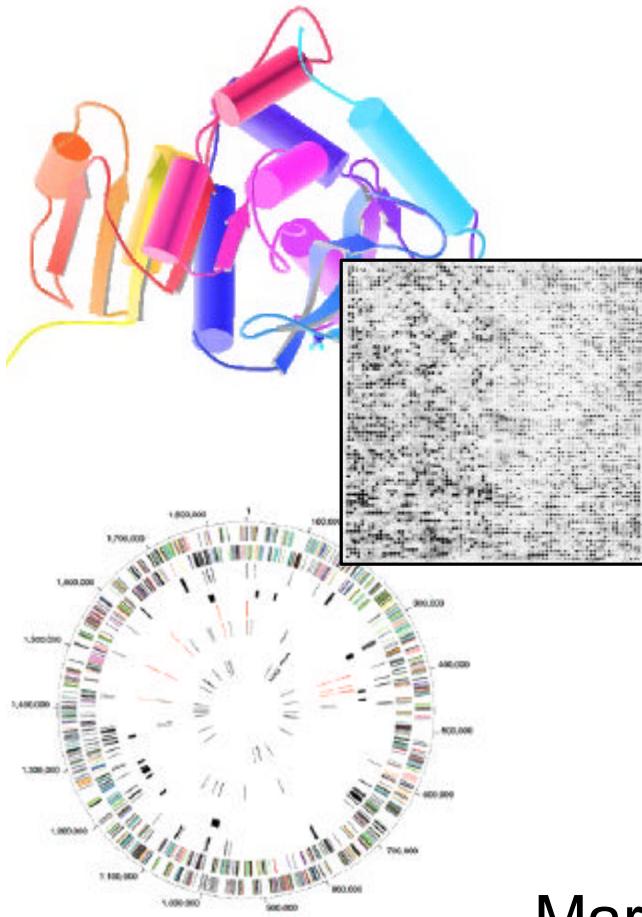


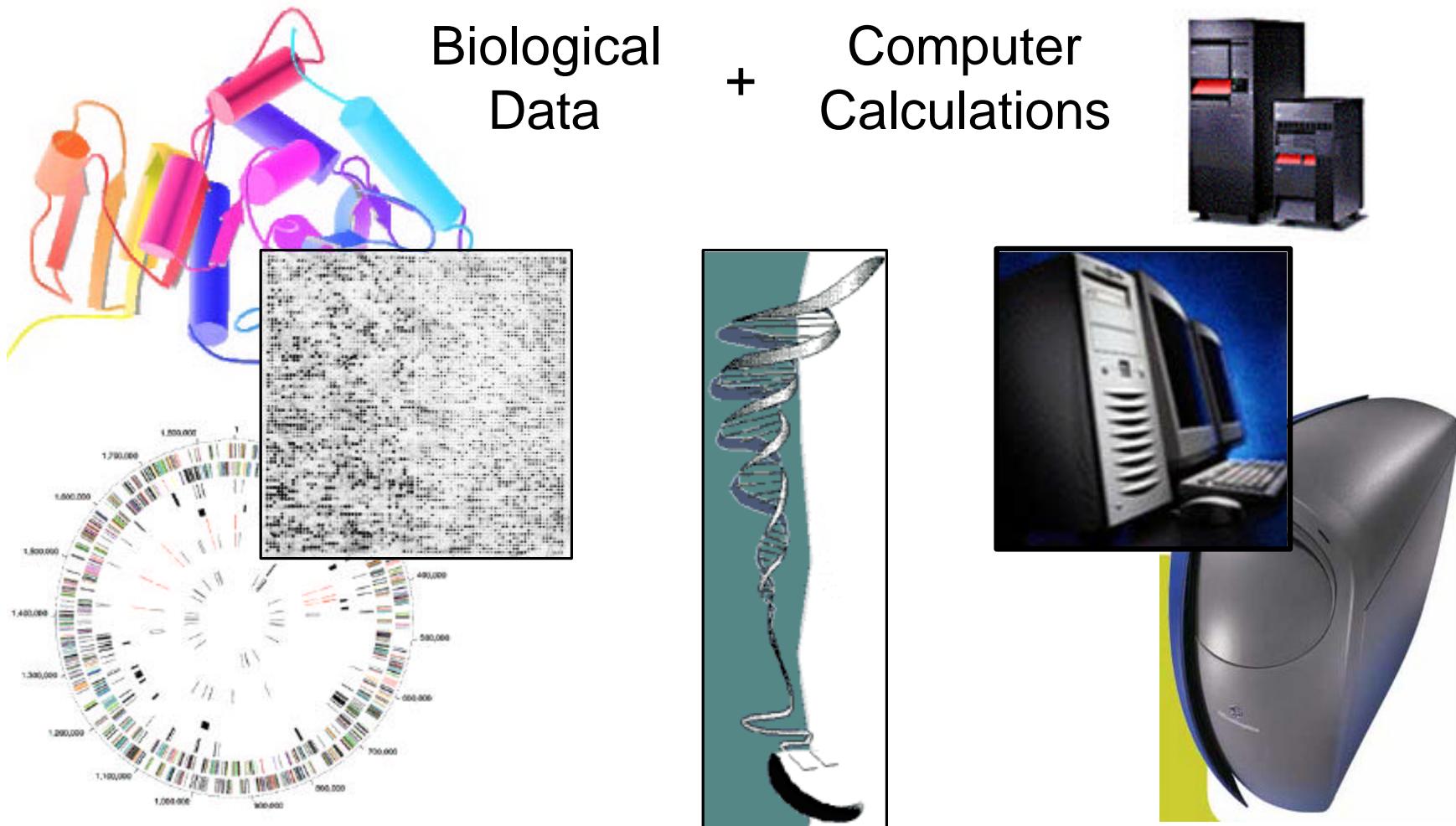
BIOINFORMATICS

Introduction



Mark Gerstein, Yale University
bioinfo.mbb.yale.edu/mgb452a

Bioinformatics



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**.

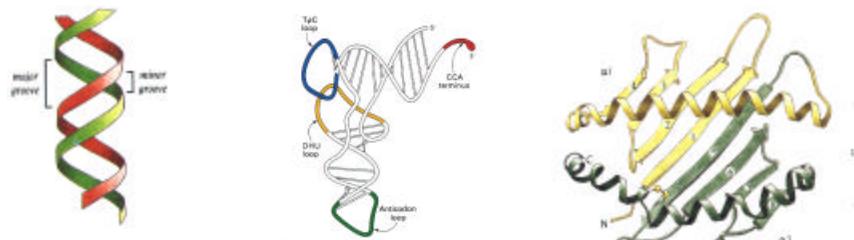
What is the **Information**?

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

atggcaattaaaatttgttatcaatgggttggctgtatcgccgtatcgattccgtgca
gcacaacaccgtgatgacattgaagtttaggtattaacgacttaatcgacgttgaaatac
atggcttatatgttggaaatatgattcaactcacggtcgttgcacggcactgttgaaagt
aaagatgttaacttagtggtaatggtaaaactatccgttaactgcagaacgtgatcca
gcaaaacttaaactggggtgcaatcggttggatatcgcttgcagcgactggtttattc
ttaactgatgaaactgctcgtaaacatatactgcaggcgaaaaaaaaagttgttattact
ggcccatctaaagatgcacccctatgttcgttgcggtaaacttcaacgcatacgca
ggtcaagatatcgttctaacgcatactgttacaacaaactgtttagtcctttagcagct
gttgcgttatggatggatggccatcagctaaagactggcgccggccgggtgca
tcacaaaacatcattccatcttcaacaggtgcagcgaaaagcagtaggttaagtattact
gcattaaacggtaattaaactggatggcttccgtgttccaaacgcacaaacgtatctgtt
gttgcgttatggatggatggccatcagctaaagactggcgccggccgggtgca
aaagatgcagcgaaaggtaaaacggttcaatggcgaattaaaaggcgtttaggttacact
gaagatgcgttgcgttactgacttcaacgcgttgcttaacttctgtatttgatgca
gacgcgttgatcgattttcgtaattggatc . . .

..... caaaaatagggttaatatgaatctcgatctccatttgttcatcgattcaa
caacaagccaaaactcgtacaaatatgaccgcacttcgtataaaagaacacggcttggtt
cgagatatactcttggaaaaacttcaagagcaactcaatcaactttctcgagcattgtt
gctcacaatattgacgtacaaagataaaatgccattttgc当地atatgaaacgttgg
gttcatgaaacttcgttatcaaagatggtaatgaccactgttacgc当地acgact
acaatcgttgcacattgc当地accttacaaattcgc当地acatcacagtgccattacgc当地acc
aatacagccc当地cagcaagc当地agaatttatcctaattacgc当地ccatgttaaaaaattctt当地cg
ggc当地gatcaagagcaatacgc当地atcaacattggaaattgctcatcattgtccaaaattacaa
aaaattt当地qcaatqaaatccaccattcaattacaacaagatctctt当地tgc当地acttq

Molecular Biology Information: Protein Sequence

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

d1dhfa_ LNCIVAVSQNMIGKNGDLPWPPLRNEFRYFQRMTTSSVEGKQ-NLVIMGKKTWFSI
d8dfr_ LNSIVAVCQNMIGKDGNLWPPLRNEYKYFQRMSTSHTVEGKQ-NAVIMGKKTWFSI
d4dfra_ ISLIAALAVDRVIGMENAMPWN-LPADLAWFKRNTL-----NKPVIMGRHTWESI
d3dfr_ TAFLWAQDRDGLIGKDGHLPWH-LPDDLHYFRAQTV-----GKIMVVGRRTYESF

d1dhfa_ LNCIVAVSQNMIGKNGDLPWPPLRNEFRYFQRMTTSSVEGKQ-NLVIMGKKTWFSI
d8dfr_ LNSIVAVCQNMIGKDGNLWPPLRNEYKYFQRMSTSHTVEGKQ-NAVIMGKKTWFSI
d4dfra_ ISLIAALAVDRVIGMENAMPW-NLPADLAWFKRNTLD-----KPVIMGRHTWESI
d3dfr_ TAFLWAQDRNGLIGKDGHLPW-HLPDDLHYFRAQTVG-----KIMVVGRRTYESF

d1dhfa_ VPEKNRPLKGRINLVLSRELKEPPQGAHFLSRSLDDALKTEQPELANKVDMWIVGGSSVYKEAMNHP
d8dfr_ VPEKNRPLKDRINIVLSRELKEAPKGAIYLSKSLLDALALLD SPELKSVDWIVGGTAVYKAAMEKP
d4dfra_ ---G-RPLPGRKNIILS-SQPGTDDRV-TWVKSVDEAIAACGDVP-----EIMVIGGGRVYEQFLPKA
d3dfr_ ---PKRPLPERTNVVLTHQEDYQAQGA-VVHDVAAVFAYAKQHLDQ---ELVIAGGAQIFTAFKDDV

