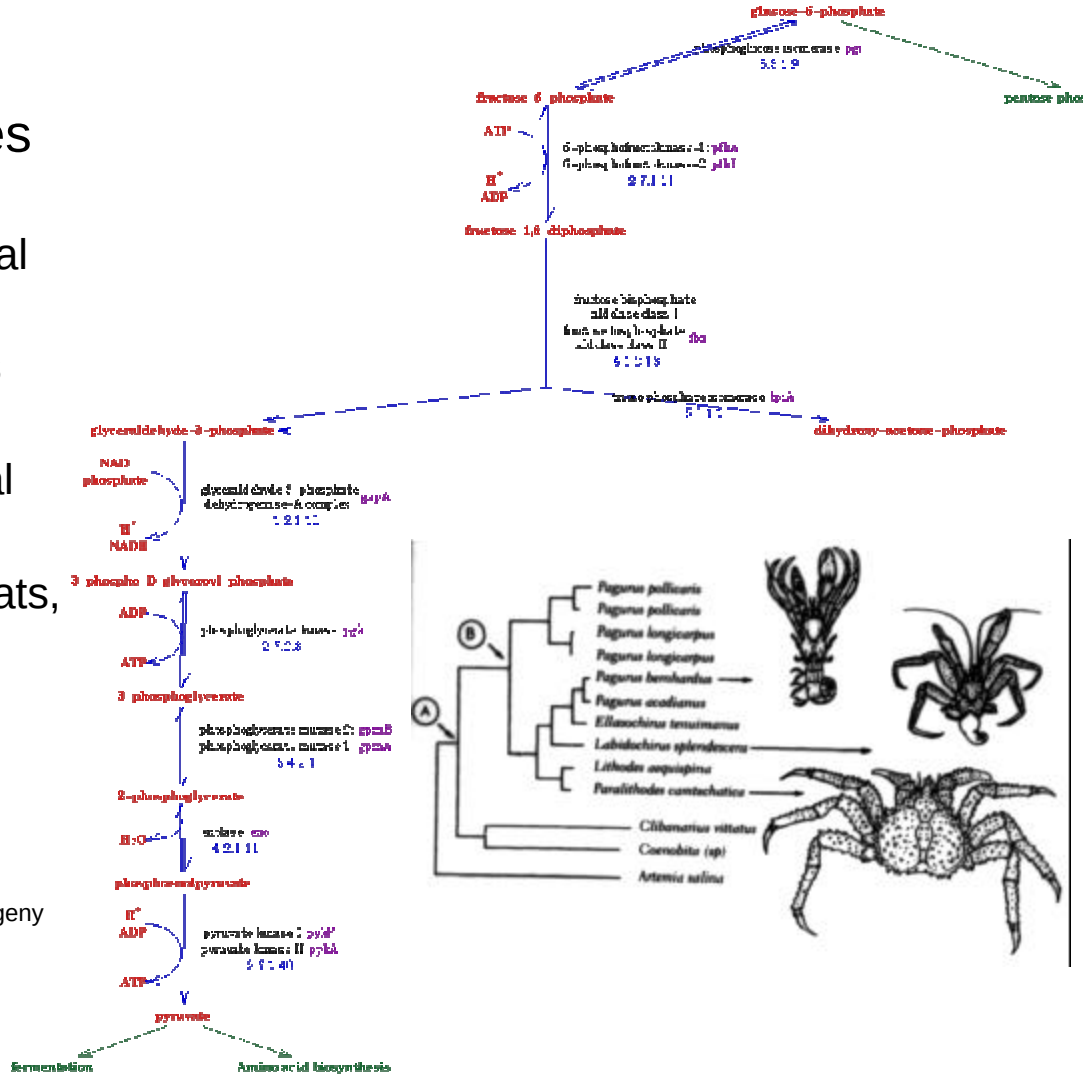
Hua, S. B., Luo, Y., Qiu, M., Chan, E., Zhou, H. & Zhu, L. (1998). Construction of a modular yeast two-hybrid cDNA library from human EST clones for the human genome protein linkage map. *Gene* **215**, 143-52

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

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....

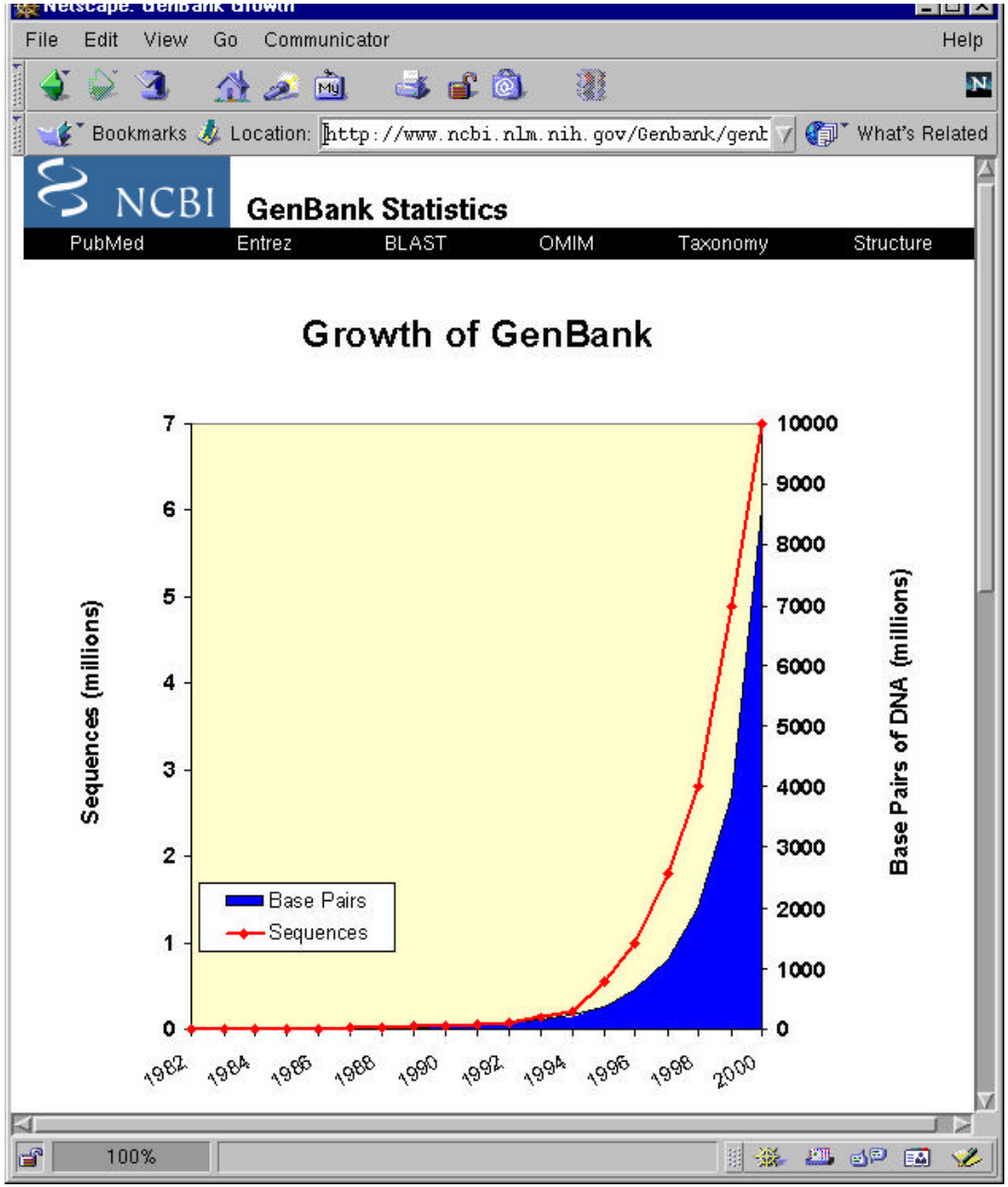
(Pathway drawing from P Karp's EcoCyc, Phylogeny from S J Gould, Dinosaur in a Haystack)



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 "MIS" for Molecular Biology Information. It is a practical discipline with many **applications**.

Large-scale Information: GenBank Growth



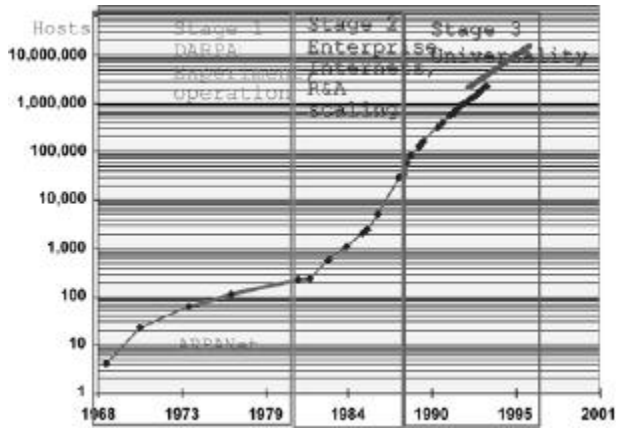
GenBank Data		
Year	**Base Pairs**	**Sequences**
1982	680338	606
1983	2274029	2427
1984	3368765	4175
1985	5204420	5700
1986	9615371	9978
1987	15514776	14584
1988	23800000	20579
1989	34762585	28791
1990	49179285	39533
1991	71947426	55627
1992	101008486	78608
1993	157152442	143492
1994	217102462	215273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

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

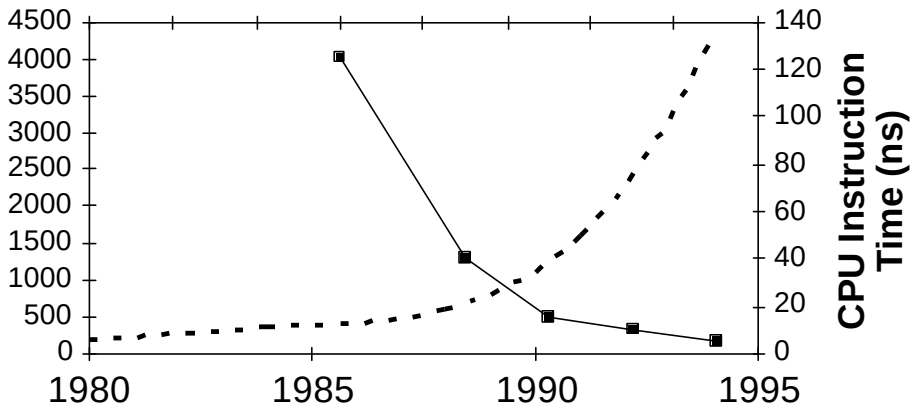
- CPU vs Disk & Net
  - ◊ As important as the increase in computer speed has been, the ability to store large amounts of information on computers is even more crucial
- Driving Force in Bioinformatics

(Internet picture adapted from D Brutlag, Stanford)

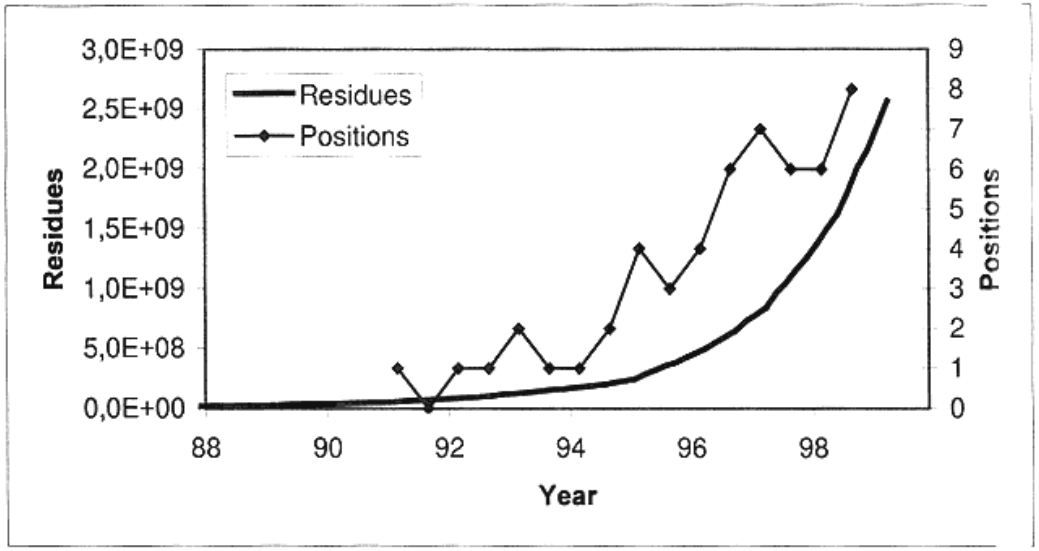
Internet Hosts



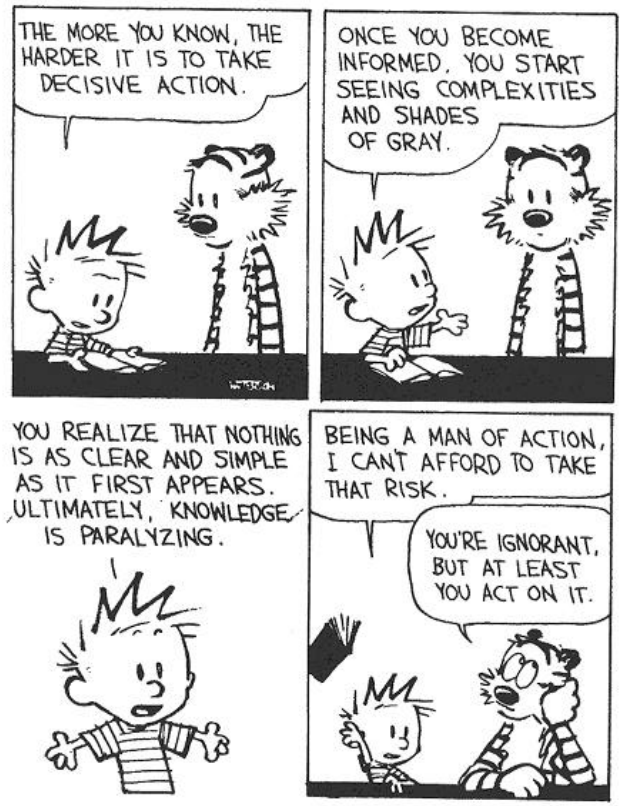
Num. Protein Domain Structures



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.



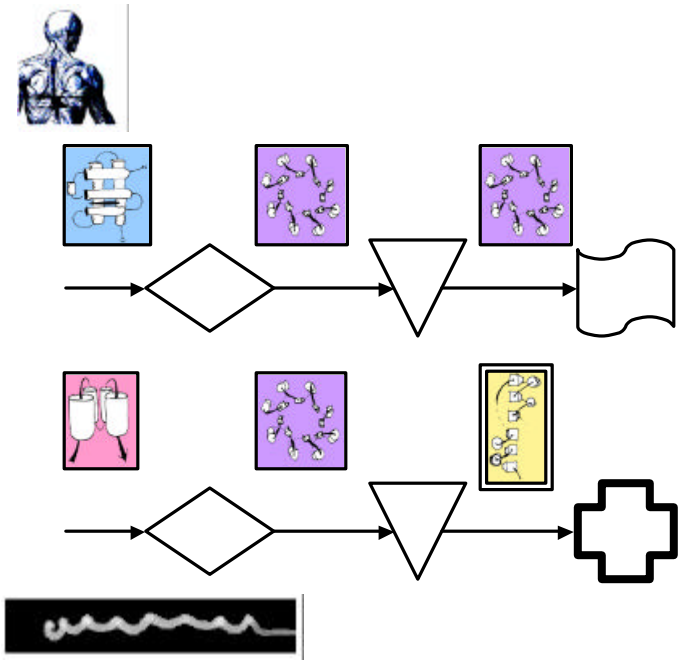
B. Watterson, "*There's treasure everywhere*", Andrews and McMeel, 1996.

(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 "MIS" for Molecular Biology Information. It is a practical discipline with many **applications**.