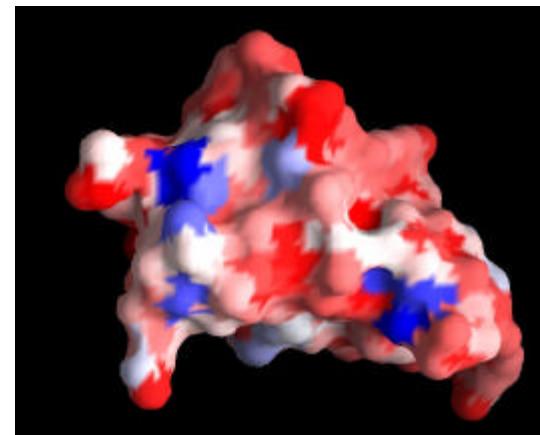
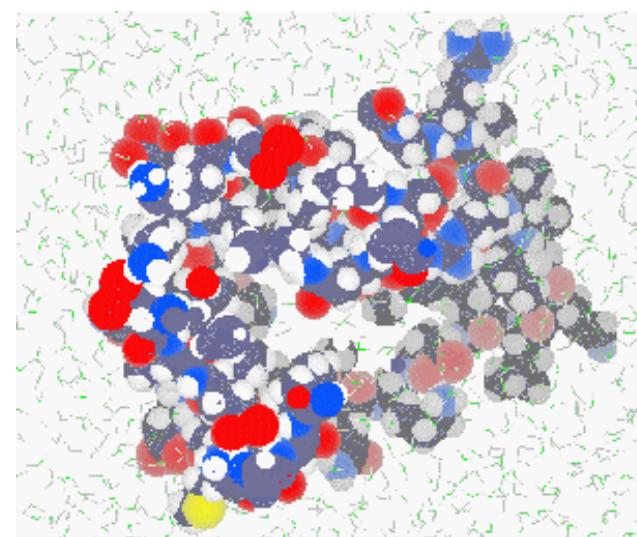
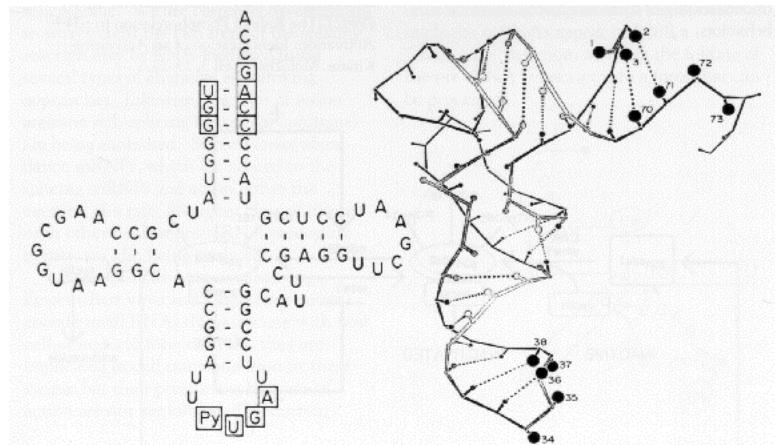
d1dhfa_ -PEKNRPLKGRINLVLSRELKEPPQGAHFLSRSLDDALKTEQPELANKVDMWIVGGSSVYKEAMNHP
d8dfr_ -PEKNRPLKDRINIVLSRELKEAPKGAIYLSKSLLDALALLD SPELKSVDWIVGGTAVYKAAMEKP
d4dfra_ -G---RPLPGRKNIILSSSQPGTDDRV-TWVKSVDEAIAACGDVPE-----IMVIGGGRVYEQFLPKA
d3dfr_ -P--KRPLPERTNVVLTHQEDYQAQGA-VVHDVAAVFAYAKQHLD---QELVIAGGAQIFTAFKDDV

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)

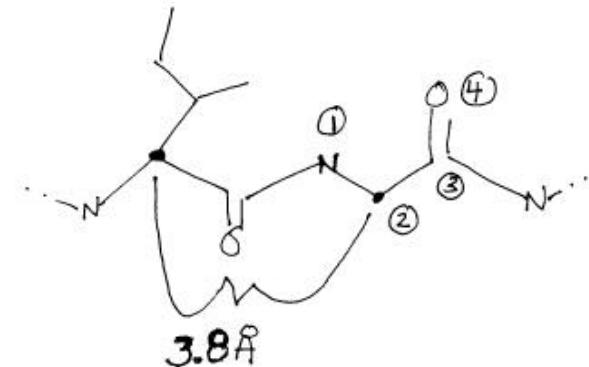


Molecular Biology Information:

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
 - ◊ 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., Geoghegan, N. S. M., Gnehm, C. L., McDonald, L. A.,

Small, K. V., Fraser, C. M., Smith, H. O. & Venter, J. C. (1995). "Whole-

genome random sequencing and assembly of *Haemophilus influenzae* rd." Science 269: 496-512.

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

- Integrative Data

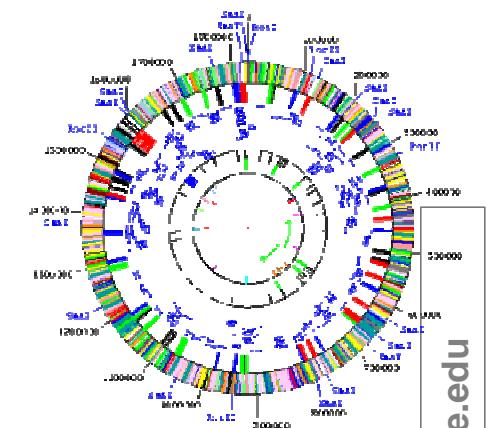
1995, HI (bacteria): 1.6 Mb & 1600 genes done

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

1998, worm: ~100Mb with 19 K genes

1999: >30 completed genomes!

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



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

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

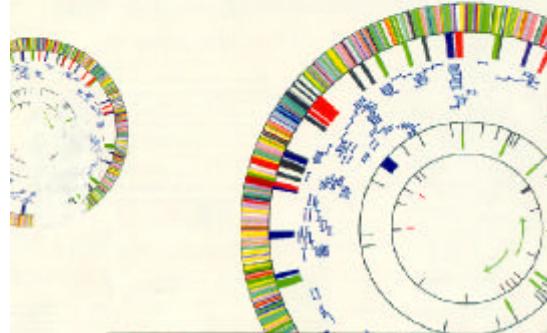
1995

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



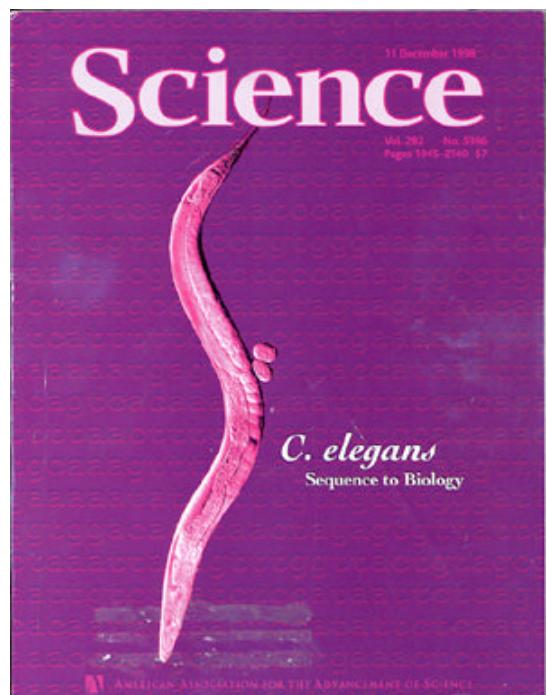
1997

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



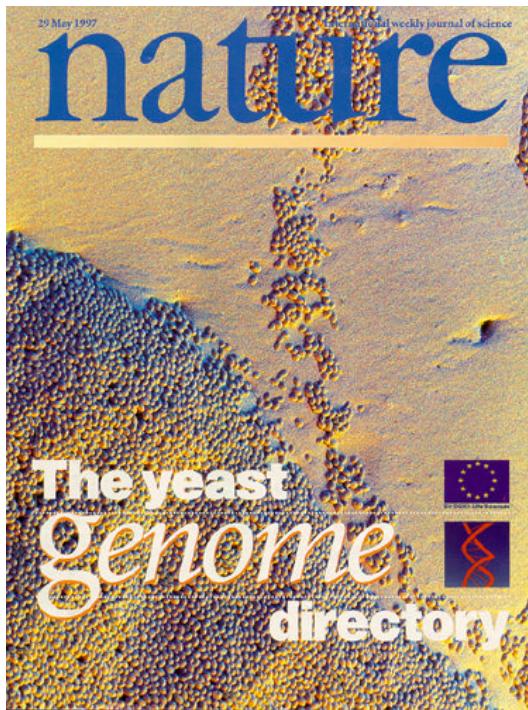
1998

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



2000?

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



Genomes
highlight
the
Finiteness
of the
"Parts" in
Biology

real thing, Apr '00



'98 spoof

Dissecting the Regulatory Circuitry of a Eukaryotic Genome

Frank C. P. Holstege,* Ezra G. Jennings,*¹
John J. Wyrick,¹ Tong Ihn Lee,¹
Christoph J. Hengartner,¹ Michael R. Green,¹
Todd R. Golub,^{1,6} Eric S. Lander,^{1,2}
and Richard A. Young^{1,11}

¹Whitehead Institute for Biomedical Research
Cambridge, Massachusetts 02142

²Department of Biology
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

³Howard Hughes Medical Institute
Program in Molecular Medicine

University of Massachusetts Medical Center
Worcester, Massachusetts 01655

⁴Dana-Farber Cancer Institute and
Harvard Medical School

Boston, Massachusetts 02115



Young/Lander, Chips, Abs. Exp.

specific transcript factors, or forces transcriptional silencing of a specific set of genes.

Figure 2. Genomewide Expression Data for Selected Components of the RNA Polymerase II Transcription Complex. The figure shows four panels (A–D) of gene expression data. Each panel contains a grid where each row represents a different gene and each column represents a different condition or sample. The colors in the grid indicate the level of mRNA expression, with red being high and blue being low. Panel A shows RP81, Panel B shows MED6, Panel C shows SRB10, and Panel D shows SWI2. The panels illustrate how specific transcription factors (RP81, MED6, SRB10, SWI2) are expressed across the genome.

The Brown Lab
Stanford University Department of Biochemistry

The MGuide

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

The transcription program in the response of
human fibroblasts to serum

The results support the hypothesis that

The Transcription of
Sporulation in
The Web Computer
Science Magazine Re-

Brown, marry,
Rel. Exp. over
Timecourse

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

Gene Expression Datasets: the Transcriptosome

Proc. Natl. Acad. Sci. USA
Vol. 95, pp. 199–203, January 1997
Genetics

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

PETRA ROSS-MACHONALDI, AMY SHELLIAN, G. SEHILDEEN ROEDER, AND MICHAEL SNYDER*

Department of Biology, Yale University, P.O. Box 20810, New Haven,
Connecticut 06520; Glycolipid Laboratory, Wellcome Research Laboratories, Beckenham, Kent, BR3 4ST, United Kingdom

ABSTRACT Analysis of the function of a particular product typically involves determining the expression profile of the gene, the subcellular location of the protein, and phenotype of a null strain lacking the protein. Conditional alleles of the gene are often created as an additional tool to facilitate these analyses. We describe a system that simultaneously generates constructs for all three analyses and is suitable for mutagenesis of any given *Saccharomyces cerevisiae* gene. Depending on the transposon used, the yeast gene is fused to a coding region for β-galactosidase or GFP, or to a promoter element that directs expression of the gene in a tissue-specific manner. Each construct contains the coding region for the gene of interest, the yeast promoter and the yeast ribosomal DNA, and a TR cassette containing a unique restriction site (SphI, KpnI, SalI, SacI, EcoRI, XbaI, or NotI). In addition, there are flanking lox sites and loxP sites that flank the TR cassette. The TR cassette contains the URA3 marker gene, which is required for growth on 5-fluoroorotic acid (5-FOA). The URA3 gene is under the control of a constitutive promoter, and it is preceded by a loxP site. The loxP site allows recombination with a loxP site in the 5'-flanking region of the gene of interest, thereby creating a null allele. These constructs allow identification of recessive lesions because a transposon insertion can be inverted in the encoded protein. In addition,

transposons contain the TR cassette for induction of transposon mutagenesis. The encoded construct contains the same elements as the other constructs, except that the TR cassette contains the URA3 marker gene in its orientation. This orientation is required for growth on 5-FOA.

With both the full-length and the null alleles, about 10% of the yeast population is able to grow on 5-FOA. This frequency is similar to that of a null allele created by standard methods. The advantage of this system is that it is a multipurpose system that can be used for all three analyses simultaneously. This is particularly useful for identifying regulatory elements in a gene, because it is easier to determine the regulatory elements in a gene if the gene is mutated in more than one way. The site of the insertion provides a unique identifier for a transposon insertion.

is incorporated in *S. cerevisiae* by double recombination. DNA sequencing of the insertion site then reveals from the insertion site the location of the insertion and the nature of the insertion. This allows the insertion to be located in the 5'-flanking region by homologous recombination.

With both the full-length and the null alleles, about 10% of the yeast population is able to grow on 5-FOA. This frequency is similar to that of a null allele created by standard methods. The advantage of this system is that it is a multipurpose system that can be used for all three analyses simultaneously. This is particularly useful for identifying regulatory elements in a gene, because it is easier to determine the regulatory elements in a gene if the gene is mutated in more than one way. The site of the insertion provides a unique identifier for a transposon insertion.

With both the full-length and the null alleles, about 10% of the yeast population is able to grow on 5-FOA. This frequency is similar to that of a null allele created by standard methods. The advantage of this system is that it is a multipurpose system that can be used for all three analyses simultaneously. This is particularly useful for identifying regulatory elements in a gene, because it is easier to determine the regulatory elements in a gene if the gene is mutated in more than one way. The site of the insertion provides a unique identifier for a transposon insertion.

With both the full-length and the null alleles, about 10% of the yeast population is able to grow on 5-FOA. This frequency is similar to that of a null allele created by standard methods. The advantage of this system is that it is a multipurpose system that can be used for all three analyses simultaneously. This is particularly useful for identifying regulatory elements in a gene, because it is easier to determine the regulatory elements in a gene if the gene is mutated in more than one way. The site of the insertion provides a unique identifier for a transposon insertion.

With both the full-length and the null alleles, about 10% of the yeast population is able to grow on 5-FOA. This frequency is similar to that of a null allele created by standard methods. The advantage of this system is that it is a multipurpose system that can be used for all three analyses simultaneously. This is particularly useful for identifying regulatory elements in a gene, because it is easier to determine the regulatory elements in a gene if the gene is mutated in more than one way. The site of the insertion provides a unique identifier for a transposon insertion.

With both the full-length and the null alleles, about 10% of the yeast population is able to grow on 5-FOA. This frequency is similar to that of a null allele created by standard methods. The advantage of this system is that it is a multipurpose system that can be used for all three analyses simultaneously. This is particularly useful for identifying regulatory elements in a gene, because it is easier to determine the regulatory elements in a gene if the gene is mutated in more than one way. The site of the insertion provides a unique identifier for a transposon insertion.

With both the full-length and the null alleles, about 10% of the yeast population is able to grow on 5-FOA. This frequency is similar to that of a null allele created by standard methods. The advantage of this system is that it is a multipurpose system that can be used for all three analyses simultaneously. This is particularly useful for identifying regulatory elements in a gene, because it is easier to determine the regulatory elements in a gene if the gene is mutated in more than one way. The site of the insertion provides a unique identifier for a transposon insertion.

With both the full-length and the null alleles, about 10% of the yeast population is able to grow on 5-FOA. This frequency is similar to that of a null allele created by standard methods. The advantage of this system is that it is a multipurpose system that can be used for all three analyses simultaneously. This is particularly useful for identifying regulatory elements in a gene, because it is easier to determine the regulatory elements in a gene if the gene is mutated in more than one way. The site of the insertion provides a unique identifier for a transposon insertion.

With both the full-length and the null alleles, about 10% of the yeast population is able to grow on 5-FOA. This frequency is similar to that of a null allele created by standard methods. The advantage of this system is that it is a multipurpose system that can be used for all three analyses simultaneously. This is particularly useful for identifying regulatory elements in a gene, because it is easier to determine the regulatory elements in a gene if the gene is mutated in more than one way. The site of the insertion provides a unique identifier for a transposon insertion.

With both the full-length and the null alleles, about 10% of the yeast population is able to grow on 5-FOA. This frequency is similar to that of a null allele created by standard methods. The advantage of this system is that it is a multipurpose system that can be used for all three analyses simultaneously. This is particularly useful for identifying regulatory elements in a gene, because it is easier to determine the regulatory elements in a gene if the gene is mutated in more than one way. The site of the insertion provides a unique identifier for a transposon insertion.

With both the full-length and the null alleles, about 10% of the yeast population is able to grow on 5-FOA. This frequency is similar to that of a null allele created by standard methods. The advantage of this system is that it is a multipurpose system that can be used for all three analyses simultaneously. This is particularly useful for identifying regulatory elements in a gene, because it is easier to determine the regulatory elements in a gene if the gene is mutated in more than one way. The site of the insertion provides a unique identifier for a transposon insertion.

With both the full-length and the null alleles, about 10% of the yeast population is able to grow on 5-FOA. This frequency is similar to that of a null allele created by standard methods. The advantage of this system is that it is a multipurpose system that can be used for all three analyses simultaneously. This is particularly useful for identifying regulatory elements in a gene, because it is easier to determine the regulatory elements in a gene if the gene is mutated in more than one way. The site of the insertion provides a unique identifier for a transposon insertion.

With both the full-length and the null alleles, about 10% of the yeast population is able to grow on 5-FOA. This frequency is similar to that of a null allele created by standard methods. The advantage of this system is that it is a multipurpose system that can be used for all three analyses simultaneously. This is particularly useful for identifying regulatory elements in a gene, because it is easier to determine the regulatory elements in a gene if the gene is mutated in more than one way. The site of the insertion provides a unique identifier for a transposon insertion.

With both the full-length and the null alleles, about 10% of the yeast population is able to grow on 5-FOA. This frequency is similar to that of a null allele created by standard methods. The advantage of this system is that it is a multipurpose system that can be used for all three analyses simultaneously. This is particularly useful for identifying regulatory elements in a gene, because it is easier to determine the regulatory elements in a gene if the gene is mutated in more than one way. The site of the insertion provides a unique identifier for a transposon insertion.

With both the full-length and the null alleles, about 10% of the yeast population is able to grow on 5-FOA. This frequency is similar to that of a null allele created by standard methods. The advantage of this system is that it is a multipurpose system that can be used for all three analyses simultaneously. This is particularly useful for identifying regulatory elements in a gene, because it is easier to determine the regulatory elements in a gene if the gene is mutated in more than one way. The site of the insertion provides a unique identifier for a transposon insertion.

With both the full-length and the null alleles, about 10% of the yeast population is able to grow on 5-FOA. This frequency is similar to that of a null allele created by standard methods. The advantage of this system is that it is a multipurpose system that can be used for all three analyses simultaneously. This is particularly useful for identifying regulatory elements in a gene, because it is easier to determine the regulatory elements in a gene if the gene is mutated in more than one way. The site of the insertion provides a unique identifier for a transposon insertion.

With both the full-length and the null alleles, about 10% of the yeast population is able to grow on 5-FOA. This frequency is similar to that of a null allele created by standard methods. The advantage of this system is that it is a multipurpose system that can be used for all three analyses simultaneously. This is particularly useful for identifying regulatory elements in a gene, because it is easier to determine the regulatory elements in a gene if the gene is mutated in more than one way. The site of the insertion provides a unique identifier for a transposon insertion.

With both the full-length and the null alleles, about 10% of the yeast population is able to grow on 5-FOA. This frequency is similar to that of a null allele created by standard methods. The advantage of this system is that it is a multipurpose system that can be used for all three analyses simultaneously. This is particularly useful for identifying regulatory elements in a gene, because it is easier to determine the regulatory elements in a gene if the gene is mutated in more than one way. The site of the insertion provides a unique identifier for a transposon insertion.

With both the full-length and the null alleles, about 10% of the yeast population is able to grow on 5-FOA. This frequency is similar to that of a null allele created by standard methods. The advantage of this system is that it is a multipurpose system that can be used for all three analyses simultaneously. This is particularly useful for identifying regulatory elements in a gene, because it is easier to determine the regulatory elements in a gene if the gene is mutated in more than one way. The site of the insertion provides a unique identifier for a transposon insertion.

With both the full-length and the null alleles, about 10% of the yeast population is able to grow on 5-FOA. This frequency is similar to that of a null allele created by standard methods. The advantage of this system is that it is a multipurpose system that can be used for all three analyses simultaneously. This is particularly useful for identifying regulatory elements in a gene, because it is easier to determine the regulatory elements in a gene if the gene is mutated in more than one way. The site of the insertion provides a unique identifier for a transposon insertion.

With both the full-length and the null alleles, about 10% of the yeast population is able to grow on 5-FOA. This frequency is similar to that of a null allele created by standard methods. The advantage of this system is that it is a multipurpose system that can be used for all three analyses simultaneously. This is particularly useful for identifying regulatory elements in a gene, because it is easier to determine the regulatory elements in a gene if the gene is mutated in more than one way. The site of the insertion provides a unique identifier for a transposon insertion.

bioinfo.mbb.yale.edu

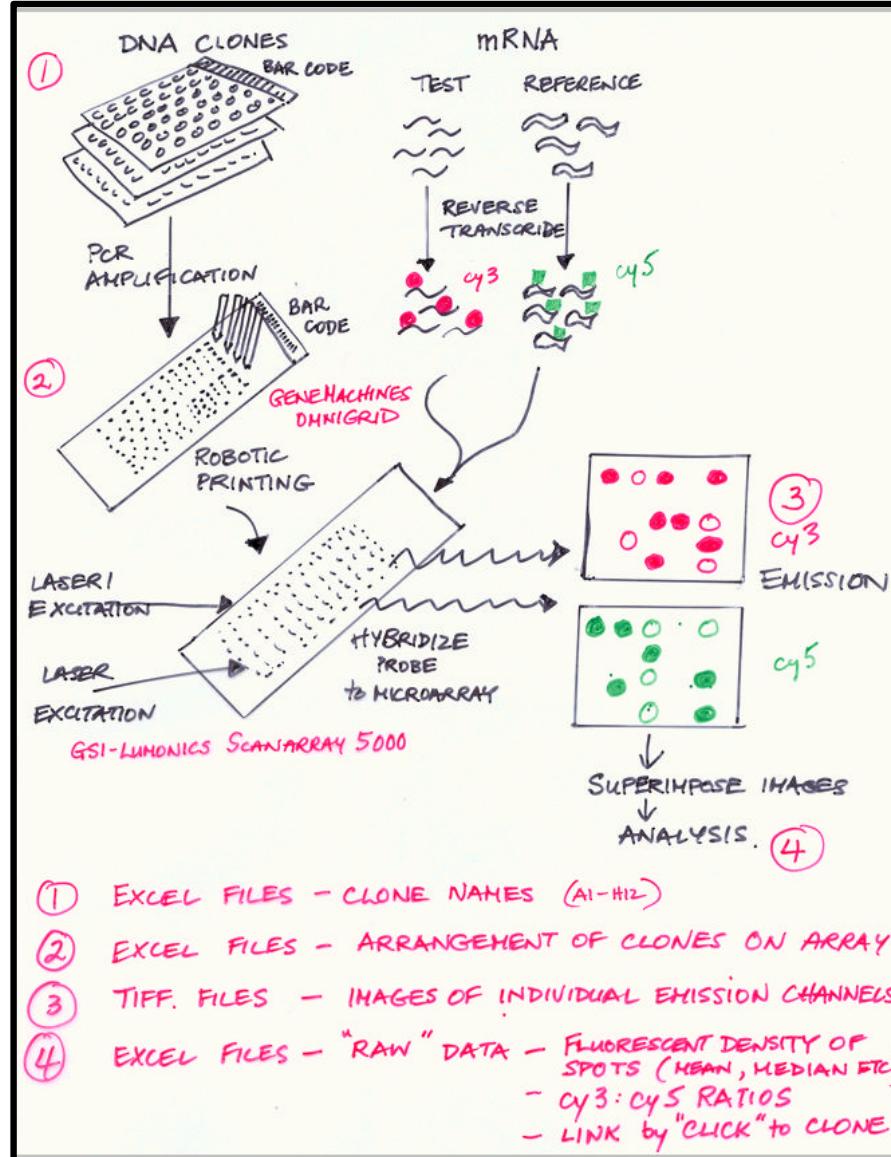
Array Data

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,
 $6000 \times 10 = 60K$ floats

telling signal from background



(courtesy of J Hager)

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 Rocío Benito,⁵ Jef D. Boeke,⁶ Howard Burd,⁷ Carla Connelly,⁸ Karen Davis,¹ Fred Dietrich,⁹ Mohamed El Bakkoury,⁹ Françoise Foury,¹⁰ Erik Gentalen,¹¹ Guri Giaever,⁷ Johan Ted Jones,¹ Michael Laub,¹ Hong Liao,¹¹ David J. Lockhart,¹¹ Anca Lucau-Dan,¹² Nasiba M'Rabet,³ Patrice Menard,⁷ Michael Chai Pai,¹ Corinne Rebischung,⁸ Jose L. Ross-Macdonald,¹³ Christopher J. Roberts,² Petra Ross-Madon,¹⁴ Michael Snyder,⁴ Sharon Sookhai-Mahadeo,¹⁵ Steeve Véronneau,⁷ Marleen Voet,¹⁴ Teresa R. Ward,² Robert Wysocki,¹⁰ Grzegorz Katja Zimmermann,¹² Peter Mark Johnston,¹² Ronald W. Davis,¹¹

The functions of many open reading frames (ORFs) in sequencing projects are unknown. New, whole-genome approaches are needed to systematically determine their function. A total of 4200 *S. cerevisiae* strains were constructed, by a high-throughput deletion of one of 2026 ORFs (more than 10% of the genome). Of the deleted ORFs, 17 percent were medium. The phenotypes of more than 500 deletion strains were parallel. Of the deletion strains, 40 percent showed growth 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

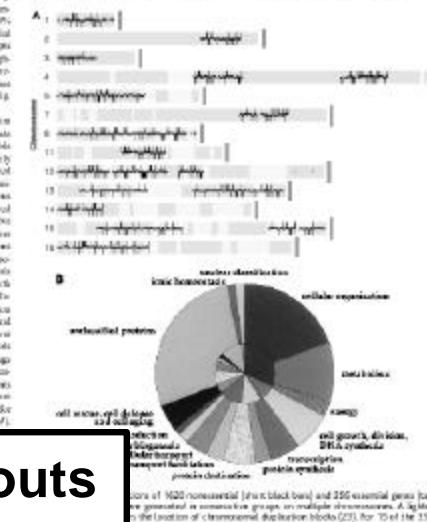
identify essential genes (6%), genes with 1 kb of another genes (47% of nonessential genes), and genes with 5 kb of the reference (1%).