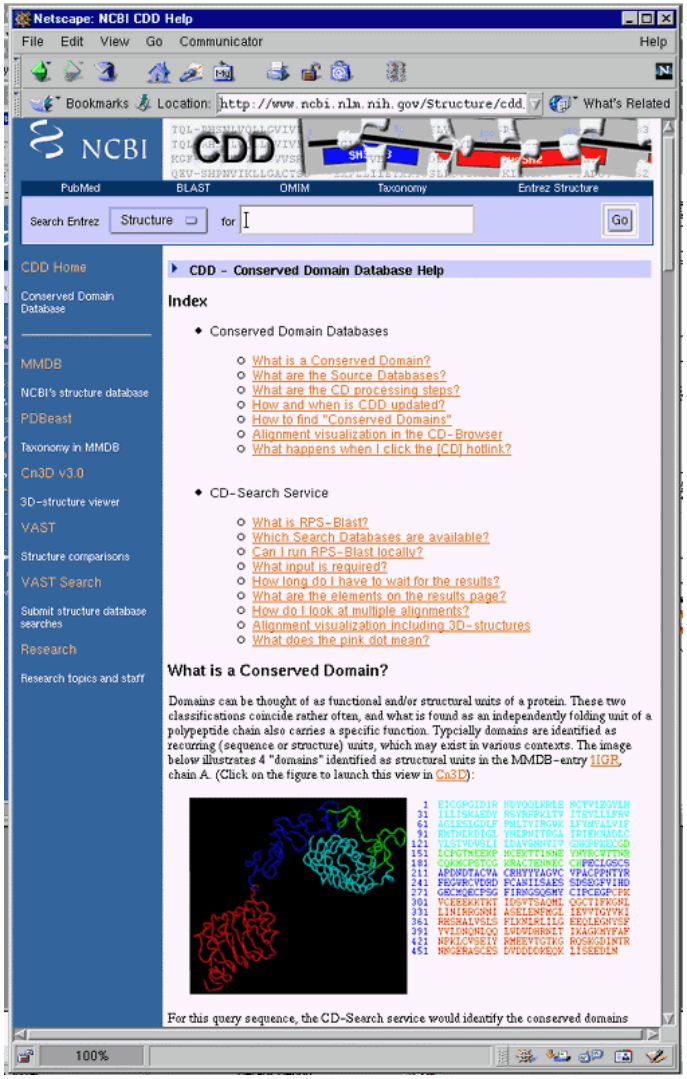
Organizing Molecular Biology Information: Redundancy and Multiplicity

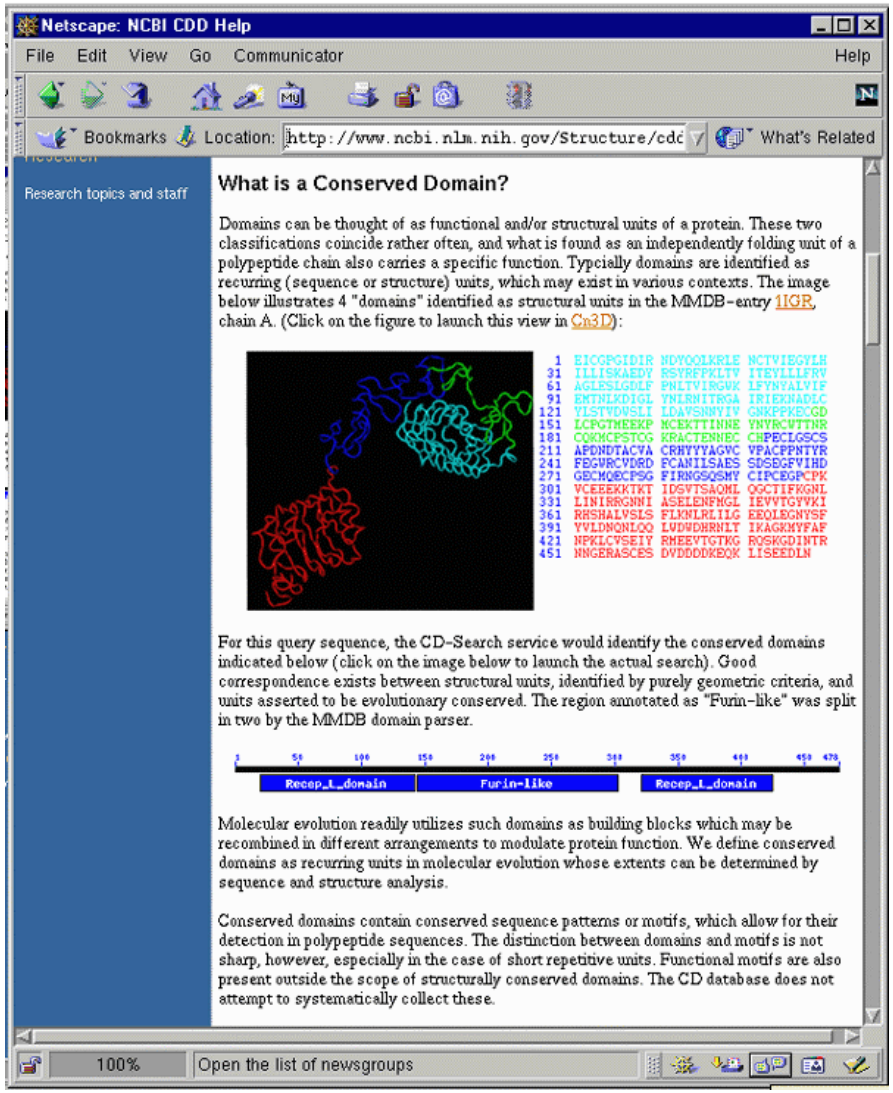


- Different Sequences Have the Same Structure
- Organism has many similar genes
- Single Gene May Have Multiple Functions
- Genes are grouped into Pathways
- Genomic Sequence Redundancy due to the Genetic Code
- **How do we find the similarities?.....**

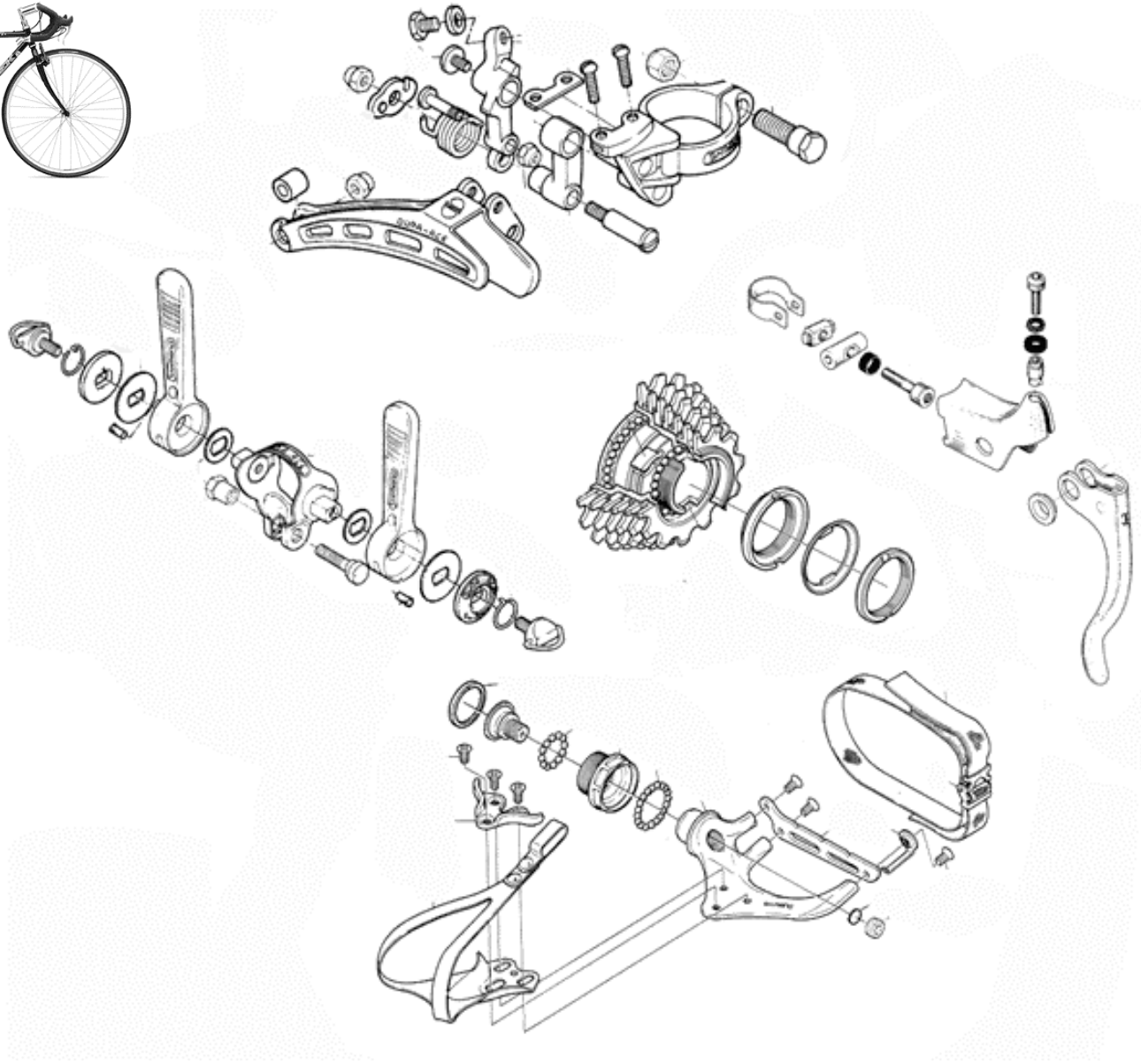
Integrative Genomics - genes ↔ structures ↔ **functions** ↔ **pathways** ↔ expression levels ↔ regulatory systems ↔ ….

Molecular Parts = Conserved Domains, Folds, &c

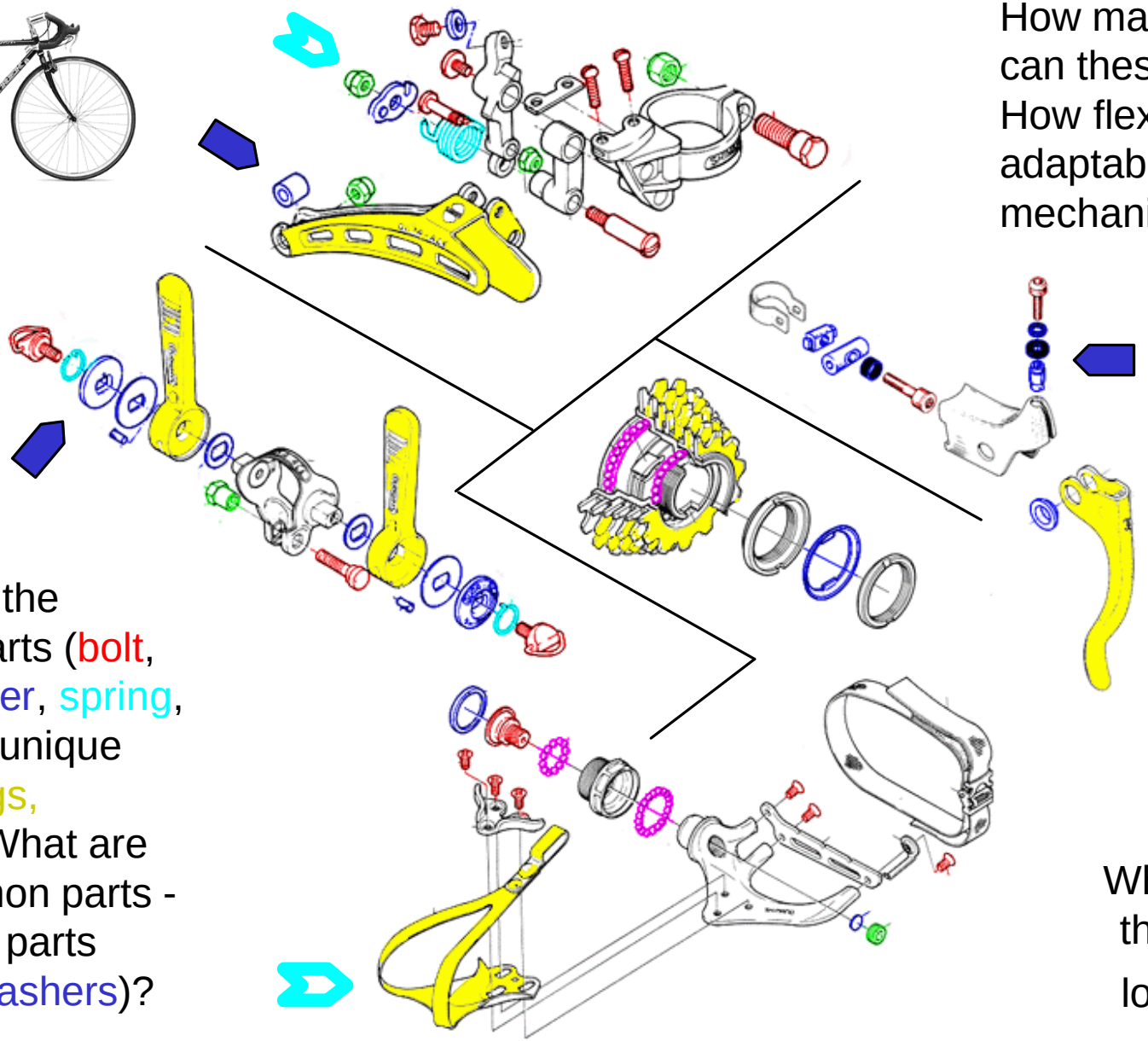




A Parts List Approach to Bike Maintenance



A Parts List Approach to Bike Maintenance

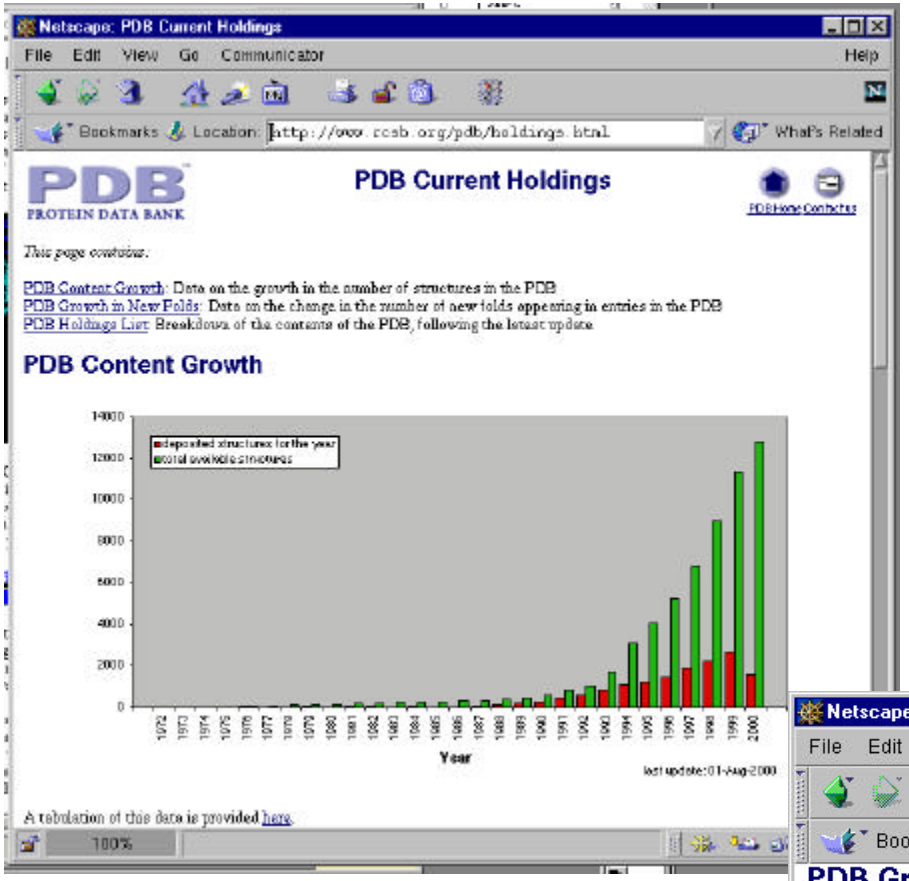


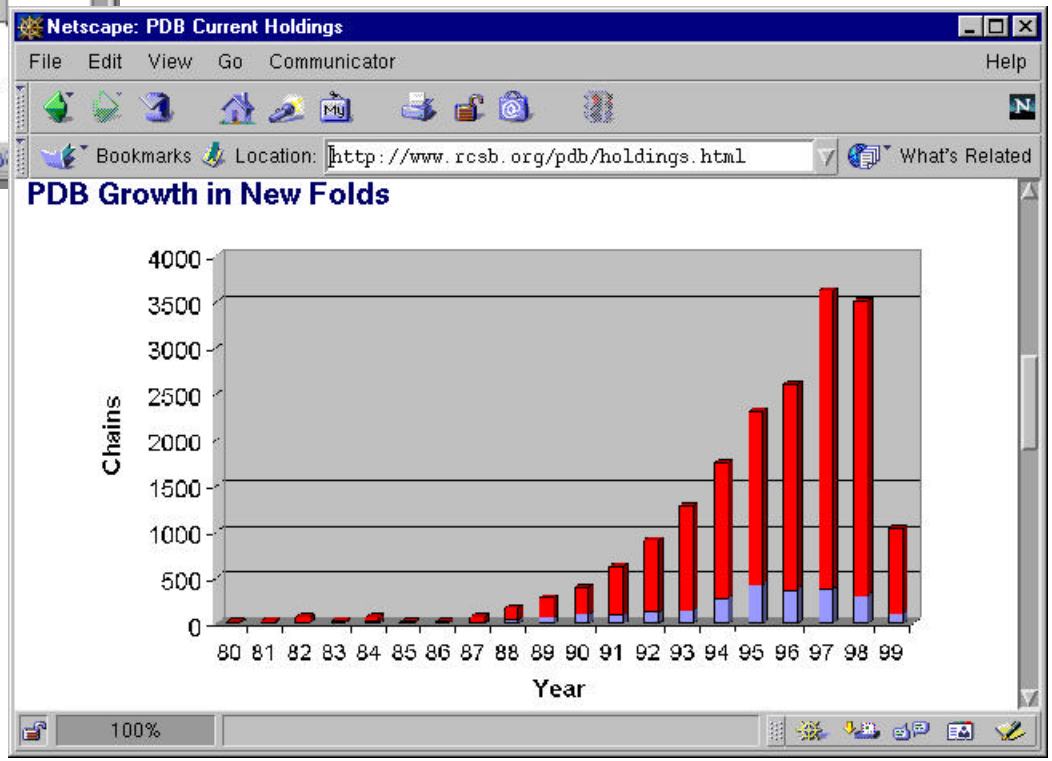
How many roles can these play? How flexible and adaptable are they mechanically?

What are the shared parts (bolt, nut, washer, spring, bearing), unique parts (cogs, levers)? What are the common parts -- types of parts (nuts & washers)?

Where are the parts located?

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

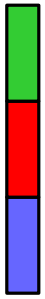




Total in Databank

New Submissions

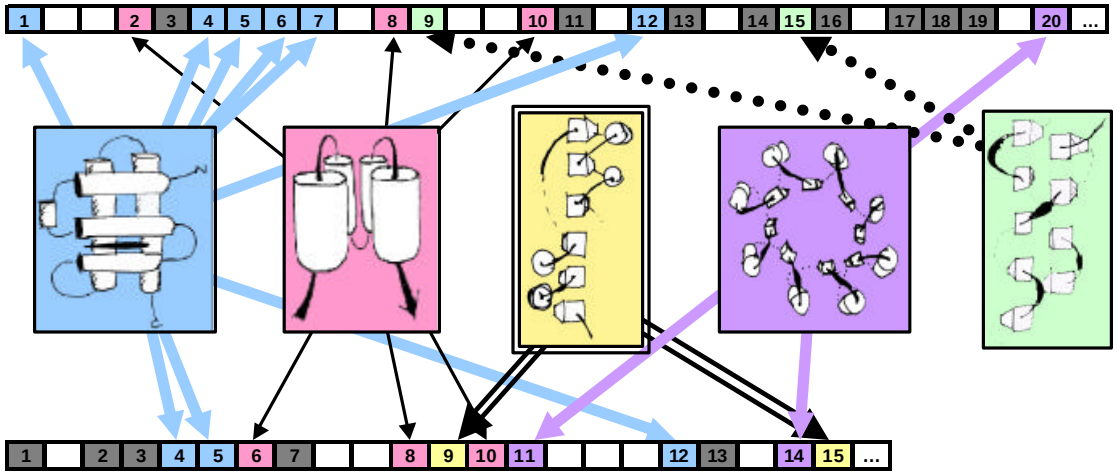
New Folds



World of Structures is even more Finite, providing a valuable simplification

(human)

(*T. pallidum*)



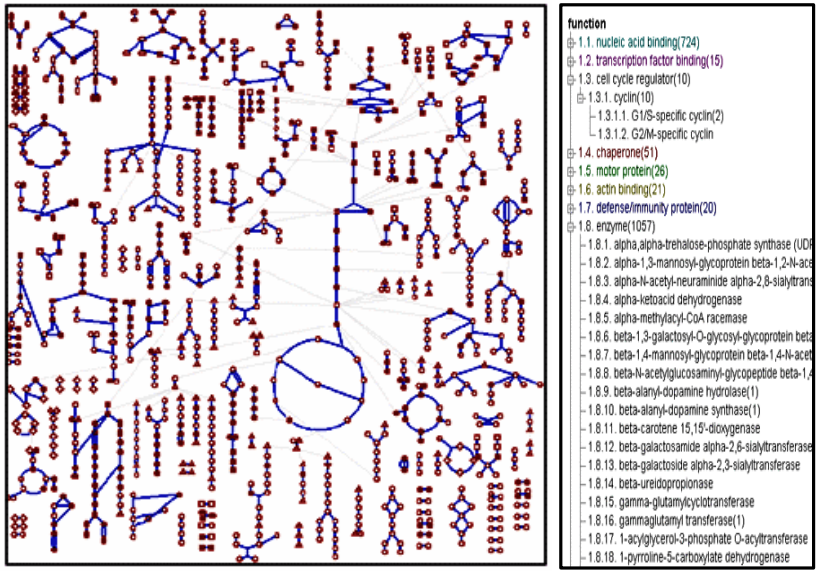
~100000 genes

~1000 folds

~1000 genes

Same logic for pathways, functions, sequence families, blocks, motifs....