Genes that are more highly expressed than others were deleted for >99%

of strains. Some genes are expressed in only a subset of strains, and these genes will likely be under very specific control, necessitating the use of different conditions. Proteins encoded by these genes may have distinct roles in their respective cellular processes. In addition, other 12 strains were used for many processes, including those that are poorly understood. The phenotypic analysis of these strains will allow us to be surprised when we find new interactions.

During this time, aliquots from the two pools, the rags and lymphocytes, were taken at regular intervals.

The lag phase represents the time between the relative growth rates for each in the population (7).



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

Other Whole-Genome Experiments



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 Luo 1, Mengsheng Qiu 1,3, Eva Chan 2, Helen Zhou 4, Li Zhu

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

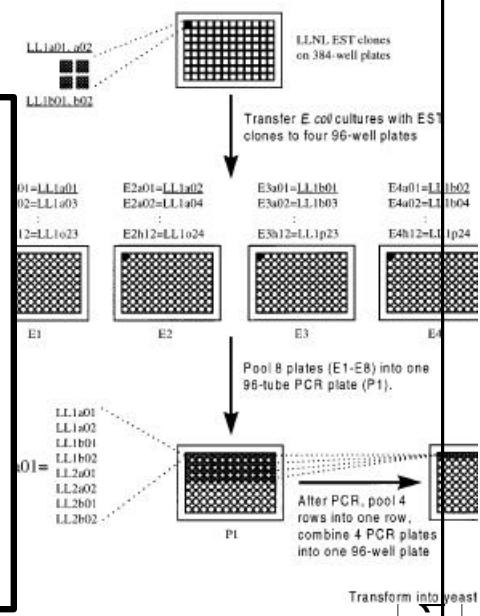
Received 1 February 1998; received in revised form 28 April 1998; accepted 29 April 1998; Received by E.Y. Chen

Abstract

Identification of all human proteins is important information for functional studies. Protein-protein interactions can be studied by constructing modular yeast two-hybrid cDNA libraries.

148

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



2 hybrids, linkage maps

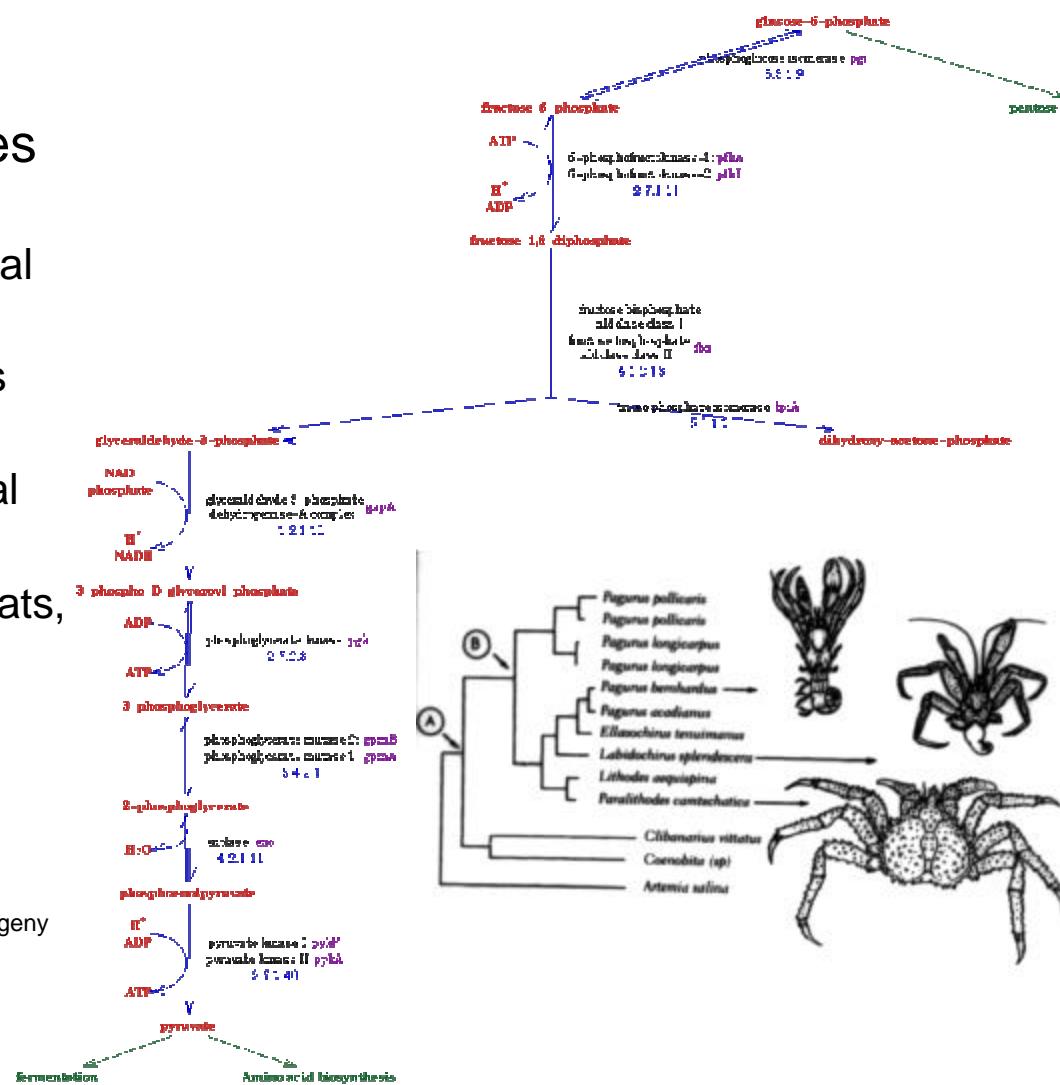
Hua, S. B., Luo, Y., Qiu, M., Chan, E., Zhou, H. & Zhu, L. (1998). Construction of a modular yeast two-hybrid cDNA library from human EST clones for the human genome protein linkage map. *Gene* **215**, 143-52

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

Molecular Biology Information: Other Integrative Data

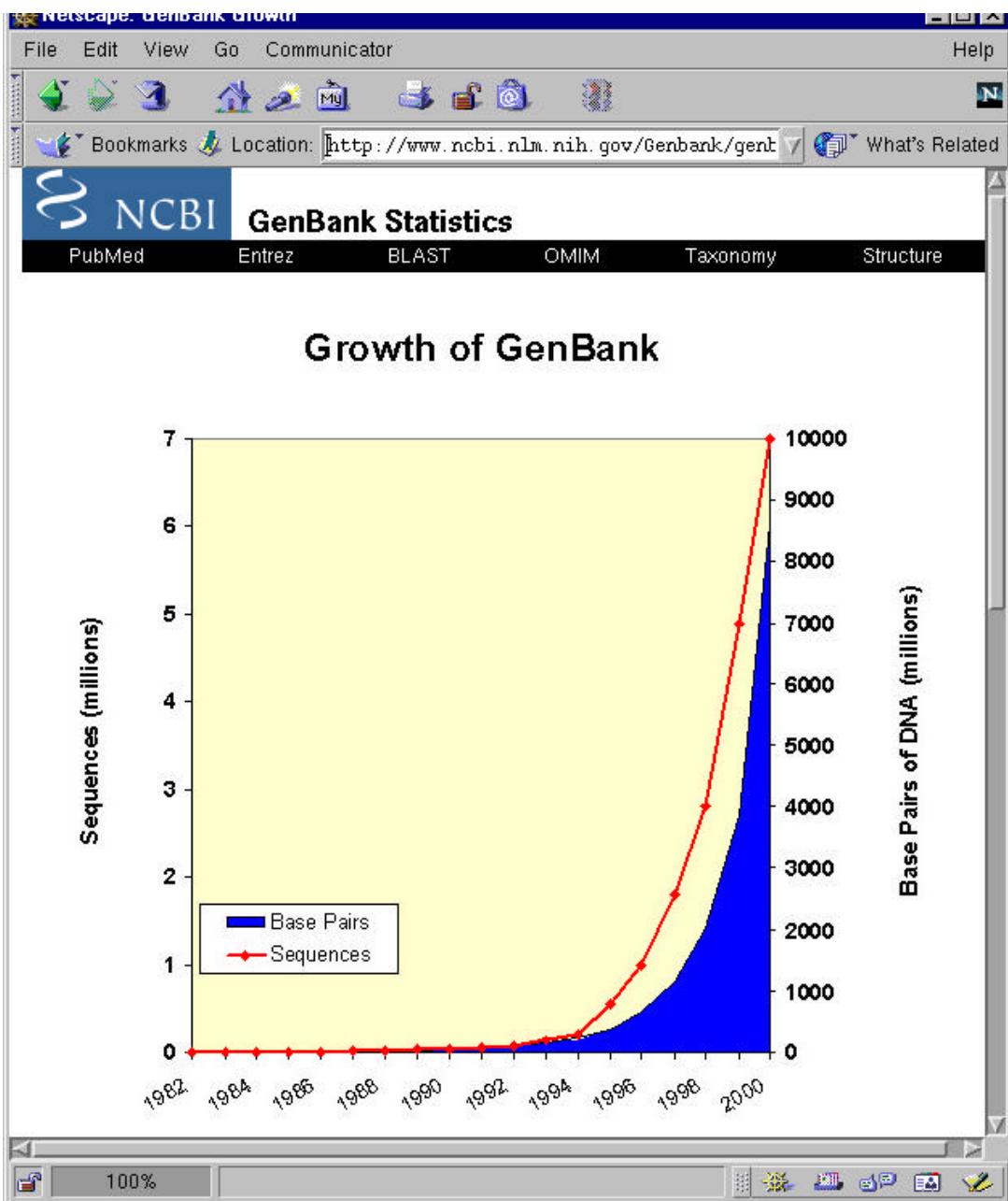
- Information to understand genomes
 - ◊ Metabolic Pathways (glycolysis), traditional biochemistry
 - ◊ Regulatory Networks
 - ◊ Whole Organisms Phylogeny, traditional zoology
 - ◊ Environments, Habitats, ecology
 - ◊ The Literature (MEDLINE)
- The Future....

(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**.



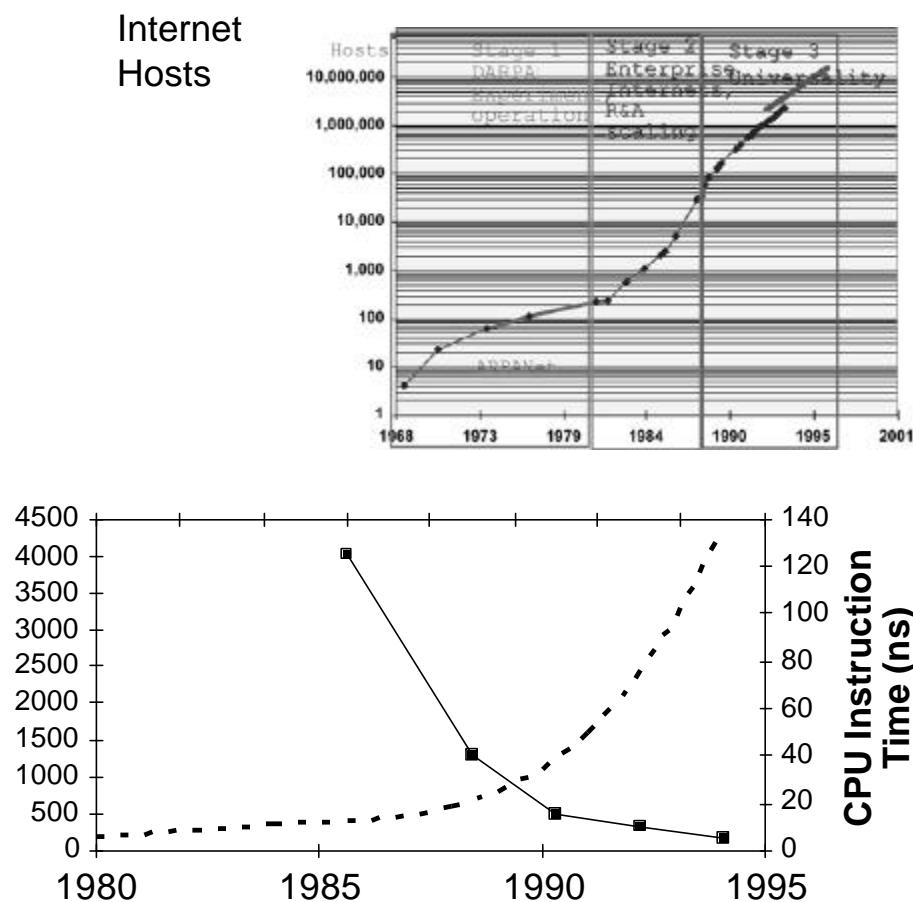
Large-scale Information: GenBank Growth

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

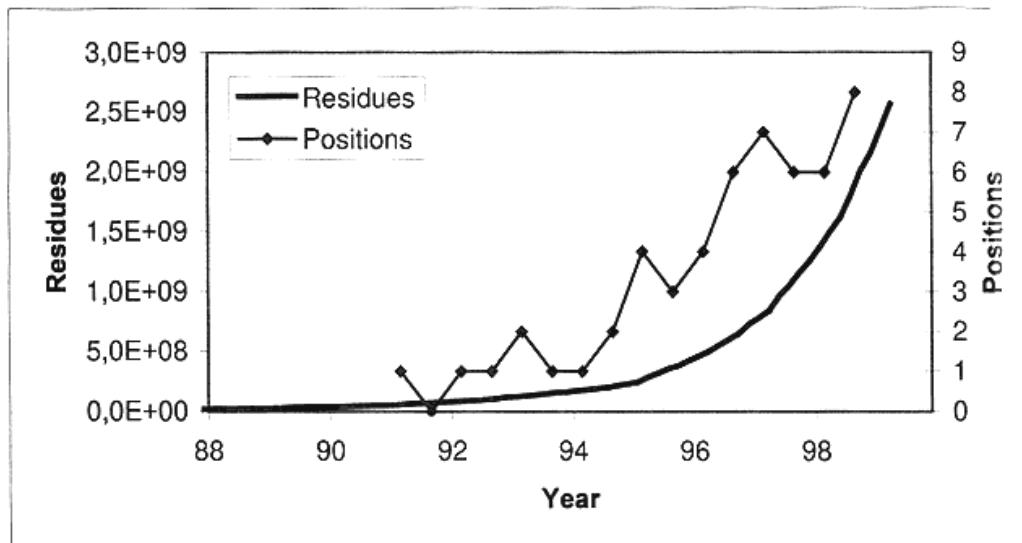
- CPU vs Disk & Net
 - ◊ As important as the increase in computer speed has been, the ability to store large amounts of information on computers is even more crucial
- Driving Force in Bioinformatics

(Internet picture adapted from D Brutlag, Stanford)

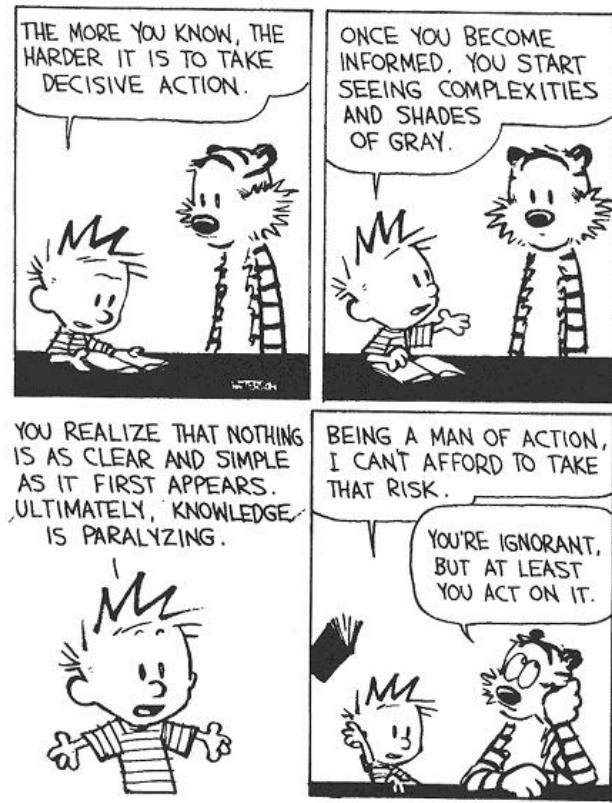
Num.
Protein
Domain
Structures



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.



(courtesy of Finn Drablos)

Weber
Cartoon



"Don't just sit there! If you've processed all the data there is, go out and find more data!"

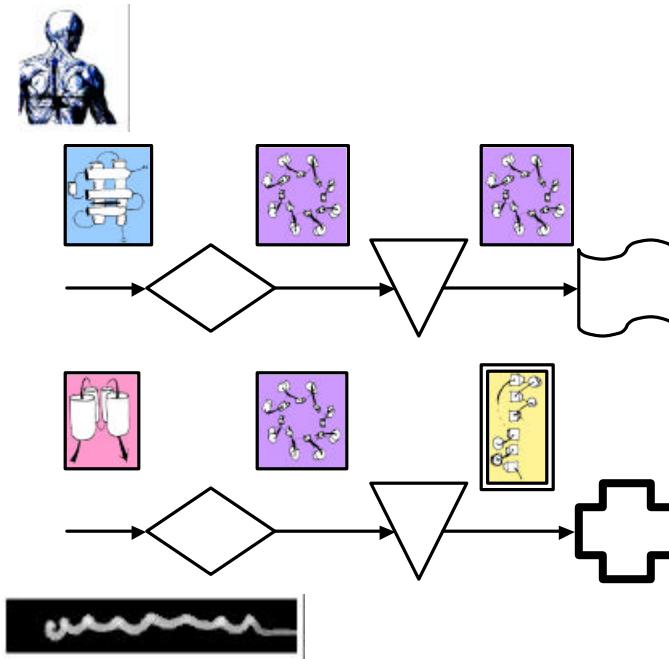
Reproduced in R.L. Weber, "A random walk in science", IOP Publishing, 1973

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

Netscape: NCBI CDD Help

File Edit View Go Communicator Help

Bookmarks Location: http://www.ncbi.nlm.nih.gov/Structure/cdd/ What's Related

NCBI CDD

PubMed BLAST OMIM Taxonomy Entrez Structure

Search Entrez Structure for [] Go

CDD Home

Conserved Domain Database

MMDB

NCBI's structure database

PDBeast

Taxonomy in MMDB

Ch3D v3.0

3D-structure viewer

VAST

Structure comparisons

VAST Search

Submit structure database searches

Research

Research topics and staff

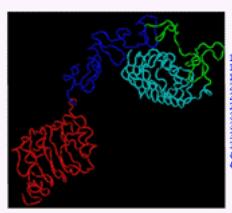
CDD - Conserved Domain Database Help

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 [Ch3D](#)):



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.

1 EICGGPIDIIR HNGVQNLNLK NCTVIEGVLH
31 LILISIAEAEV RSEYRFELTY ITETYLLEFLWV
61 AGCCELGOLF PHVETVIRGK LFVTVALVIF
91 EKTHLKDGLI YHWRNITRGQ IRIENNAALC
121 VLTSTVNSLLI LDAVSNHYIV GNPKPKKGD
151 LCFCGTMEKEK NCEKTTLINHE YHVRCVTINR
181 COKMCPESTG KRACTEENEC CIPCELGSCS
211 AFDNHTACVA CRHTYYAGVC VPACPPNTYR
241 FEGWRCVDRG FCANLISABE SDSEGFIVHD
271 GBCHOECPG FIRGSOSMY CIPCEGCPK
301 VCEERKKTTK IDSTVSAOMI OGCTIFPGNL
331 LINIRRGNNI ASELENFHQL IEVFTGVK1
361 VSHASLHLSLIS ESELGKSYSE ESELGKSYSE
391 FVLDONLQO LUVDWBRLLT IKAQMYFAF
421 NPKLCPSEIY RHEEVVTGKG ROSGDQINTR
451 NNGERASCES DVIDDQDEBK LISEEDLN

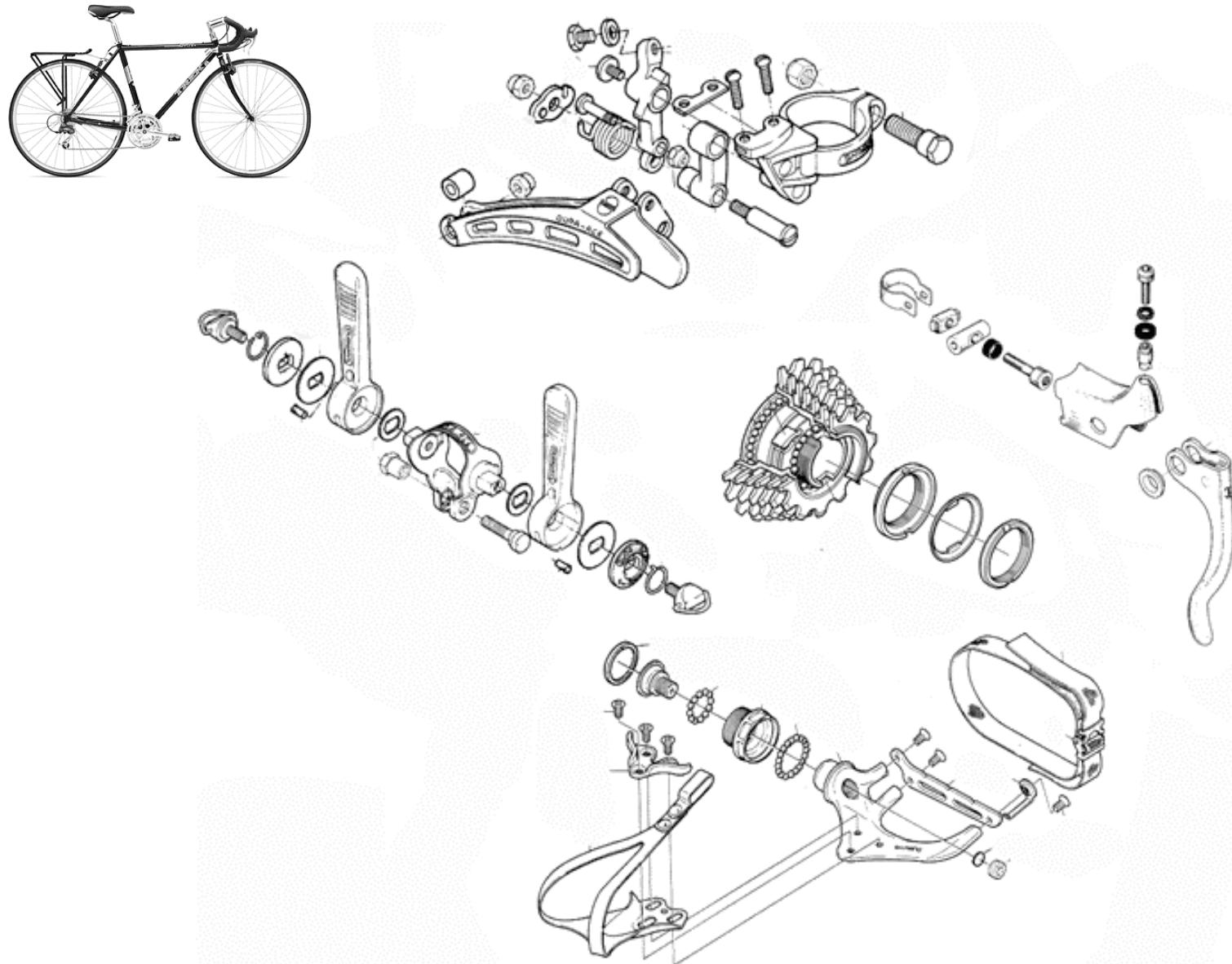
Recep_L-domain Furin-like Recep_L-domain

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.