Global Surveys of a Finite Set of Parts from Many Perspectives



Functions picture from www.fruitfly.org/~suzi (Ashburner); Pathways picture from, ecocyc.pangeasystems.com/ecocyc (Karp, Riley). Related resources: COGS, ProDom, Pfam, Blocks, Domo, WIT, CATH, Scop....

What is Bioinformatics?

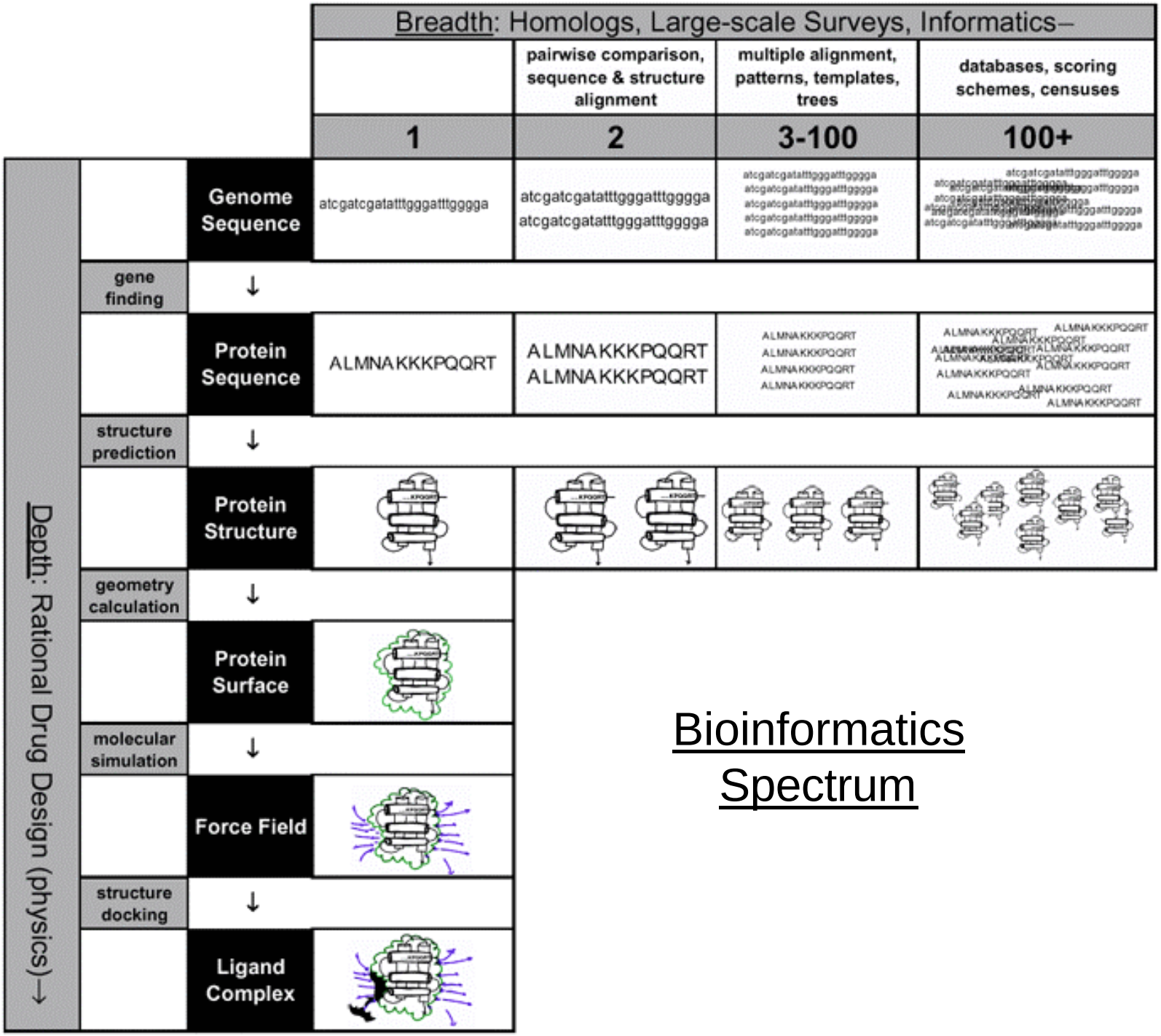
- *(Molecular)* **Bio** - **informatics**
- One idea for a definition?
  Bioinformatics is conceptualizing **biology in terms of molecules** (in the sense of physical-chemistry) and then applying **"informatics" techniques** (derived from disciplines such as applied math, CS, and statistics) to understand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

General Types of "Informatics" techniques in Bioinformatics

- Text String Comparison
  - ◊ Optimal 1D Alignment, Probabilistic Patterns
  - ◊ Rapid non-exact search (Alta Vista, grep)
  - ◊ Significance Statistics
- Databases
  - ◊ Designing Building, Querying
  - ◊ Complex Data: Object DB, Ontologies
  - ◊ Integrative Analysis and Surveys
- Datamining
  - ◊ General Machine Learning
  - ◊ Clustering, Trees, PCA
  - ◊ Bayesian Networks, NNs, kNNs
- Geometry
  - ◊ Robotics
  - ◊ Graphics (Surfaces, Voronoi Volumes)
  - ◊ Comparison and 3D Matching (Vision, recognition)
- Physical Simulation
  - ◊ Representing Interactions: Electrostatics, QM
  - ◊ Numerical Algorithms for optimization (MC, SA)
  - ◊ Newtonian Mechanics, Macromolecular Simulation

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



Bioinformatics Spectrum

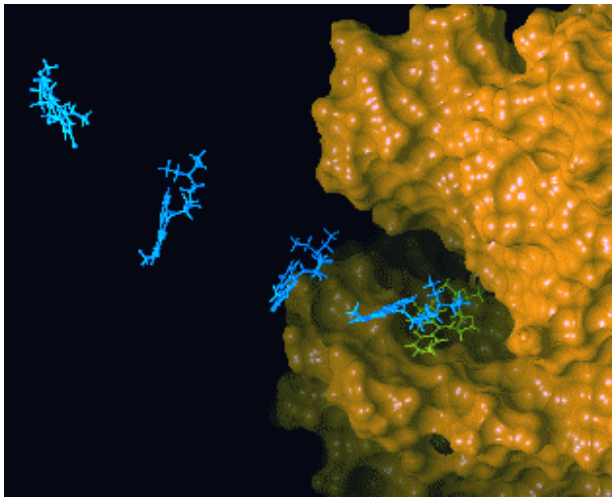
What is Bioinformatics?

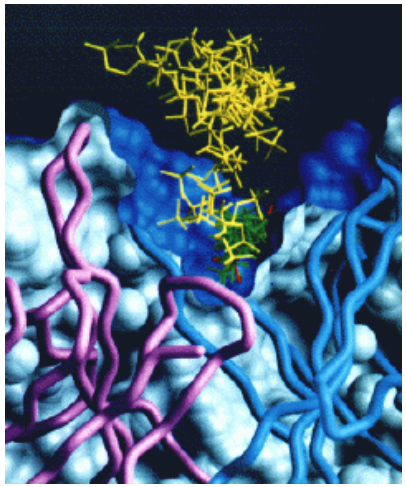
- *(Molecular)* **Bio** - **`informatics`**
- One idea for a definition?
  Bioinformatics is conceptualizing **biology in terms of molecules** (in the sense of physical-chemistry) and then applying **"informatics" techniques** (derived from disciplines such as applied math, CS, and statistics) to understand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

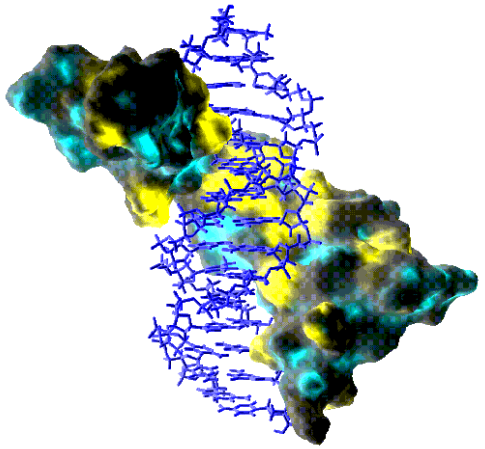
Major Application I: Designing Drugs

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

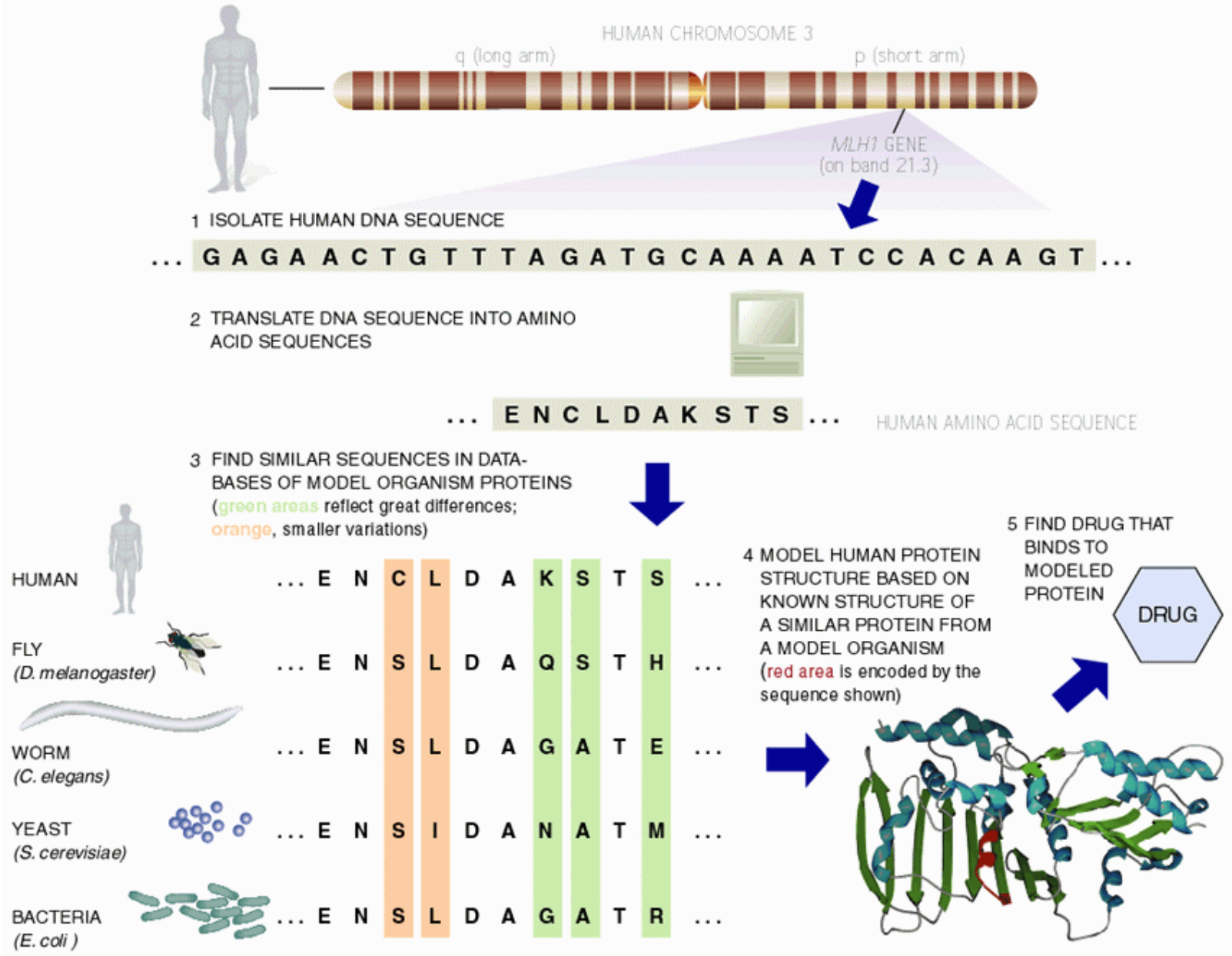
(From left to right, figures adapted from Olsen Group Docking Page at Scripps, Dyson NMR Group Web page at Scripps, and from Computational Chemistry Page at Cornell Theory Center).







Major Application II: Finding Homologs



Major Application II: Finding Homologues

- Find Similar Ones in Different Organisms
- Human vs. Mouse vs. Yeast
  - ◊ Easier to do Expts. on latter!

(Section from NCBI Disease Genes Database Reproduced Below.)

Best Sequence Similarity Matches to Date Between Positionally Cloned Human Genes and S. cerevisiae Proteins

Human Disease	MIM #	Human Gene	GenBank Acc# for Human cDNA	BLASTX P-value	Yeast Gene	GenBank Acc# for Yeast cDNA	Yeast Gene Description
Hereditary Non-polyposis Colon Cancer	120436	MSH2	U03911	9.2e-261	MSH2	M84170	DNA repair protein
Hereditary Non-polyposis Colon Cancer	120436	MLH1	U07418	6.3e-196	MLH1	U07187	DNA repair protein
Cystic Fibrosis	219700	CFTR	M28668	1.3e-167	YCF1	L35237	Metal resistance protein
Wilson Disease	277900	WND	U11700	5.9e-161	CCC2	L36317	Probable copper transporter
Glycerol Kinase Deficiency	307030	GK	L13943	1.8e-129	GUT1	X69049	Glycerol kinase
Bloom Syndrome	210900	BLM	U39817	2.6e-119	SGS1	U22341	Helicase
Adrenoleukodystrophy, X-linked	300100	ALD	Z21876	3.4e-107	PXA1	U17065	Peroxisomal ABC transporter
Ataxia Telangiectasia	208900	ATM	U26455	2.8e-90	TEL1	U31331	PI3 kinase
Amyotrophic Lateral Sclerosis	105400	SOD1	K00065	2.0e-58	SOD1	J03279	Superoxide dismutase
Myotonic Dystrophy	160900	DM	L19268	5.4e-53	YPK1	M21307	Serine/threonine protein kinase
Lowe Syndrome	309000	OCRL	M88162	1.2e-47	YIL002C	Z47047	Putative IPP-5-phosphatase
Neurofibromatosis, Type 1	162200	NF1	M89914	2.0e-46	IRA2	M33779	Inhibitory regulator protein
Choroideremia	303100	CHM	X78121	2.1e-42	GDI1	S69371	GDP dissociation inhibitor
Diastrophic Dysplasia	222600	DTD	U14528	7.2e-38	SUL1	X82013	Sulfate permease
Lissencephaly	247200	LIS1	L13385	1.7e-34	MET30	L26505	Methionine metabolism
Thomsen Disease	160800	CLC1	Z25884	7.9e-31	GEF1	Z23117	Voltage-gated chloride channel
Wilms Tumor	194070	WT1	X51630	1.1e-20	FZF1	X67787	Sulphite resistance protein
Achondroplasia	100800	FGFR3	M58051	2.0e-18	IPL1	U07163	Serine/threoinine protein kinase
Menkes Syndrome	309400	MNK	X69208	2.1e-17	CCC2	L36317	Probable copper transporter

Major Application II: Finding Homologues (cont.)