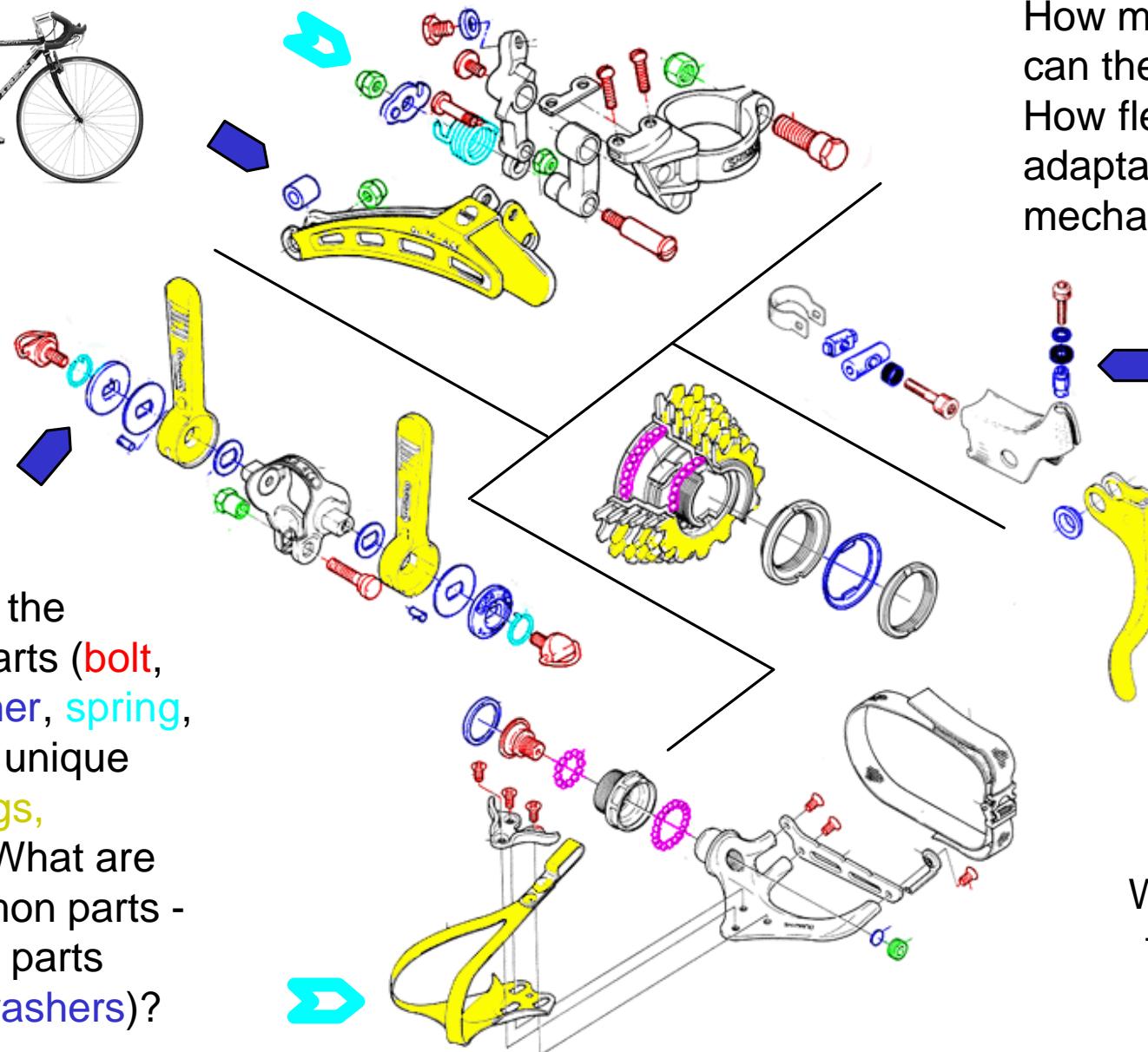
Conserved domains contain conserved sequence patterns or motifs, which allow for their detection in polypeptide sequences. The distinction between domains and motifs is not sharp, however, especially in the case of short repetitive units. Functional motifs are also present outside the scope of structurally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

100% Open the list of newsgroups

A Parts List Approach to Bike Maintenance



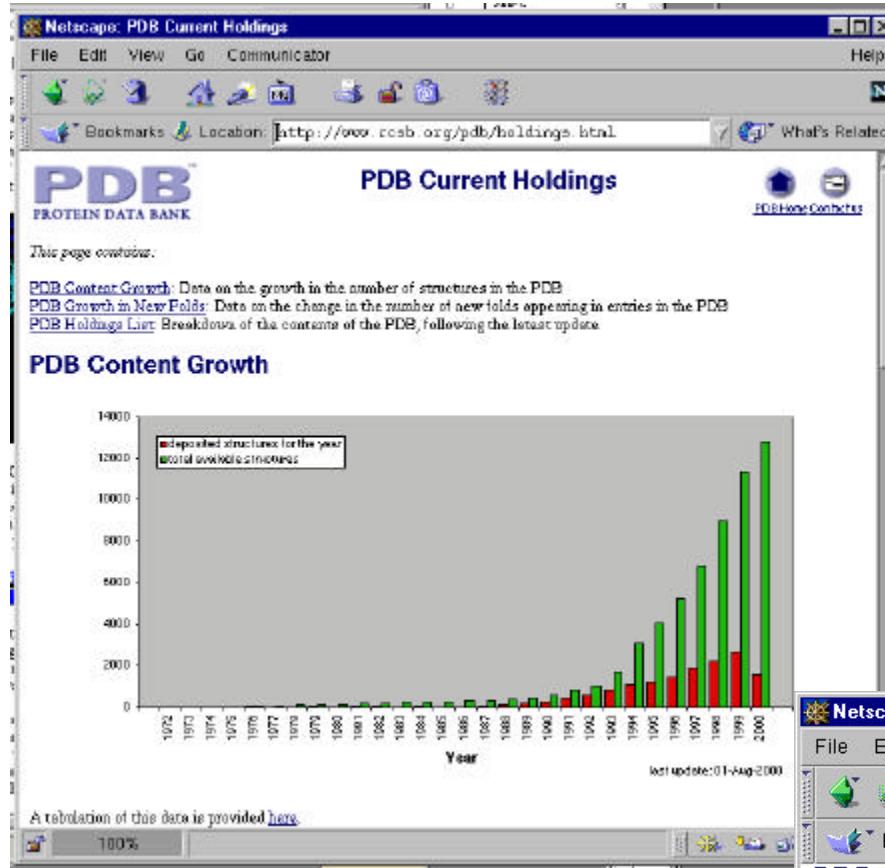
A Parts List Approach to Bike Maintenance



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

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

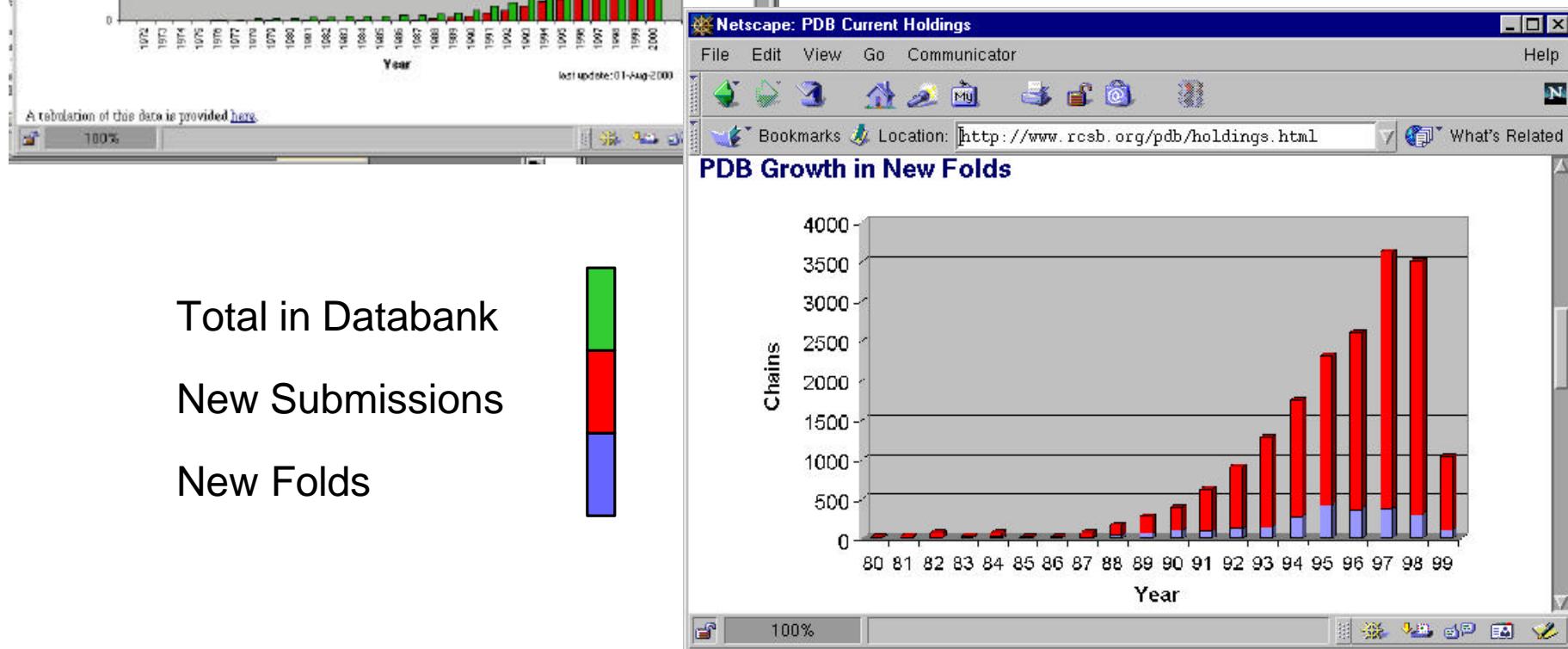
Where are the parts located?



Total in Databank

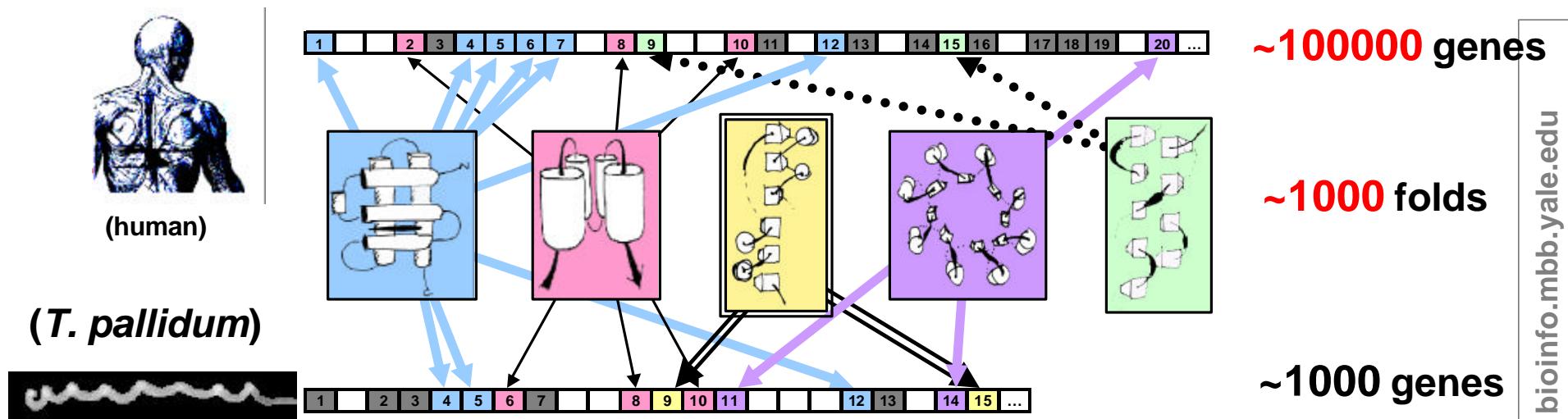
New Submissions

New Folds



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

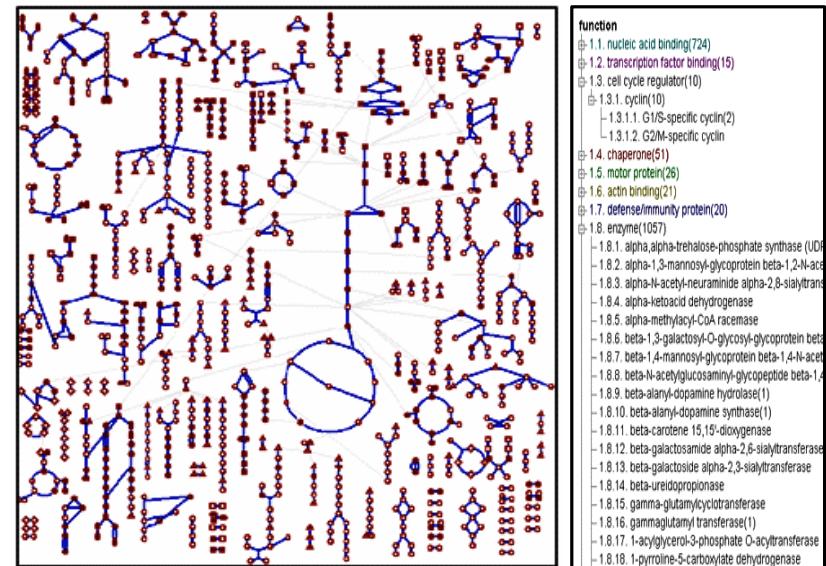
World of Structures is even more Finite, providing a valuable simplification