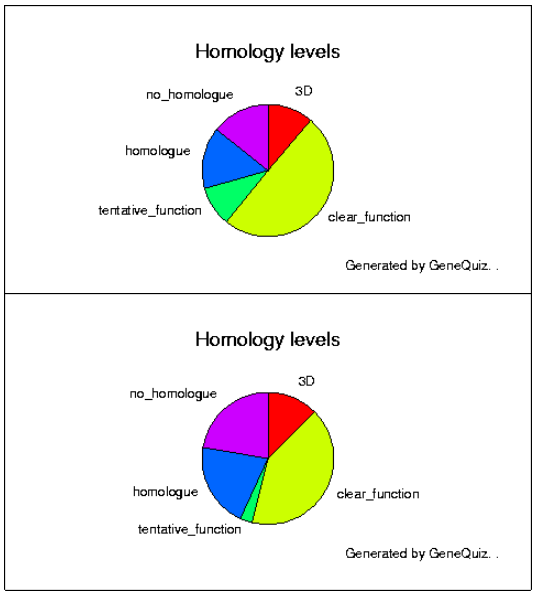
- Cross-Referencing, one thing to another thing
- Sequence Comparison and Scoring
- Analogous Problems for Structure Comparison
- Comparison has two parts:
  (1) Optimally **Aligning** 2 entities to get a Comparison **Score**
  (2) Assessing **Significance** of this score in a given **Context**

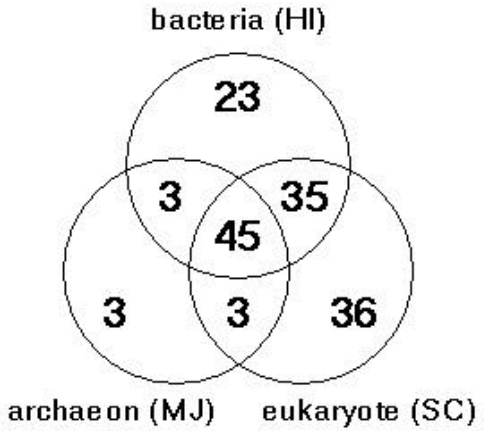
- **Integrated Presentation**
  ◊ Align Sequences
  ◊ Align Structures
  ◊ Score in a Uniform Framework

Major Application I|I:
Overall Genome Characterization

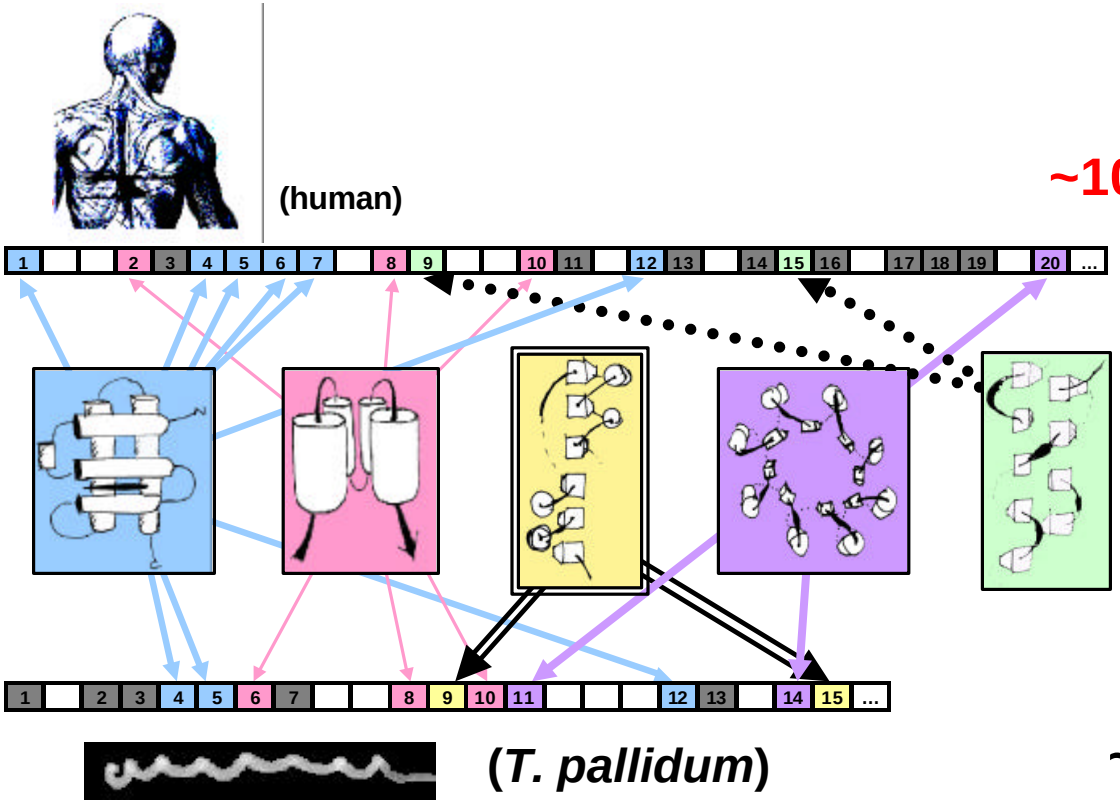
- Overall Occurrence of a Certain Feature in the Genome
  - ◊ e.g. how many kinases in Yeast
- Compare Organisms and Tissues
  - ◊ Expression levels in Cancerous vs Normal Tissues
- Databases, Statistics

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





Simplfying Genomes with Folds, Pathways, &c

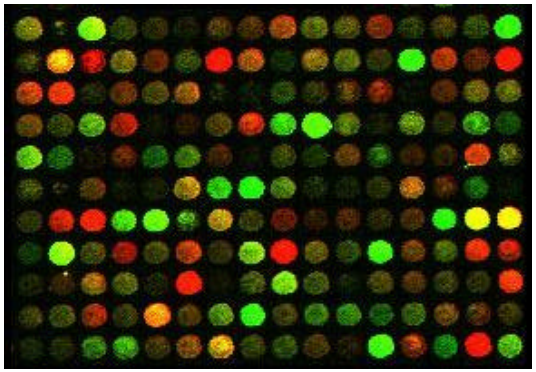


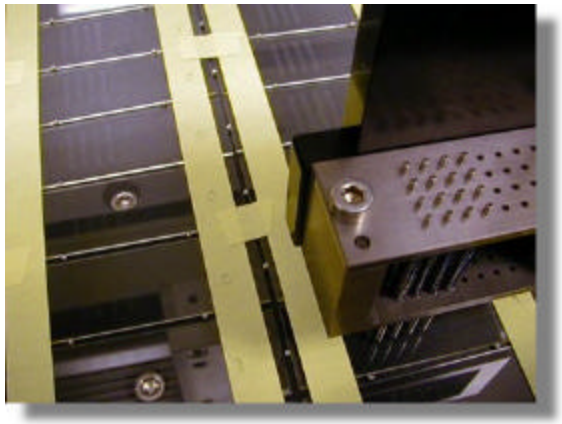
Bioinformatics

- A Very Broad Overview: What is Bioinformatics?
  - ◊ Types of Information, Organizing Principles, Informatics Techniques, Real-world Applications
- Example Calculation 1: Datamining Genome Information
  - ◊ Representing expression data and other features in high-dimensional space; Discriminants
  - ◊ Simple Bayesian analysis
- Example Calculation 2: Aligning Text Strings
  - ◊ Simple dynamic programming
  - ◊ Adding in gaps and other complexities

microarrays

- Affymetrix
  - Oligos
    - Don’t have to know sequence
- Glass slides
  - ◊ Pat brown





Typical Predictors and Response for Yeast

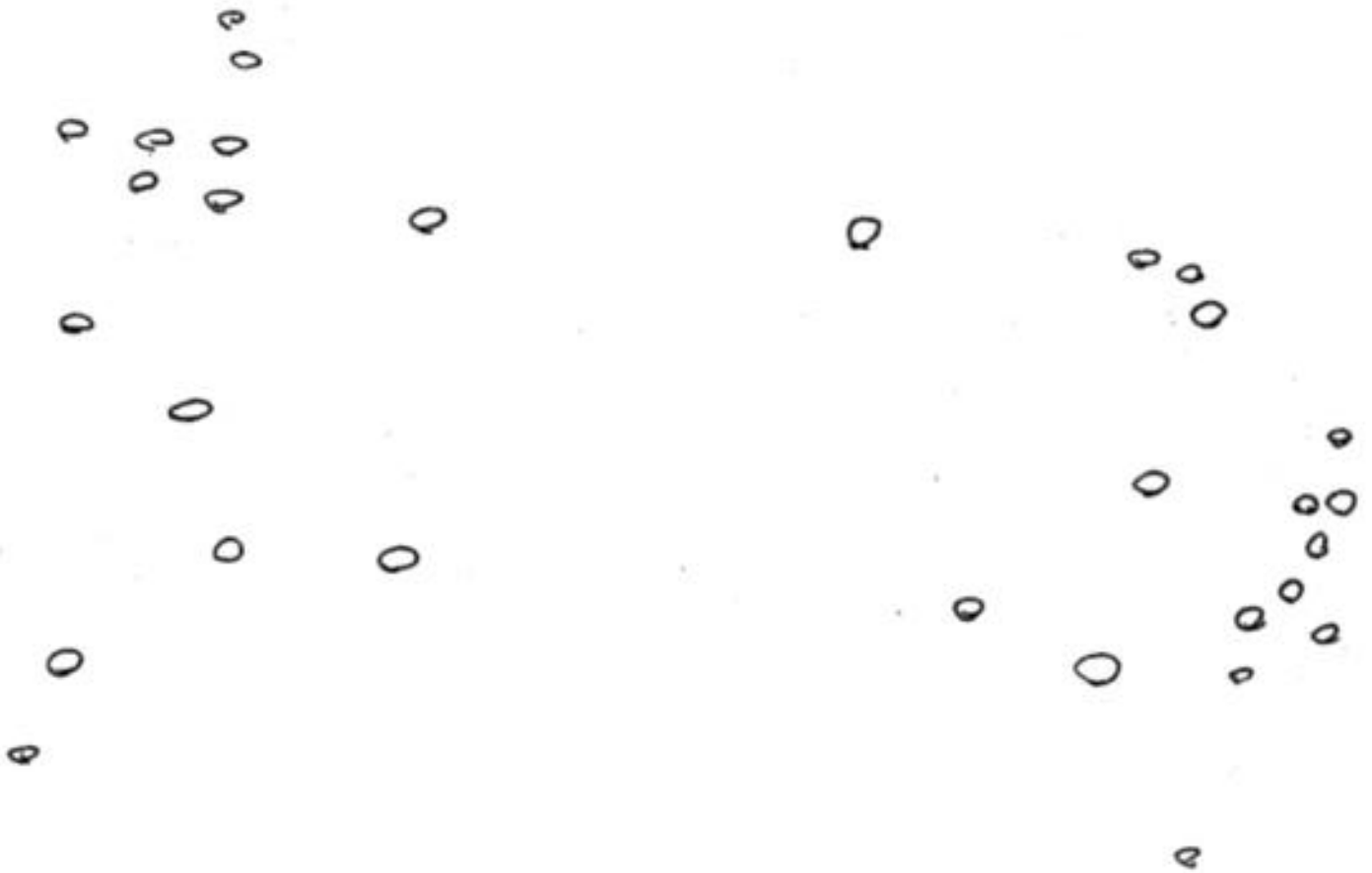
Basics		Predictors																				Response			
		Sequence Features													Genomic Features										
		seq. length	Amino Acid Composition							How many times does the sequence have these motif features?						Abs. expr. Level (mRNA copies / cell)		Prot. Abun-dance	Cell cycle timecourse				Function		Localization
Yeast Gene ID	Sequence		A	C	D	E	F	W	Y	farn site	NLS	hdel motif		nuc2	signalp	tms1	Gene-Chip expt. from RY Lab	sage tag freq.	(1000 copies /cell)	t=0	t=1	t=15	t=16	function ID(s) (from MIPS)	function description	5-compartment
YAL001C	MNIFEMLRI	1160	.08	.02	.06			.01	.04	0	1	0		1	0	0	0.3	0	?	5	3	4	5	04.01.01;04.03	TFIIIC (transcription initia	N
YAL002W	KVFGRCELA	1176	.09	.02	.06			.01	.04	0	0	0		0	0	1	0.2	?	?	8	4	4	3	06.04;08.13	vacuolar sorting protein,	C
YAL003W	KMLQFNLRW	206	.08	.02	.06			.01	.04	0	0	0		0	0	0	19.1	19	23	70	73	98	126	05.04;30.03	translation elongation fac	N
YAL004W	RPDFCLEPP	215	.08	.02	.06			.01	.04	0	0	0		0	0	0	?	0	?	18	12	4	6	01.01.01	0	N
YAL005C	VINTFDGVA	641	.08	.02	.06			.01	.04	0	0	0		0	0	1	13.4	16	17	39	38	8	14	06.01;06.04;08	heat shock protein of HS	????
YAL007C	KKAVINGEQ	190	.08	.02	.06			.01	.04	0	0	0		0	1	4	2.2	8	?	15	20	16	17	99	????	????
YAL008W	HPETLVKVK	198	.08	.02	.06			.01	.04	0	0	0		0	0	3	1.2	?	?	9	6	2	3	99	????	????
YAL009W	PTLEWFLSH	259	.08	.02	.06			.01	.04	0	2	0		0	0	3	0.6	?	?	6	2	3	5	03.10;03.13	meiotic protein	????
YAL010C	MEQRITLKD	493	.08	.02	.06			.02	.04	0	0	0		0	0	1	0.3	?	?	11	6	6	6	30.16	involved in mitochondrial	????
YAL011W	KSFPEVVGK	616	.08	.02	.06			.01	.04	0	8	0		1	0	0	0.4	?	?	6	5	5	6	30.16;99	protein of unknown funct	????
YAL012W	GVQVETISP	393	.08	.02	.06			.01	.04	0	0	0		0	0	1	8.9	4	6.7	29	26	23	29	01.01.01;30.03	cystathionine gamma-lya	C
YAL013W	RTDCYGNVN	362	.08	.02	.06			.01	.04	0	0	0		0	0	0	0.6	?	?	7	9	6	10	01.06.10;30.03	regulator of phospholipid	N
YAL014C	GDVEKGKKI	202	.08	.02	.06			.01	.04	0	0	0		0	0	0	1.1	?	?	12	13	9	12	99	????	N
YAL015C	MTPAVTTYK	399	.08	.02	.06			.01	.04	0	1	0		0	0	0	0.7	0	1	19	18	12	13	11.01;11.04	DNA repair protein	N
YAL016W	KKPLTQEQL	635	.08	.02	.06			.01	.04	0	0	0		0	0	1	3.3	5	?	15	20	16	16	03.01;03.04;03	ser/thr protein phosphata	????

Arrange data in a tabulated form, each row representing an example and each column representing a feature, including the dependent experimental quantity to be predicted.

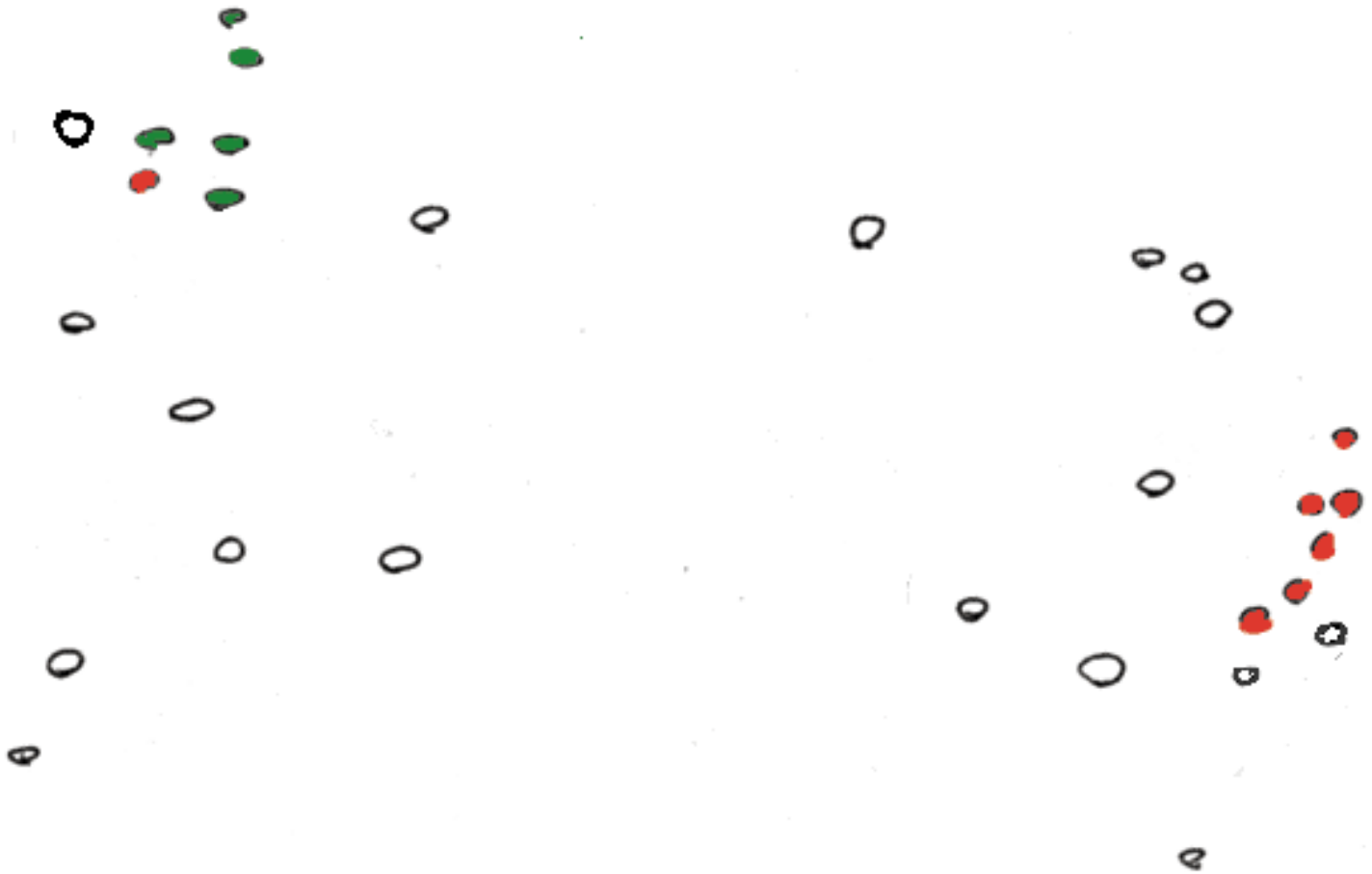
	predictor1	Predictor2	predictor3	predictor4	response
G1	A(1,1)	A(1,2)	A(1,3)	A(1,4)	Class A
G2	A(2,1)	A(2,2)	A(2,3)	A(2,4)	Class A
G3	A(3,1)	A(3,2)	A(3,3)	A(3,4)	Class B

(adapted from Y Kluger)

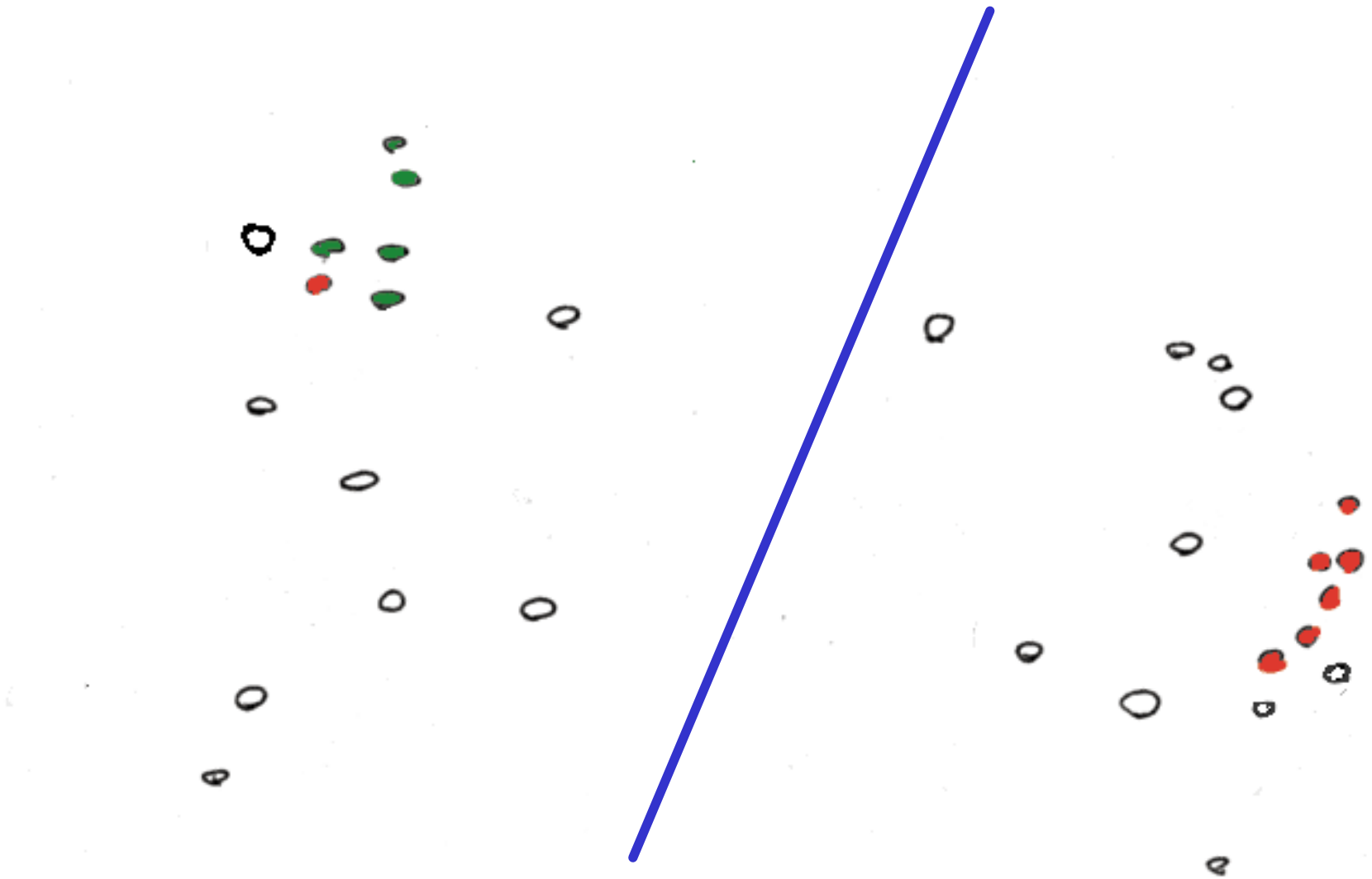
Represent predictors in abstract high dimensional space



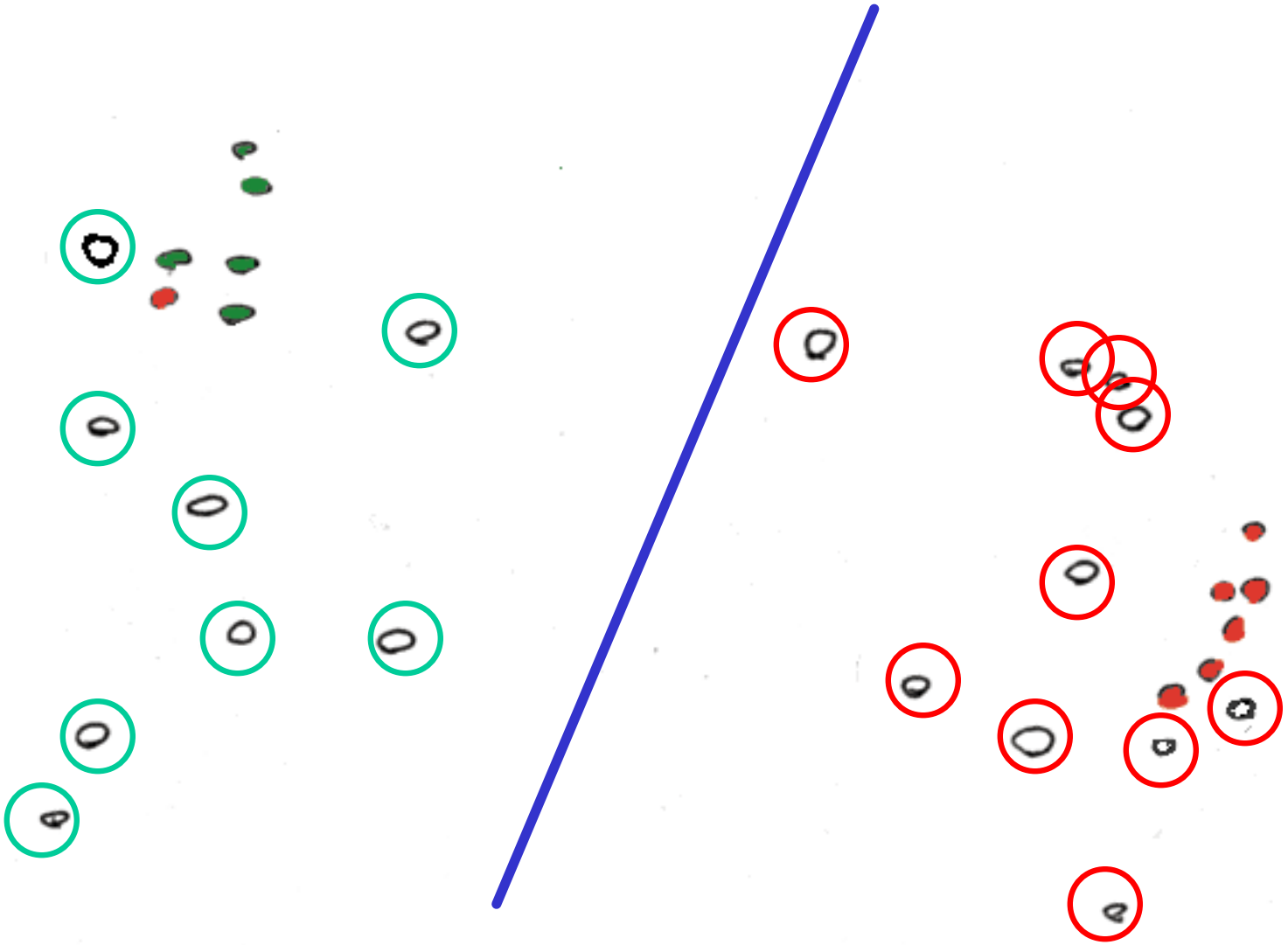
“Tag” Certain Points



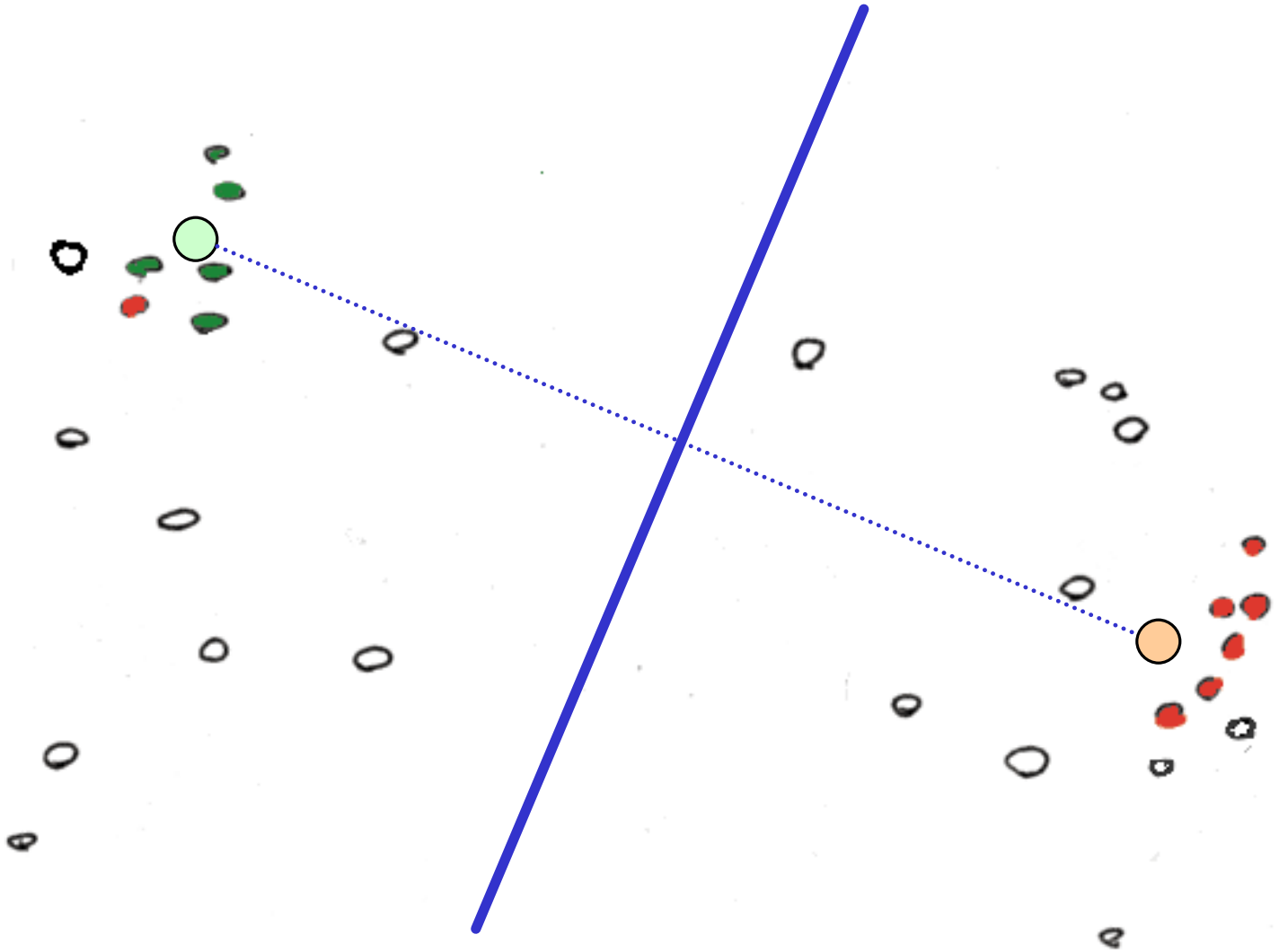
Find a Division to Separate Tagged Points



Extrapolate to Untagged Points



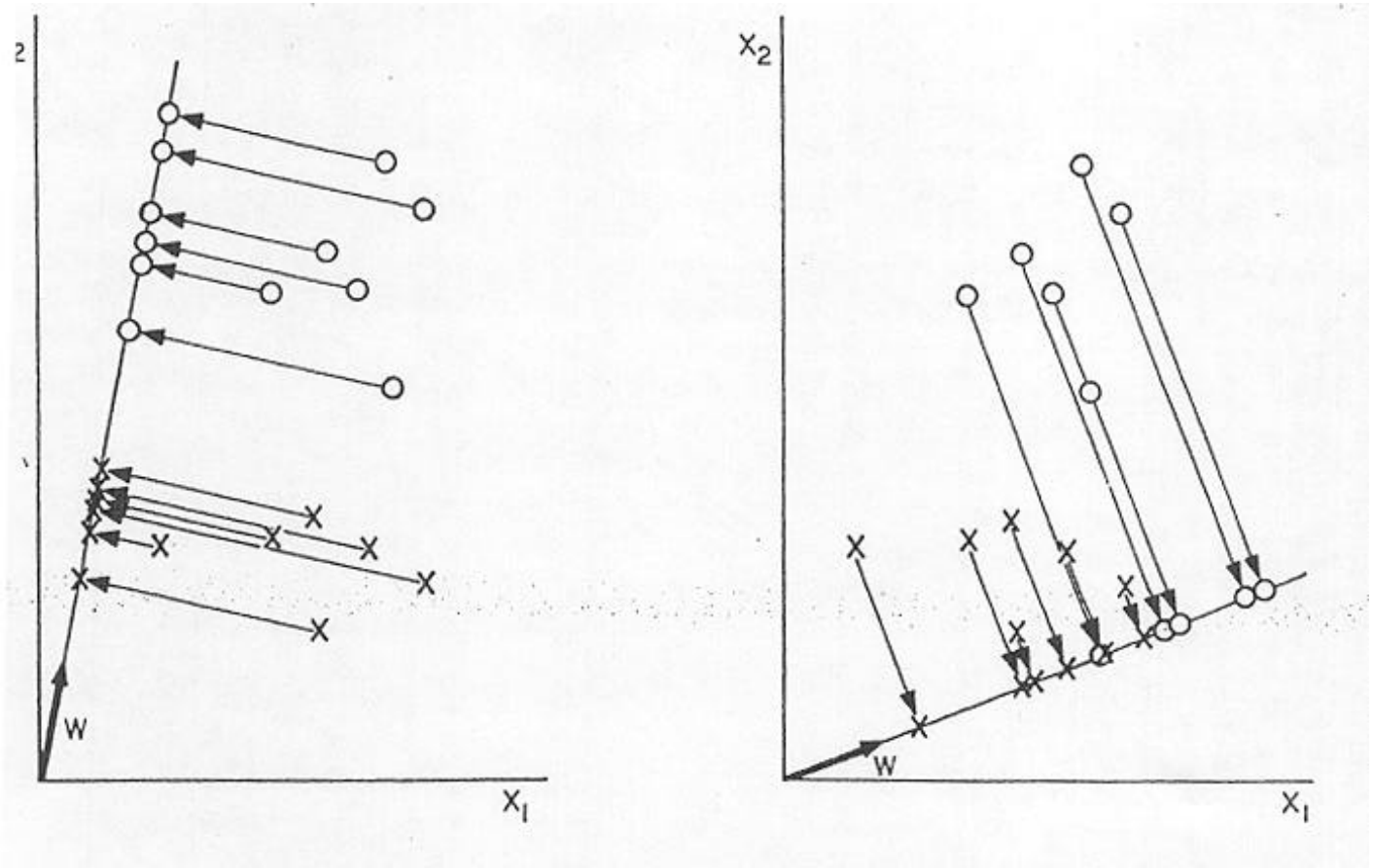
Discriminant to Position Plane



Fisher discriminant analysis

- Use the training set to reveal the structure of class distribution by seeking a linear combination
- $y = w_1x_1 + w_2x_2 + ... + w_nx_n$ which maximizes the ratio of the separation of the class means to the sum of each class variance (within class variance). This linear combination is called the first linear discriminant or first canonical variate. Classification of a future case is then determined by choosing the nearest class in the space of the first linear discriminant and significant subsequent discriminants, which maximally separate the class means and are constrained to be uncorrelated with previous ones.

Fischer's Discriminant



(Adapted from ???)

Fisher cont.

$$m_i = \vec{w} \cdot \vec{m}_i \qquad s_i^2 = \sum_{y \in Y_i} (y - m_i)^2$$

Solution of 1st variate

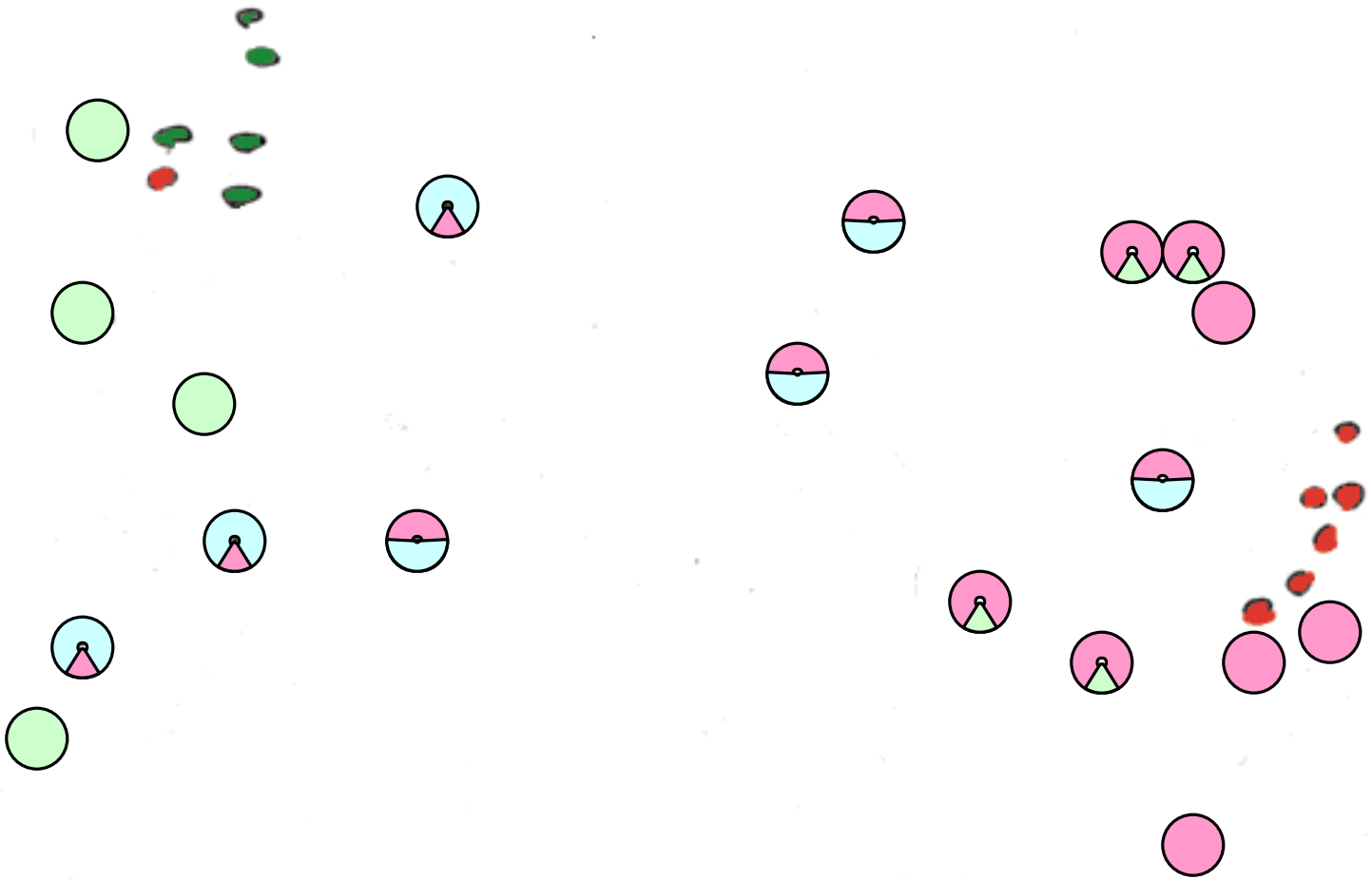
$$\vec{w} = S_W^{-1}(\vec{m}_1 - \vec{m}_2)$$