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

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

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 Topics --

Genome Sequence

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

- 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

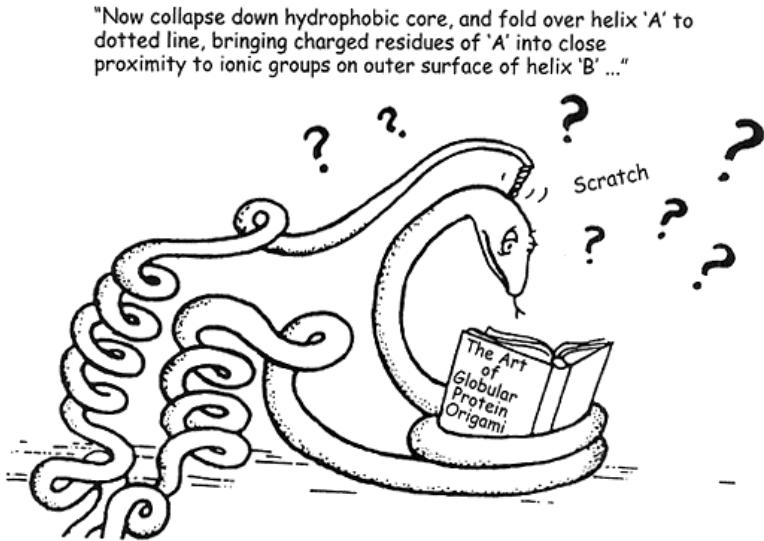
Bioinformatics

Topics --

Sequence /

Structure

- Secondary Structure “Prediction”
 - ◊ via Propensities
 - ◊ Neural Networks, Genetic Alg.
 - ◊ Simple Statistics
 - ◊ TM-helix finding
 - ◊ Assessing Secondary Structure Prediction



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

- 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

- 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?

Topics -- Databases

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

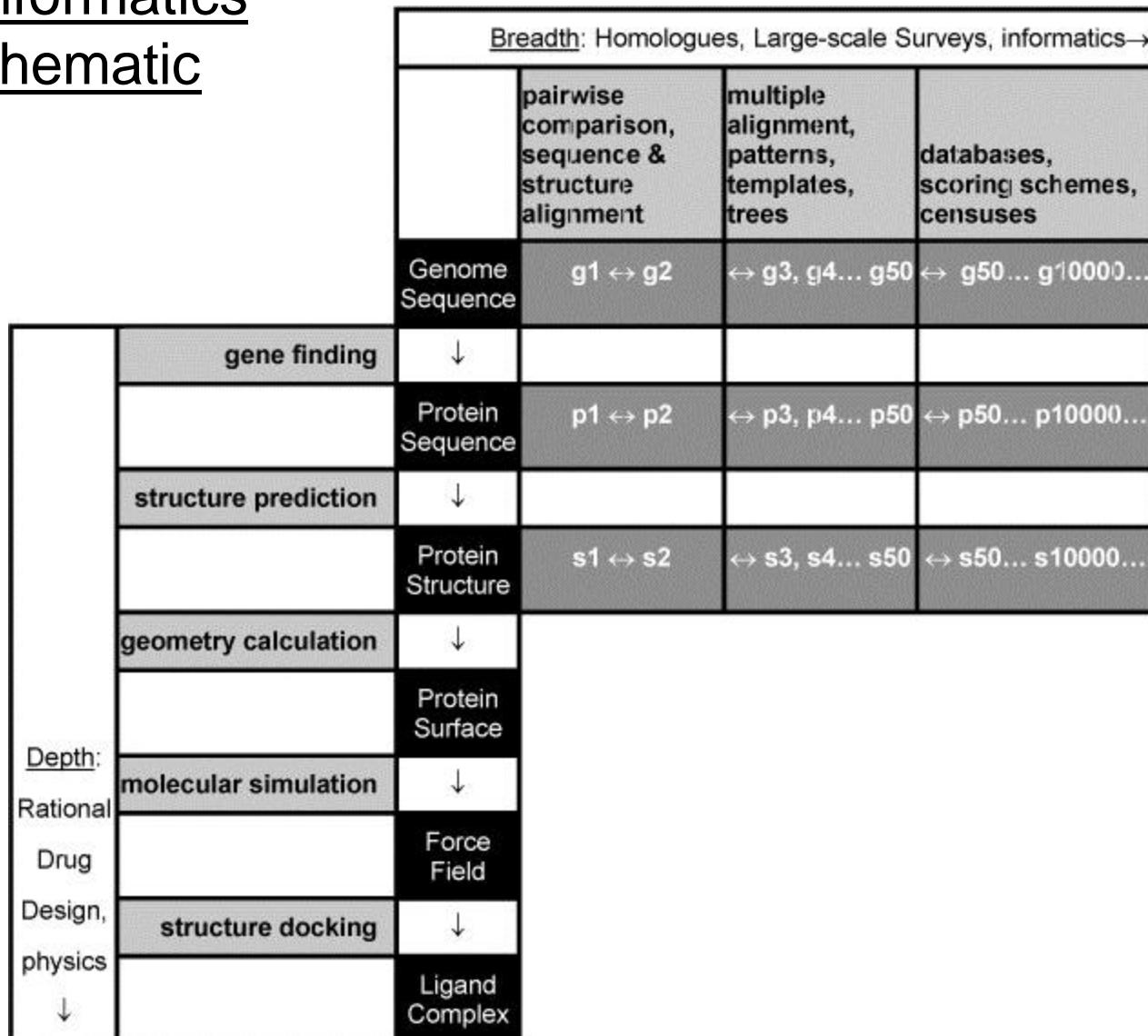
Topics -- Genomics

- 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 Schematic



Background

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

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)

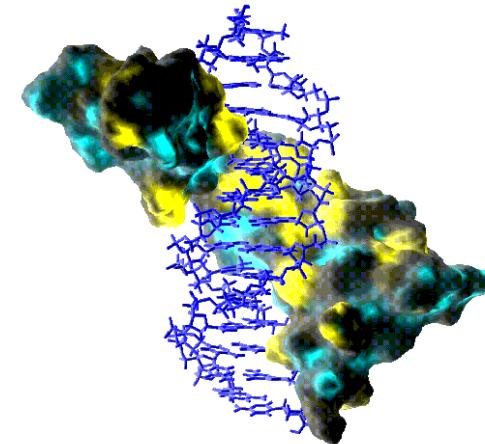
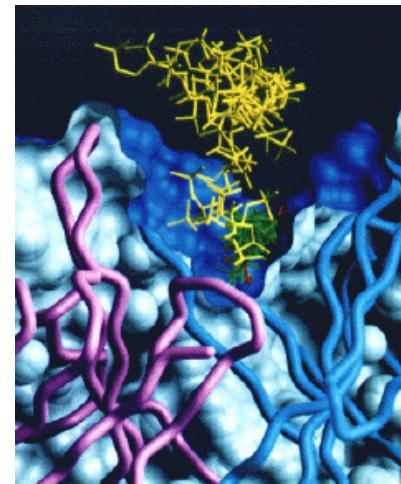
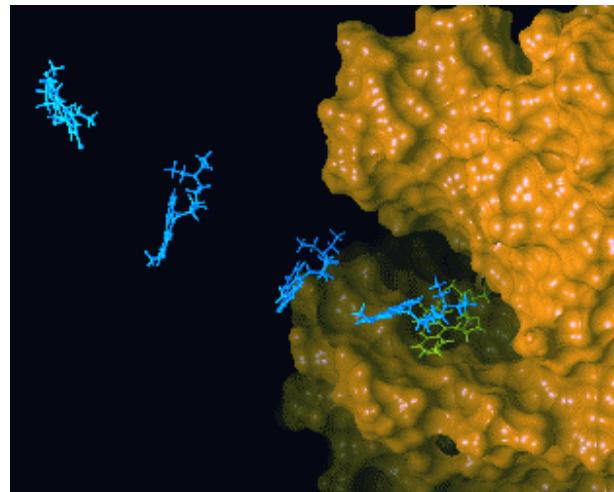
What is Bioinformatics?

- (*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.**

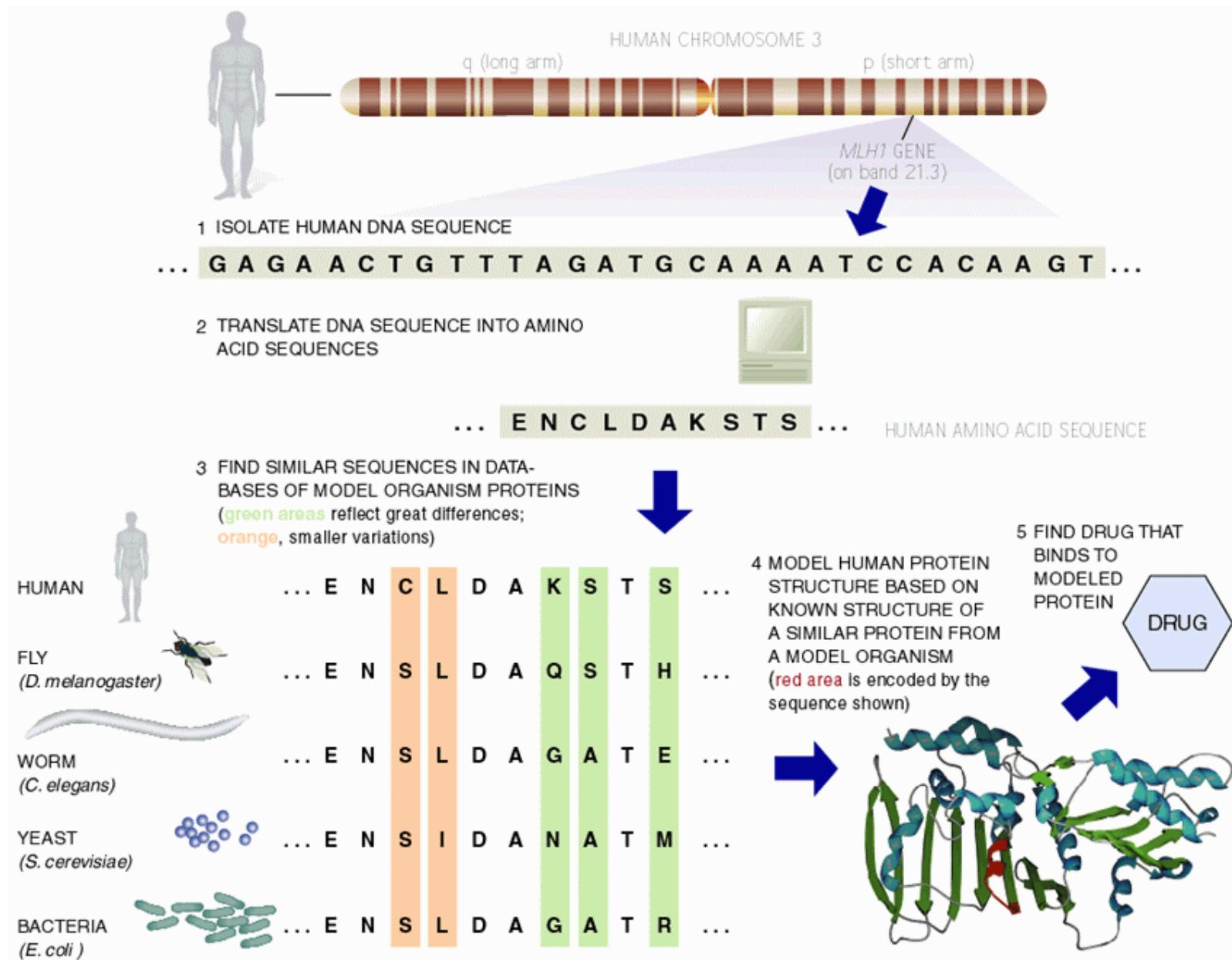
Major Application I: Designing Drugs

- 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.)