Bioinformatics

- A Very Broad Overview: What is Bioinformatics?
  - ◊ Types of Information, Organizing Principles, Informatics Techniques, Real-world Applications
- Example Calculation 1: Datamining Genome Information
  - ◊ Representing expression data and other features in high-dimensional space; Discriminants
  - ◊ Simple Bayesian analysis
- Example Calculation 2: Aligning Text Strings
  - ◊ Simple dynamic programming
  - ◊ Adding in gaps and other complexities

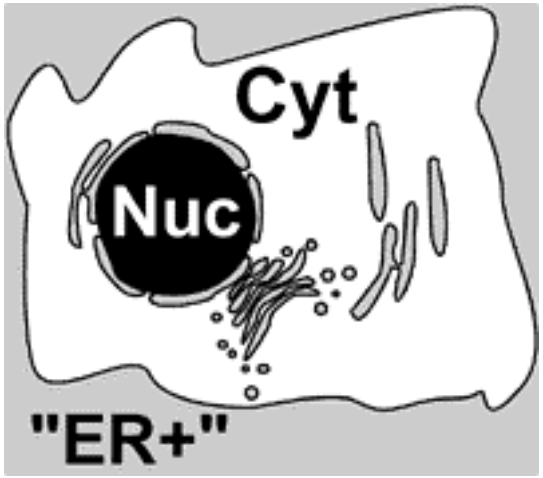
Tagged Data

Probabilistic Predictions of Class



Bayesian System for Localizing Proteins

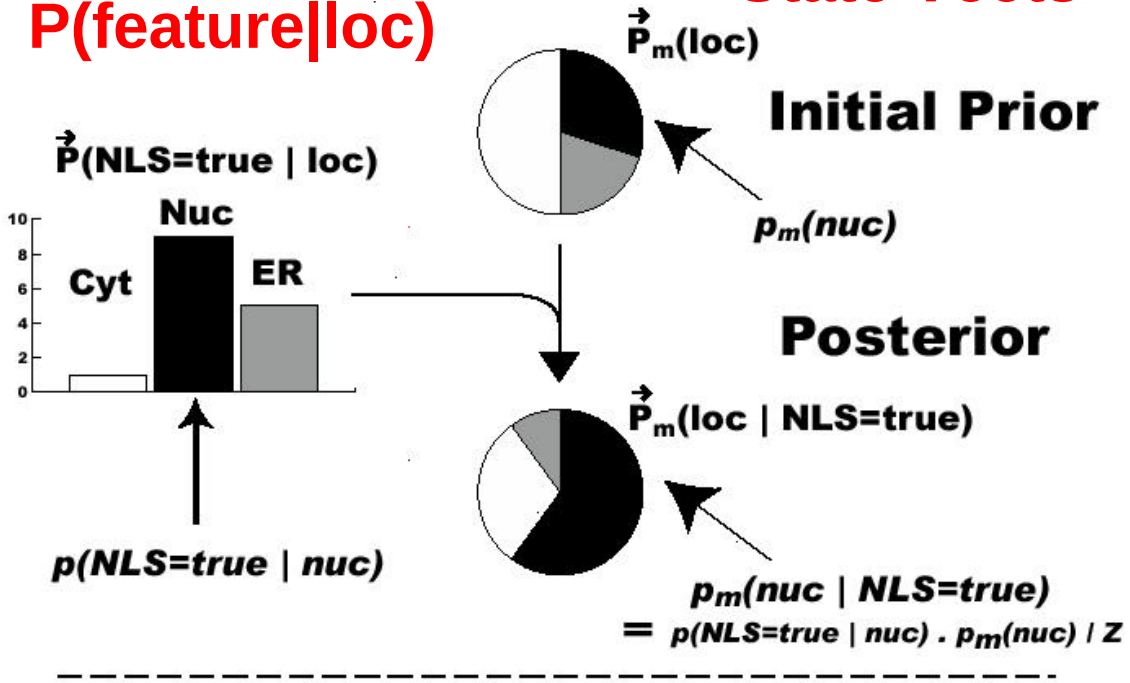
loc=



Represent localization of each protein by the state vector **P**(loc) and each feature by the feature vector P(feature|loc). Use Bayes rule to update.

Feature Vects
P(feature|loc)

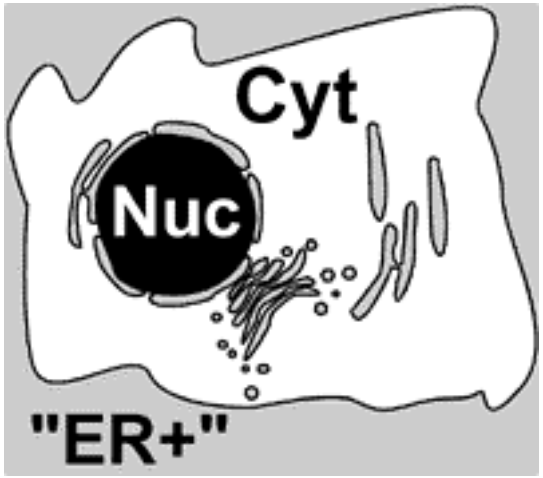
State Vects



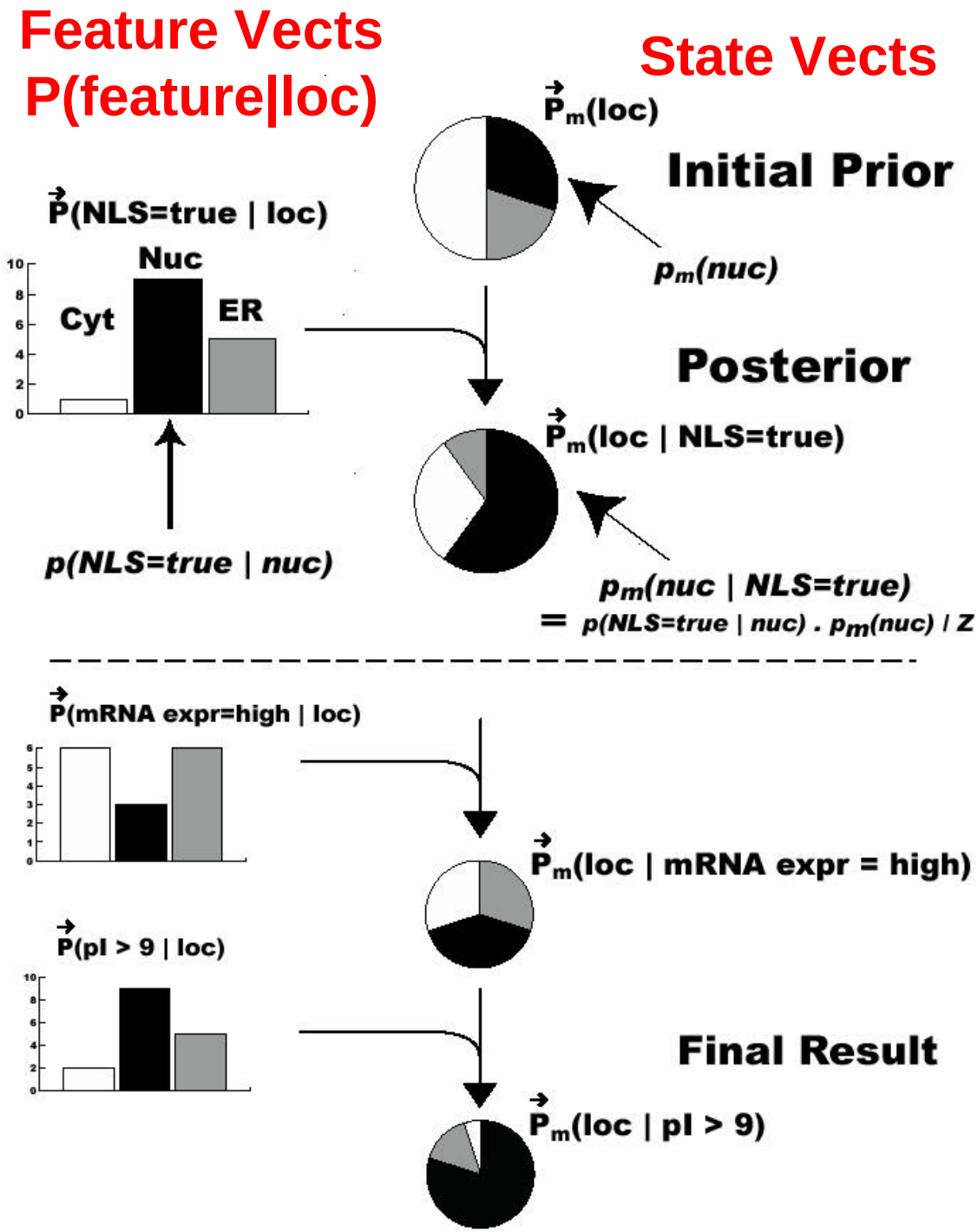
18 Features: Expression Level (absolute and fluctuations), signal seq., KDEL, NLS, Essential?, aa composition

Bayesian System for Localizing Proteins

loc=



Represent localization of each protein by the state vector **P**(loc) and each feature by the feature vector P(feature|loc). Use Bayes rule to update.



P(c|F) = P(F|c) P(c) / P(F)

P(c|F): Probability that protein is in class c given it has feature F

P(F|c): Probability in training data that a protein has feature F if it is class c

P(c): Prior probability that that protein is in class c

P(F): Normalization factor set so that sum over all classes c and ~c is 1 – i.e. P(c|F) + P(~c|F) = 1

Bayes Rule

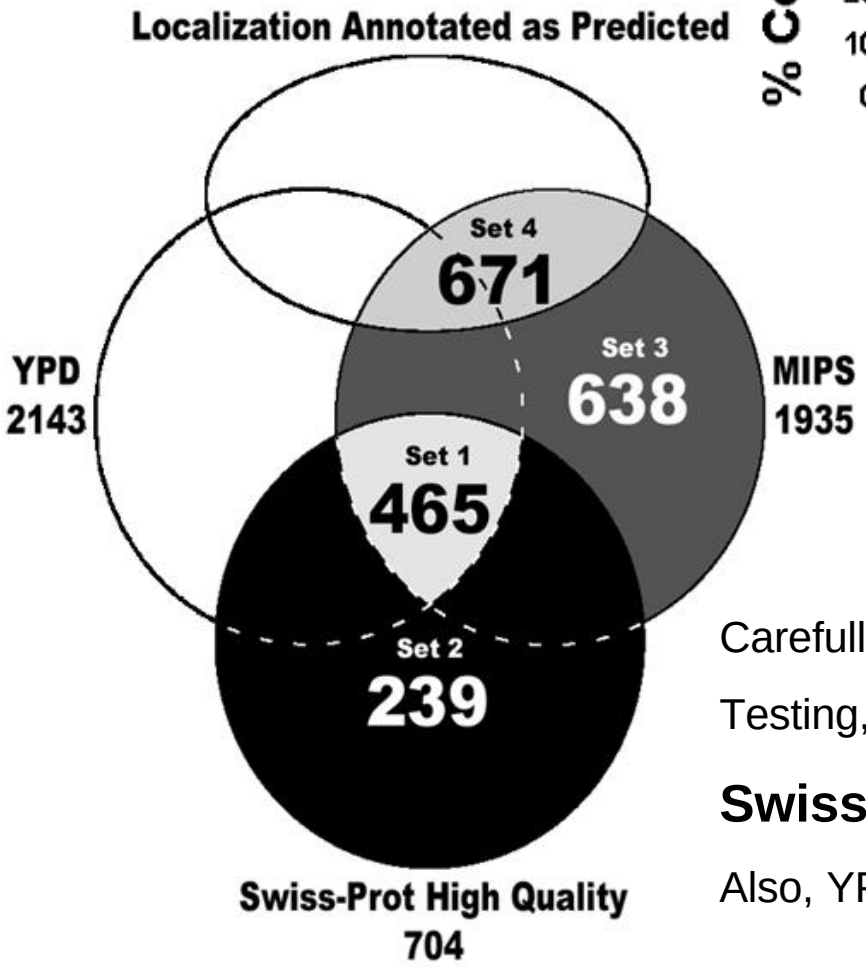
If features are independent, this formula can be iterated with
P(c) [at iter. i+1] <= P(c|F) [at iter. i]
("Naïve Bayesian Net")

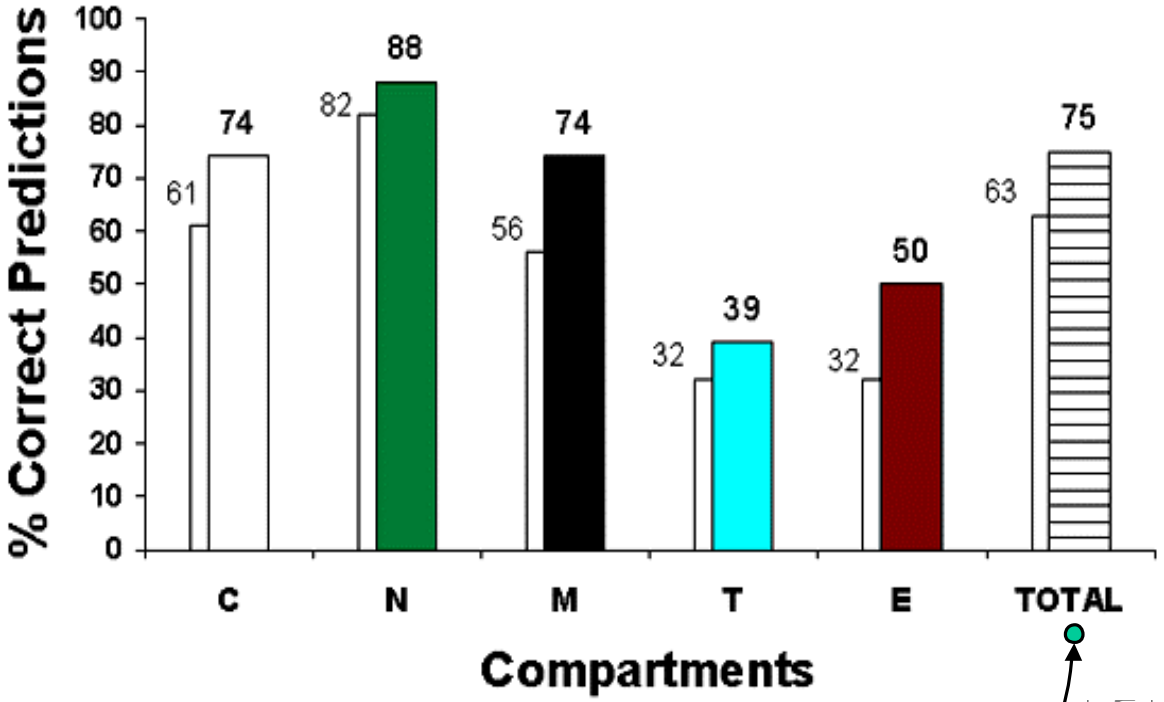
$$C_{MAP} = \arg\max_{C_j \in \{C_1, C_2\}} P(c_j) \prod_{i=1}^{n} P(x_i \mid c_j)$$

Yeast Tables for Localization Prediction

Basics		Predictors																				Response			Bayesian Localization								
		Sequence Features															Genomic Features																	
		seq. length	Amino Acid Composition								How many times does the sequence have these motif features?						Abs. expr. Level (mRNA copies / cell)		Cell cycle timecourse						Function	Localization	State Vector giving localization prediction					Collapsed Prediction	
Yeast Gene ID	Sequence		A	C	D			W	Y	farn site	NLS	hdel motif		nuc2	signalp	tms1	Gene-Chip expt. from RY Lab	sage tag freq.	t=0	t=1		t=15	t=16	function	5-compartment	C	N	M	T	E	Training	Extrapolation
YAL001C	M	1160	.08	.02	.06			.01	.04	0	1	0		1	0	0	0.3	0	5	3		4	5	04	TI	N	0%	100%	0%	0%	0%	N	
YAL002W	K	1176	.09	.02	.06			.01	.04	0	0	0		0	0	1	0.2	?	8	4		4	3	06	va	C	95%	3%	2%	0%	0%	C	
YAL003W	K	206	.08	.02	.06			.01	.04	0	0	0		0	0	0	19.1	19	70	73		98	126	05	tra	N	67%	33%	0%	0%	0%	C	
YAL004W	F	215	.08	.02	.06			.01	.04	0	0	0		0	0	0	?	0	18	12		4	6	01	0	N	41%	59%	0%	0%	0%	N	
YAL005C	V	641	.08	.02	.06			.01	.04	0	0	0		0	0	1	13.4	16	39	38		8	14	06	he	????	68%	32%	0%	0%	0%		C
YAL007C	K	190	.08	.02	.06			.01	.04	0	0	0		0	1	4	2.2	8	15	20		16	17	#	??	????	26%	43%	31%	0%	0%		-
YAL008W	H	198	.08	.02	.06			.01	.04	0	0	0		0	0	3	1.2	?	9	6		2	3	#	??	????	37%	60%	3%	0%	0%		-
YAL009W	F	259	.08	.02	.06			.01	.04	0	2	0		0	0	3	0.6	?	6	2		3	5	03	m	????	2%	98%	0%	0%	0%		N
YAL010C	M	493	.08	.02	.06			.02	.04	0	0	0		0	0	1	0.3	?	11	6		6	6	#	in	????	6%	90%	4%	0%	0%		N
YAL011W	K	616	.08	.02	.06			.01	.04	0	8	0		1	0	0	0.4	?	6	5		5	6	30	pr	????	28%	62%	10%	0%	0%		N
YAL012W	G	393	.08	.02	.06			.01	.04	0	0	0		0	0	1	8.9	4	29	26		23	29	01	cy	C	92%	5%	4%	0%	0%	C	
YAL013W	F	362	.08	.02	.06			.01	.04	0	0	0		0	0	0	0.6	?	7	9		6	10	01	re	N	0%	98%	0%	0%	1%	N	
YAL014C	G	202	.08	.02	.06			.01	.04	0	0	0		0	0	0	1.1	?	12	13		9	12	#	??	N	1%	96%	4%	0%	0%	N	
YAL015C	M	399	.08	.02	.06			.01	.04	0	1	0		0	0	0	0.7	0	19	18		12	13	11	D	N	4%	96%	0%	0%	0%	N	
YAL016W	K	635	.08	.02	.06			.01	.04	0	0	0		0	0	1	3.3	5	15	20		16	16	03	se	????	74%	26%	0%	0%	0%		C
YAL017W	V	1356	.08	.02	.06			.01	.04	0	0	0		0	0	0	0.4	?	14	3		4	7	#	??	????	0%	1%	99%	0%	0%		M
YAL018C	K	325	.08	.02	.06			.01	.04	0	0	0		0	0	4	?	?	4	2		2	1	#	??	????	0%	100%	0%	0%	0%		N

Results on Testing Data





Individual proteins: 75% with cross-validation

Carefully clean training dataset to **avoid circular logic**

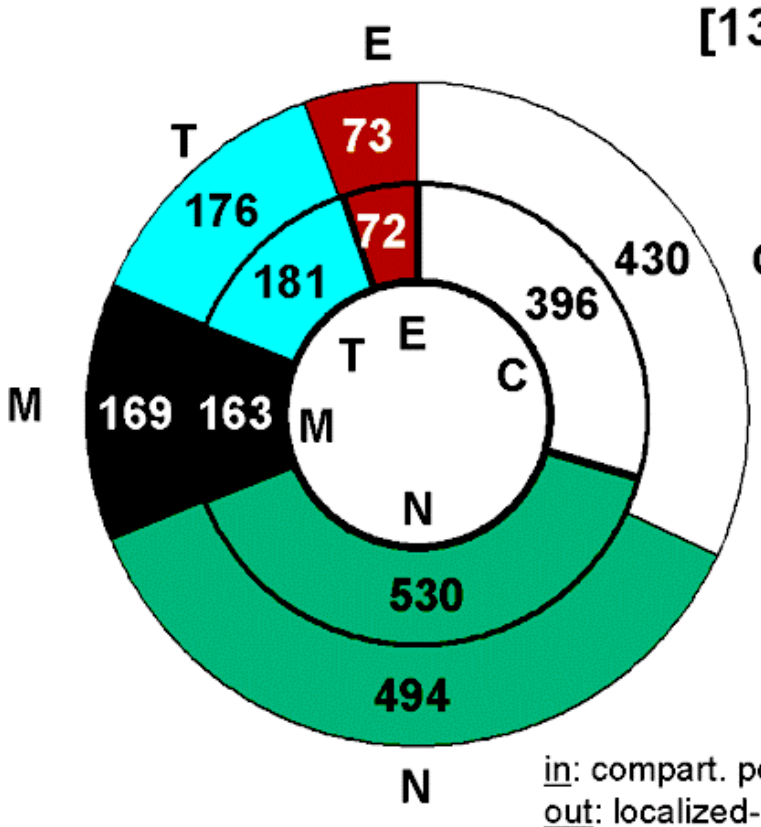
Testing, training data, Priors: ~2000 proteins from

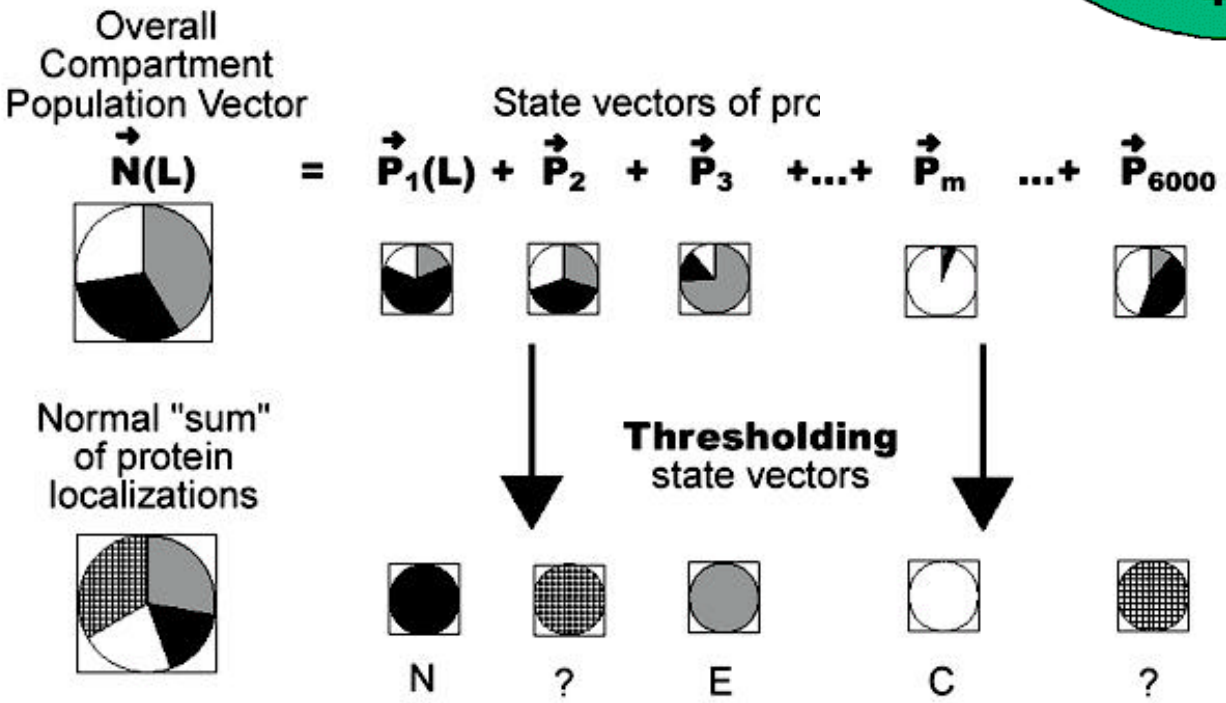
Swiss-Prot Master List

Also, YPD, MIPS, Snyder Lab

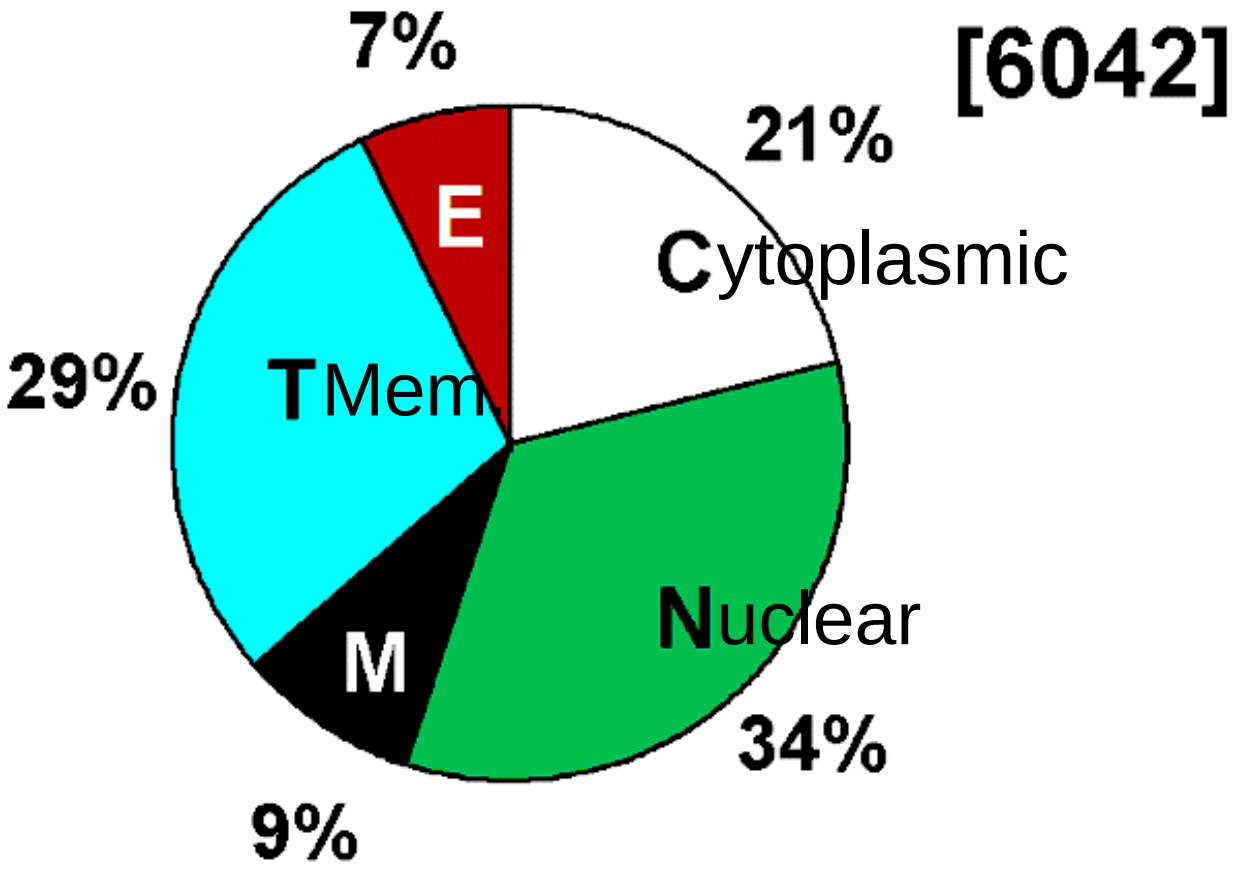
Results on Testing Data #2

Compartment Populations. Like QM, directly sum state vectors to get population. Gives **96%** pop. similarity.





Extrapolation to Compartment Populations of Whole Yeast Genome: ~4000 predicted + ~2000 known



Bioinformatics

- A Very Broad Overview: What is Bioinformatics?
  - ◊ Types of Information, Organizing Principles, Informatics Techniques, Real-world Applications
- Example Calculation 1: Datamining Genome Information
  - ◊ Representing expression data and other features in high-dimensional space; Discriminants
  - ◊ Simple Bayesian analysis
- Example Calculation 2: Aligning Text Strings
  - ◊ Simple dynamic programming
  - ◊ Adding in gaps and other complexities

Aligning Text Strings

```
Raw Data ???
  T  C  A  T  G
    C  A  T  T  G
```

```
2 matches, 0 gaps
  T  C  A  T  G
           |  |
  C  A  T  T  G
```

```
3 matches (2 end gaps)
  T  C  A  T  G  .
     |  |  |
  .  C  A  T  T  G
```

```
4 matches, 1 insertion
  T  C  A  -  T  G
     |  |     |  |
  .  C  A  T  T  G
```

```
4 matches, 1 insertion
  T  C  A  T  -  G
     |  |  |     |
  .  C  A  T  T  G
```

Dynamic Programming

- What to do for Bigger String?

```
SSDSEREEHVKRFRQALDDTGMKVPMATTNLFTHPVFKDGGFTANDRDVRRYALRKTIRNIDLAVELGAETYVAWGGREGAESGGAKDVRDALDRMKEAFDLLGEYVTSQGYDIRFAIEP
KPNEPRGDILLPTVGHALAFIERLERPELYGVNPEVGHEQMAGLNFPHGIAQALWAGKLFHIDLNGQNGIKYDQDLRFGAGDLRAAFWLVDLLESAGYSGPRHFDFKPPRTEDFDGVWAS
```

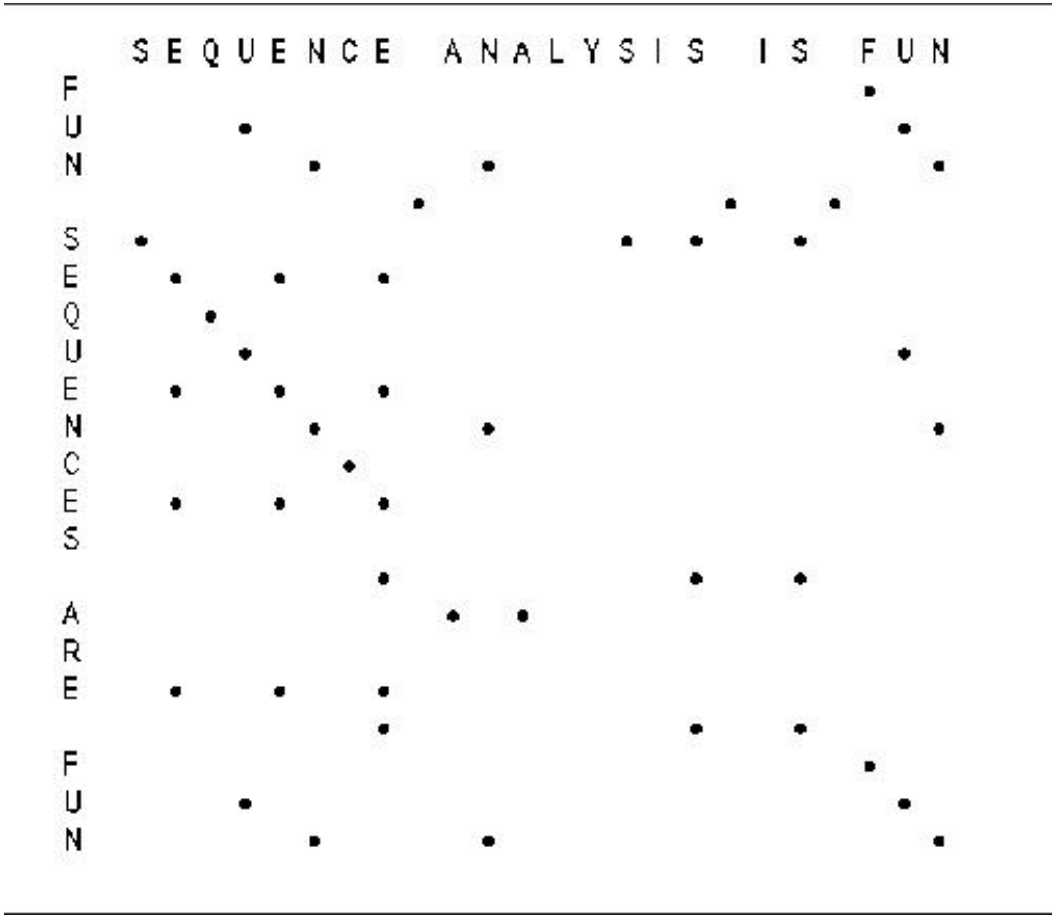
- Needleman-Wunsch (1970) provided first automatic method
  - ◊ Dynamic Programming to Find Global Alignment
- Their Test Data (J->Y)
  - ◊ ABCNYRQCLCRPM
    AYCYNRCKCRBP

Step 1 -- Make a Dot Plot (Similarity Matrix)

Put 1's where characters are identical.

	A	B	C	N	Y	R	Q	C	L	C	R	P	M
A	1												
Y					1								
C			1					1		1			
Y					1								
N				1									
R						1					1		
C			1					1		1			
K													
C			1					1		1			
R						1					1		
B		1											
P												1	

A More Interesting Dot Matrix



(adapted from R Altman)

Step 2 -- Start Computing the Sum Matrix

```
new_value_cell(R,C) <=
  cell(R,C)                          { Old value, either 1 or 0    }
  + Max[
        cell (R+1, C+1),             { Diagonally Down, no gaps    }
        cells(R+1, C+2 to C_max),{ Down a row, making col. gap }
        cells(R+2 to R_max, C+1) { Down a col., making row gap }
       ]
```

	A	B	C	N	Y	R	Q	C	L	C	R	P	M
A	**1**												
Y					**1**								
C			**1**					**1**		**1**			
Y					**1**								
N				**1**									
R						**1**					**1**		
C			**1**					**1**		**1**			
K													
C			**1**					**1**		**1**			
R						**1**					**1**		
B		**1**											
P												**1**	

	A	B	C	N	Y	R	Q	C	L	C	R	P	M
A	1												
Y					1								
C			1					1		1			
Y					1								
N				1									
R						1					**1**		
C			1					1		1			
K													
C			1					1		1			
R						1					***2***	**0**	**0**
B	**1**	**2**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**0**	**0**
P	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**1**	**0**

Step 3 -- Keep Going

	A	B	C	N	Y	R	Q	C	L	C	R	P	M
A	1												
Y					1								
C			1					1		1			
Y					1								
N				1									
R						1					**1**		
C			1					1		1			
K													
C			1					1		1			
R						1					***2***	**0**	**0**
B	**1**	**2**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**0**	**0**
P	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**1**	**0**

	A	B	C	N	Y	R	Q	C	L	C	R	P	M
A	1												
Y					1								
C			1					1		1			
Y					1								
N				1									
R						***5***	**4**	**3**	**3**	**2**	**2**	**0**	**0**
C	**3**	**3**	**4**	**3**	**3**	**3**	**3**	**4**	**3**	**3**	**1**	**0**	**0**
K	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**2**	**1**	**0**	**0**
C	**2**	**2**	**3**	**2**	**2**	**2**	**2**	**3**	**2**	**3**	**1**	**0**	**0**
R	**2**	**1**	**1**	**1**	**1**	**2**	**1**	**1**	**1**	**1**	**2**	**0**	**0**
B	**1**	**2**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**0**	**0**
P	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**1**	**0**

Step 4 -- Sum Matrix All Done

Alignment Score is 8 matches.