Human Disease	MIM #	Human Gene	GenBank Acc# for Human cDNA	BLAST X 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	MSH2	U07418	6.3e-196	MSH1	U07187	DNA repair protein
Cystic Fibrosis	219700	CFTP	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	TELL	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/threonine protein kinase
Menkes Syndrome	309400	MNK	X69208	2.1e-17	CCC2	L36317	Probable copper transporter

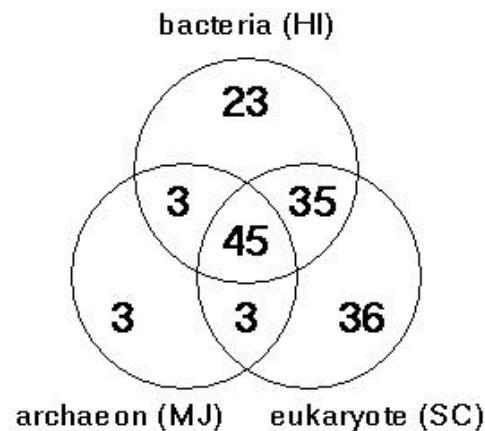
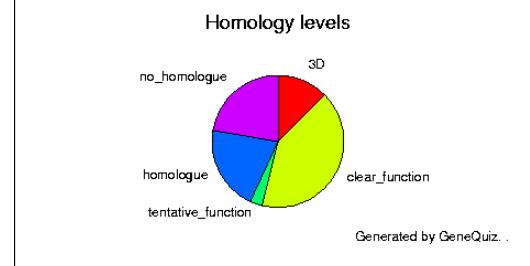
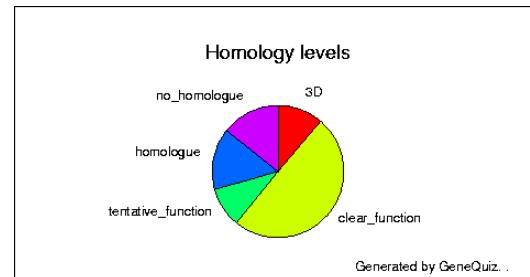
Major Application II: Finding Homologues (cont.)

- 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

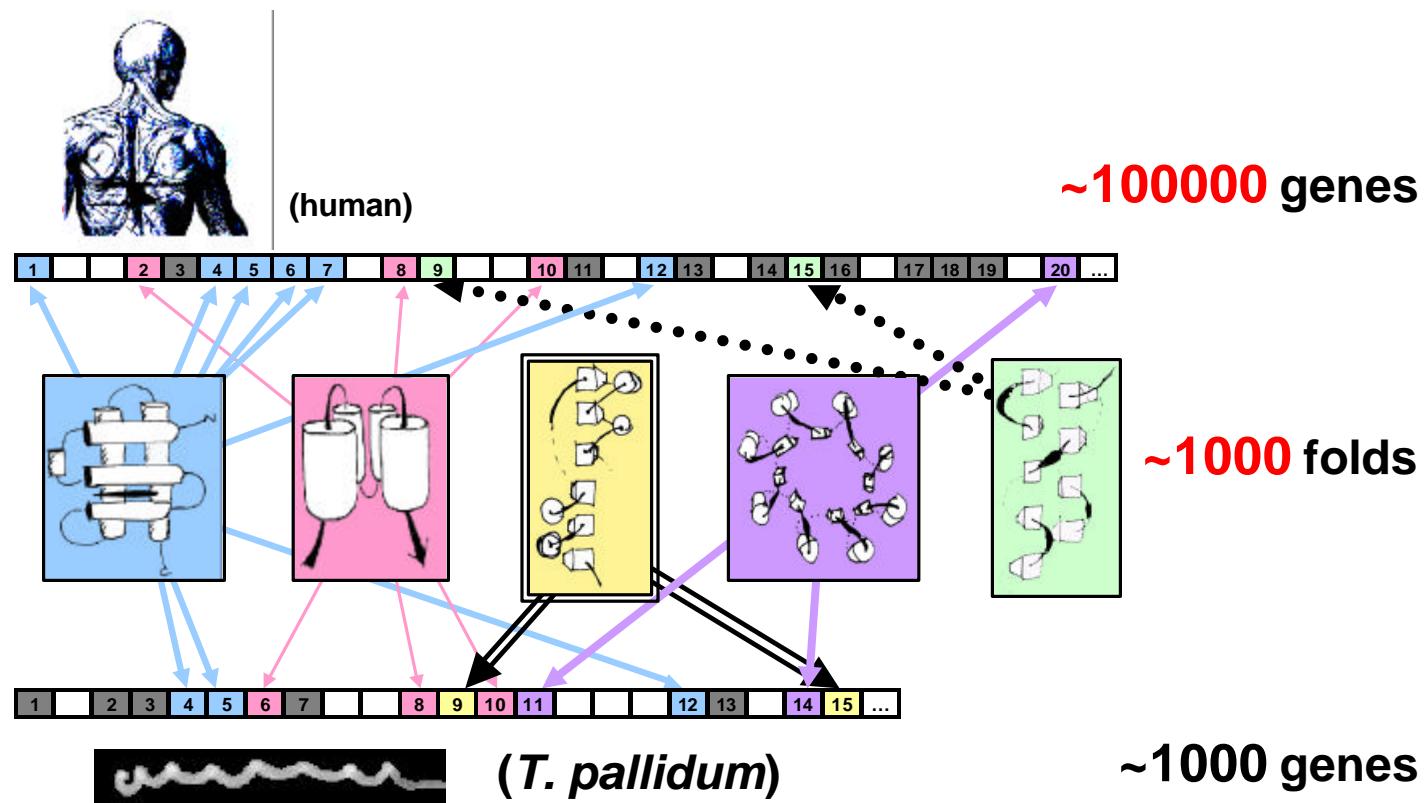
Major Application III: Overall Genome Characterization

- 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

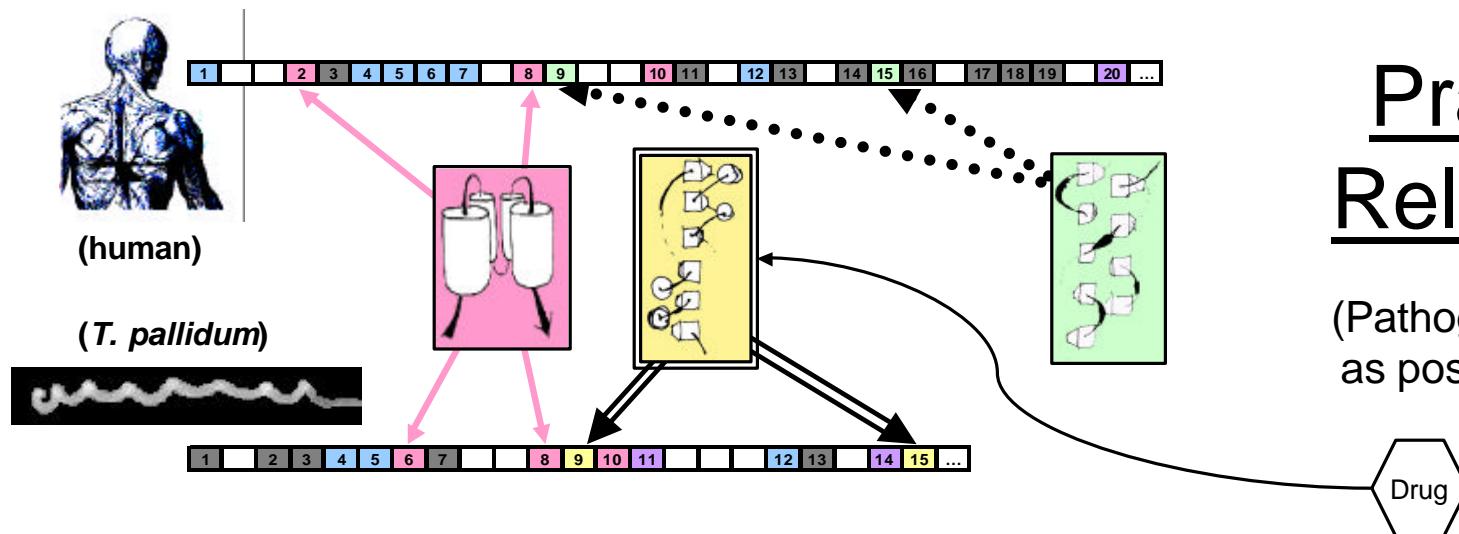
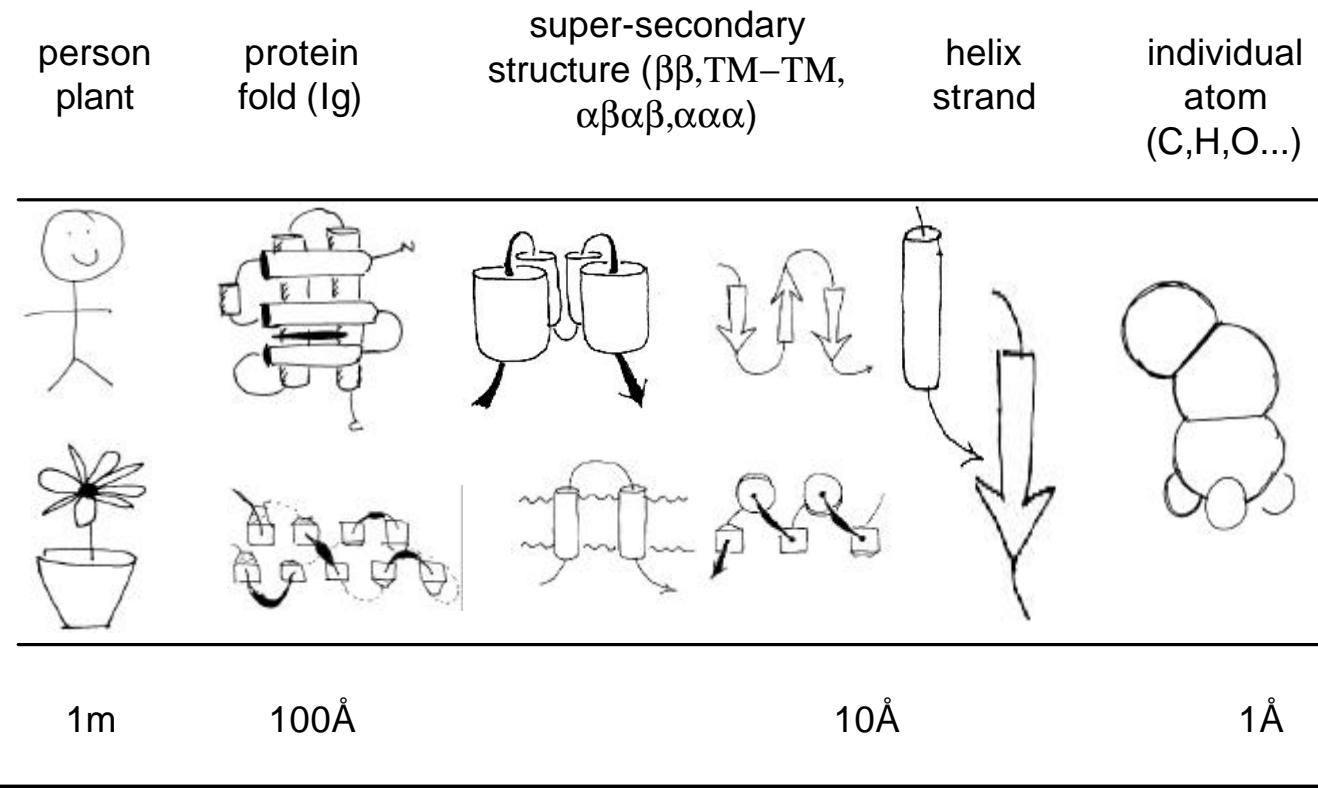
(Clock figures, yeast v. Synechocystis,
adapted from GeneQuiz Web Page, Sander Group, EBI)



Simplifying Genomes with Folds, Pathways, &c



At What Structural Resolution Are Organisms Different?



Practical Relevance

(Pathogen only folds as possible targets)