	A	B	C	N	Y	R	Q	C	L	C	R	P	M
A	1												
Y					1								
C			1					1		1			
Y					1								
N				1									
R						***5***	4	3	3	2	2	0	0
C	**3**	**3**	**4**	**3**	**3**	**3**	**3**	**4**	**3**	**3**	**1**	**0**	**0**
K	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**2**	**1**	**0**	**0**
C	**2**	**2**	**3**	**2**	**2**	**2**	**2**	**3**	**2**	**3**	**1**	**0**	**0**
R	**2**	**1**	**1**	**1**	**1**	**2**	**1**	**1**	**1**	**1**	**2**	**0**	**0**
B	**1**	**2**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**0**	**0**
P	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**1**	**0**

	A	B	C	N	Y	R	Q	C	L	C	R	P	M
A	***8***	**7**	**6**	**6**	**5**	**4**	**4**	**3**	**3**	**2**	**1**	**0**	**0**
Y	**7**	**7**	**6**	**6**	**6**	**4**	**4**	**3**	**3**	**2**	**1**	**0**	**0**
C	**6**	**6**	**7**	**6**	**5**	**4**	**4**	**4**	**3**	**3**	**1**	**0**	**0**
Y	**6**	**6**	**6**	**5**	**6**	**4**	**4**	**3**	**3**	**2**	**1**	**0**	**0**
N	**5**	**5**	**5**	**6**	**5**	**4**	**4**	**3**	**3**	**2**	**1**	**0**	**0**
R	**4**	**4**	**4**	**4**	**4**	**5**	**4**	**3**	**3**	**2**	**2**	**0**	**0**
C	**3**	**3**	**4**	**3**	**3**	**3**	**3**	**4**	**3**	**3**	**1**	**0**	**0**
K	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**2**	**1**	**0**	**0**
C	**2**	**2**	**3**	**2**	**2**	**2**	**2**	**3**	**2**	**3**	**1**	**0**	**0**
R	**2**	**1**	**1**	**1**	**1**	**2**	**1**	**1**	**1**	**1**	**2**	**0**	**0**
B	**1**	**2**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**0**	**0**
P	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**1**	**0**

Step 5 -- Traceback

Find Best Score (8) and Trace Back

A B C N Y - R Q C L C R - P M

A Y C - Y N R - C K C R B P

	A	B	C	N	Y	R	Q	C	L	C	R	P	M
A	**8**	7	6	6	5	4	4	3	3	2	1	0	0
Y	7	**7**	6	6	6	4	4	3	3	2	1	0	0
C	6	6	**7**	6	5	4	4	4	3	3	1	0	0
Y	6	6	6	5	**6**	4	4	3	3	2	1	0	0
N	5	5	5	6	5	4	4	3	3	2	1	0	0
R	4	4	4	4	4	**5**	4	3	3	2	2	0	0
C	3	3	4	3	3	3	3	**4**	3	3	1	0	0
K	3	3	3	3	3	3	3	3	**3**	2	1	0	0
C	2	2	3	2	2	2	2	3	2	**3**	1	0	0
R	2	1	1	1	1	2	1	1	1	1	**2**	0	0
B	1	2	1	1	1	1	1	1	1	1	1	0	0
P	0	0	0	0	0	0	0	0	0	0	0	**1**	0

Step 5 -- Traceback

A B C N Y - R Q C L C R - P M
A Y C - Y N R - C K C R B P

	A	B	C	N	Y	R	Q	C	L	C	R	P	M
A	**8**	7	6	6	5	4	4	3	3	2	1	0	0
Y	7	**7**	6	6	6	4	4	3	3	2	1	0	0
C	6	6	**7**	6	5	4	4	4	3	3	1	0	0
Y	6	6	6	5	**6**	4	4	3	3	2	1	0	0
N	5	5	5	6	5	4	4	3	3	2	1	0	0
R	4	4	4	4	4	**5**	4	3	3	2	2	0	0
C	3	3	4	3	3	3	3	**4**	3	3	1	0	0
K	3	3	3	3	3	3	3	3	**3**	2	1	0	0
C	2	2	3	2	2	2	2	3	2	**3**	1	0	0
R	2	1	1	1	1	2	1	1	1	1	**2**	0	0
B	1	2	1	1	1	1	1	1	1	1	1	0	0
P	0	0	0	0	0	0	0	0	0	0	0	**1**	0

Bioinformatics

- A Very Broad Overview: What is Bioinformatics?
  - ◊ Types of Information, Organizing Principles, Informatics Techniques, Real-world Applications
- Example Calculation 1: Datamining Genome Information
  - ◊ Representing expression data and other features in high-dimensional space; Discriminants
  - ◊ Simple Bayesian analysis
- Example Calculation 2: Aligning Text Strings
  - ◊ Simple dynamic programming
  - ◊ Adding in gaps and other complexities

Step 6 -- Alternate Tracebacks

A B C - N Y R Q C L C R - P M
A Y C Y N - R - C K C R B P

Also,
Suboptimal
Aligments

	A	B	C	N	Y	R	Q	C	L	C	R	P	M
A	**8**	7	6	6	5	4	4	3	3	2	1	0	0
Y	7	**7**	6	6	6	4	4	3	3	2	1	0	0
C	6	6	**7**	6	5	4	4	4	3	3	1	0	0
Y	6	6	6	5	***6***	4	4	3	3	2	1	0	0
N	5	5	5	**6**	5	4	4	3	3	2	1	0	0
R	4	4	4	4	4	**5**	4	3	3	2	2	0	0
C	3	3	4	3	3	3	3	**4**	3	3	1	0	0
K	3	3	3	3	3	3	3	3	**3**	2	1	0	0
C	2	2	3	2	2	2	2	3	2	**3**	1	0	0
R	2	1	1	1	1	2	1	1	1	1	**2**	0	0
B	1	2	1	1	1	1	1	1	1	1	1	0	0
P	0	0	0	0	0	0	0	0	0	0	0	**1**	0

Suboptimal Alignments

```
;
; Random DNA sequence generated using the seed :   -453862491
;
;   500 nucleotides
;
; A:C:G:T =   1 :   1 :   1 :   1
;
RAN  -453862491
AAATGCCAAA TCATACGAAC AGCCGACGAC GGGAGCAACC CAAGTCGCAG TTCGCTTGAG CTAGCGCGCT
CCCACCGGGA TATACACTAA TCATTACAGC AGGTCTCCTG GGCGTACAGA CTAGCTGAAC GCGCTGCGCC
AATTCCAACT TCGGTATGAA GGATCGCCTG CGGTTATCGC TGACTTGAGT AACCAGATCG CTAAGGTTAC
GCTGGGGCAA TGATGGATGT TAACCCCTTA CAGTCTCGGG AGGGACCTTA AGTCGTAATA GATGGCAGCA
TTAATACCTT CGCCGTTAAT ATACCTTTAA TCCGTTCTTG TCAATGCCGT AGCTGCAGTG AGCCTTCTGT
CACGGGCATA CCGCGGGGTA GCTGCAGCAA CCGTAGGCTG AGCATCAAGA AGACAAACAC TCCTCGCCTA
CCCCGGACAT CATATGACCA GGCAGTCTAG GCGCCGTTAG AGTAAGGAGA CCGGGGGGCC GTGATGATAG
ATGGCGTGTT 1
;
; Random DNA sequence generated using the seed :   1573438385
;
;   500 nucleotides
;
; A:C:G:T =   1 :   1 :   1 :   1
;
RAN  1573438385
CCCTCCATCG CCAGTTCCTG AAGACATCTC CGTGACGTGA ACTCTCTCCA GGCATATTAA TCGAAGATCC
CCTGTCGTGA CGCGGATTAC GAGGGGATGG TGCTAATCAC ATTGCGAACA TGTTTCGGTC CAGACTCCAC
CTATGGCATC TTCCGCTATA GGGCACGTAA CTTTCTTCGT GTGGCGGCGC GGCAACTAAA GACGAAAGGA
CCACAACGTG AATAGCCCGT GTCGTGAGGT AAGGGTCCCG GTGCAAGAGT AGAGGAAGTA CGGGAGTACG
TACGGGGCAT GACGCGGGCT GGAATTTCAC ATCGCAGAAC TTATAGGCAG CCGTGTGCCT GAGGCCGCTA
GAACCTTCAA CGCTAACTAG TGATAACTAC CGTGTGAAAG ACCTGGCCCG TTTTGTCCCT GAGACTAATC
GCTAGTTAGG CCCCATTTGT AGCACTCTGG CGCAGACCTC GCAGAGGGAC CGGCCTGACT TTTTCCGGCT
TCCTCTGAGG 1

Parameters: match weight = 10, transition weight = 1, transversion weight = -3
Gap opening penalty = 50   Gap continuation penalty = 1
Run as a local alignment (Smith-Waterman)
```

(courtesy of Michael Zucker)

Gap Penalties

The score at a position can also factor in a penalty for introducing gaps (i. e., not going from i, j to i- 1, j- 1).

Gap penalties are often of linear form:

$$GAP = a + bN$$

GAP is the gap penalty

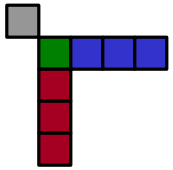
a = cost of opening a gap

b = cost of extending the gap by one (affine)

N = length of the gap

(Here assume b=0, a=1/2, so GAP = 1/2 regardless of length.)

Step 2 -- Computing the Sum Matrix **with Gaps**



```
new_value_cell(R,C) <=
  cell(R,C)                                  { Old value, either 1 or 0    }
  + Max[
        cell (R+1, C+1),                     { Diagonally Down, no gaps    }
        cells(R+1, C+2 to C_max) - GAP ,{ Down a row, making col. gap }
        cells(R+2 to R_max, C+1) - GAP  { Down a col., making row gap }
       ]
```

	A	B	C	N	Y	R	Q	C	L	C	R	P	M
A	**1**												
Y					**1**								
C			**1**					**1**		**1**			
Y					**1**								
N				**1**									
R						**1**					**1**		
C			**1**					**1**		**1**			
K													
C			**1**					**1**		**1**			
R						**1**					**1**		
B		**1**											
P												**1**	

	A	B	C	N	Y	R	Q	C	L	C	R	P	M
A	1												
Y					1								
C			1					1		1			
Y					1								
N				1									
R						1					1		
C			1					1		1			
K													
C			1					1		1			
R						1					*1.5*	**0**	**0**
B	**1**	**2**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**0**	**0**
P	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**1**	**0**

GAP =1/2

All Steps in Aligning a 4-mer

C R B P

C R P M
- C R P M
C R - P M

	C	R	P	M
C	**1**			
R		**1**		
B				
P			**1**	

	C	R	P	M
C	1			
R		***2***	**0**	**0**
B	**1**	**1**	**0**	**0**
P	**0**	**0**	**1**	**0**

	C	R	P	M
C	**3**	**1**	**0**	**0**
R	**1**	**2**	**0**	**0**
B	**1**	**1**	**0**	**0**
P	**0**	**0**	**1**	**0**

	C	R	P	M
C	**3**	1	0	0
R	1	**2**	0	0
B	1	**1**	0	0
P	0	0	**1**	0

Bottom right hand corner of previous matrices

Key Idea in Dynamic Programming

◊ The best alignment that ends at a given pair of positions (i and j) in the 2 sequences is the score of the best alignment previous to this position PLUS the score for aligning those two positions.

◊ An Example Below

- Aligning R to K does not affect alignment of previous N-terminal residues. Once this is done it is **fixed**. Then go on to align D to E.
- How could this be violated?
  Aligning R to K changes best alignment in box.

```
ACSQRP--LRV-SH  RSENCV
A-SNKPQLVKLMTH  VKDFCV

ACSQRP--LRV-SH  -R  SENCV
A-SNKPQLVKLMTH  VK  DFCV
```

Bioinformatics

- A Very Broad Overview: What is Bioinformatics?
  - ◊ Types of Information, Organizing Principles, Informatics Techniques, Real-world Applications
- Example Calculation 1: Datamining Genome Information
  - ◊ Representing expression data and other features in high-dimensional space; Discriminants
  - ◊ Simple Bayesian analysis
- Example Calculation 2: Aligning Text Strings
  - ◊ Simple dynamic programming
  - ◊ Adding in gaps and other complexities

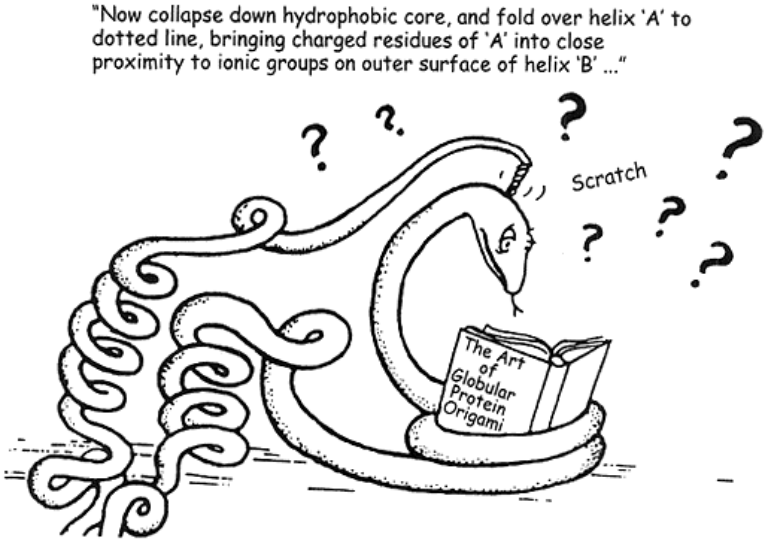
Bioinformatics Topics -- Genome Sequence

- Finding Genes in Genomic DNA
  - ◊ introns
  - ◊ exons
  - ◊ promotors
- Characterizing Repeats in Genomic DNA
  - ◊ Statistics
  - ◊ Patterns
- Duplications in the Genome

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

Bioinformatics Topics -- Sequence / Structure



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

- Secondary Structure "Prediction"
  - ◊ via Propensities
  - ◊ Neural Networks, Genetic Alg.
  - ◊ Simple Statistics
  - ◊ TM-helix finding
  - ◊ Assessing Secondary Structure Prediction
- Tertiary Structure Prediction
  - ◊ Fold Recognition
  - ◊ Threading
  - ◊ Ab initio
- Function Prediction
  - ◊ Active site identification
- Relation of Sequence Similarity to Structural Similarity

Topics -- Structures

- Basic Protein Geometry and Least-Squares Fitting
  - ◊ Distances, Angles, Axes, Rotations
    - Calculating a helix axis in 3D via fitting a line
  - ◊ LSQ fit of 2 structures
  - ◊ Molecular Graphics
- Calculation of Volume and Surface
  - ◊ How to represent a plane
  - ◊ How to represent a solid
  - ◊ How to calculate an area
  - ◊ Docking and Drug Design as Surface Matching
  - ◊ Packing Measurement
- Structural Alignment
  - ◊ Aligning sequences on the basis of 3D structure.
  - ◊ DP does not converge, unlike sequences, what to do?
  - ◊ Other Approaches: Distance Matrices, Hashing
  - ◊ Fold Library

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)
  - ◊ Forms and Reports
  - ◊ Cross-tabulation
- Protein Units?
  - ◊ What are the units of biological information?
    - sequence, structure
    - motifs, modules, domains
  - ◊ How classified: folds, motions, pathways, functions?
- Clustering and Trees
  - ◊ Basic clustering
    - UPGMA
    - single-linkage
    - multiple linkage
  - ◊ Other Methods
    - Parsimony, Maximum likelihood
  - ◊ Evolutionary implications
- The Bias Problem
  - ◊ sequence weighting
  - ◊ sampling

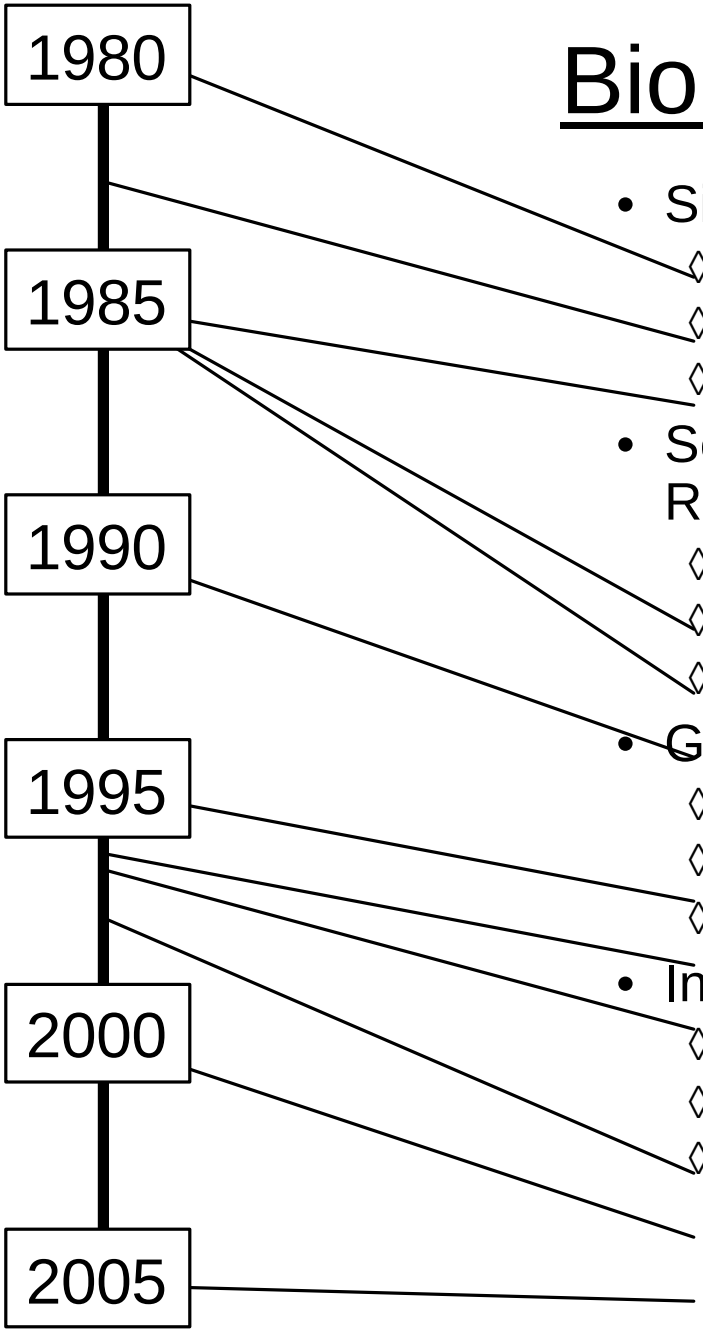
Topics -- Genomics

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

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

Topics -- Simulation

- Molecular Simulation
  - ◊ Geometry -> Energy -> Forces
  - ◊ Basic interactions, potential energy functions
  - ◊ Electrostatics
  - ◊ VDW Forces
  - ◊ Bonds as Springs
  - ◊ How structure changes over time?
    - How to measure the change in a vector (gradient)
  - ◊ Molecular Dynamics & MC
  - ◊ Energy Minimization
- Parameter Sets
- Number Density
- Poisson-Boltzman Equation
- Lattice Models and Simplification



Bioinformatics - History

- Single Structures
  - ◊ Modeling & Geometry
  - ◊ Forces & Simulation
  - ◊ Docking
- Sequences, Sequence-Structure Relationships
  - ◊ Alignment
  - ◊ Structure Prediction
  - ◊ Fold recognition
- Genomics
  - ◊ Dealing with many sequences
  - ◊ Gene finding & Genome Annotation
  - ◊ Databases
- Integrative Analysis
  - ◊ Expression & Proteomics Data
  - ◊ Datamining
  - ◊ Simulation again….

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
  - ◊ **(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)