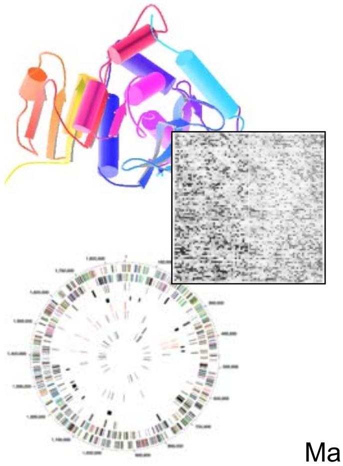
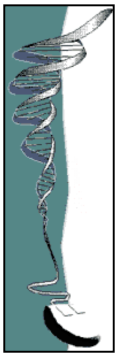
BIOINFORMATICS Introduction

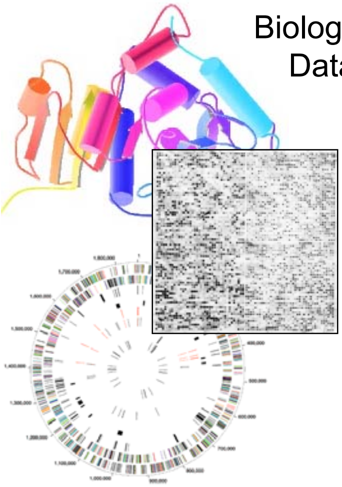


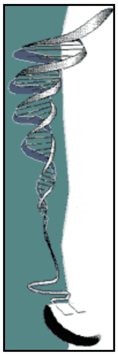


Mark Gerstein, Yale University
bioinfo.mbb.yale.edu/mbb452a
(last edit in fall 2002)

Bioinformatics

Biological Data + Computer Calculations





What is Bioinformatics?

Core

- *(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

- Central Paradigm
  for Bioinformatics

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

- Large Amounts of Information
  - ◊ Standardized
  - ◊ Statistical

•Genetic material

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

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

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

Molecular Biology Information - DNA

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

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

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

Molecular Biology Information: Protein Sequence

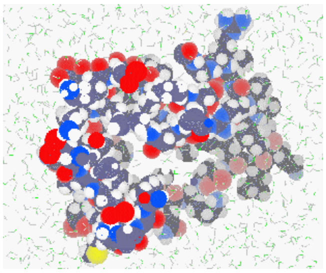
- 20 letter alphabet
  - ◊ ACDEFGHIKLMNPQRSTVWY but not BJOUXZ
- Strings of ~300 aa in an average protein (in bacteria), ~200 aa in a domain
- ~200 K known protein sequences

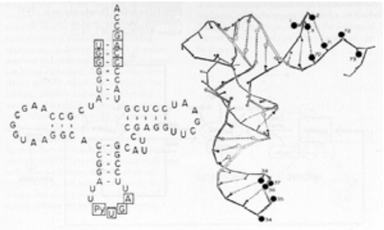
```
d1dhfa_ LNCIVAVSQNMGIGKNGDLPWPPLRNEFRYFQRMTTTSSVEGKQ-NLVIMGKKTWFSI
d8dfr__ LNSIVAVCQNMGIGKDGNLPWPPLRNEYKYFQRMTSTSHVEGKQ-NAVIMGKKTWFSI
d4dfra_ ISLIAALAVDRVIGMENAMPWN-LPADLAWFKRNTL--------NKPVIMGRHTWESI
d3dfr__ TAFLWAQDRDGLIGKDGHLPWH-LPDDLHYFRAQTV--------GKIMVVGRRTYESF

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

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

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

Molecular Biology Information: Macromolecular Structure

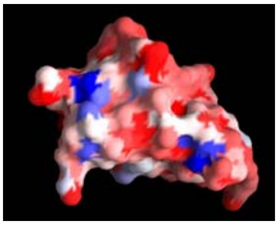
- DNA/RNA/Protein
  - ◊ Almost all protein

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





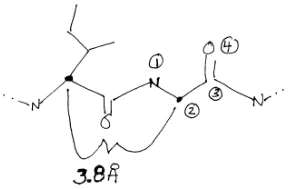
'Identity elements' in Escherichia coli glutamine tRNA.



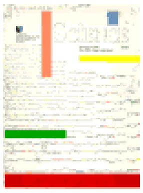
Molecular Biology Information: Protein Structure Details

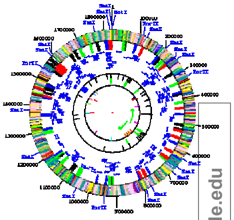
- Statistics on Number of XYZ triplets
  - ◊ 200 residues/domain -> 200 CA atoms, separated by 3.8 A
  - ◊ Avg. Residue is Leu: 4 backbone atoms + 4 sidechain atoms, 150 cubic A
    - => ~1500 xyz triplets (=8x200) per protein domain
  - ◊ 10 K known domain, ~300 folds

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



Molecular Biology Information: Whole Genomes





- The Revolution Driving Everything

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

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

- Integrative Data

  1995, HI (bacteria): 1.6 Mb & 1600 genes done
  1997, yeast: 13 Mb & ~6000 genes for yeast
  1998, 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

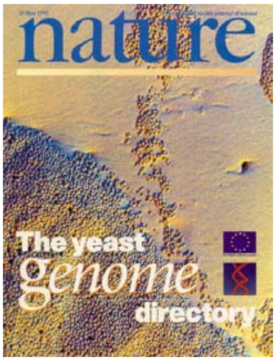
Eukaryote,
13 Mb,
~6K genes
[Nature **387**: 1]

1998

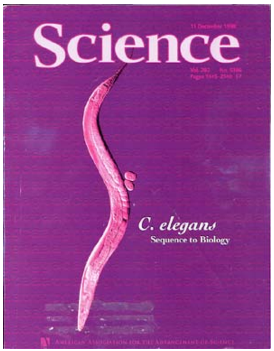
Animal,
~100 Mb,
~20K genes
[*Science* **282**: 1945]

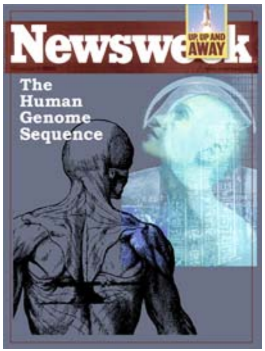
2000?

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



Genomes highlight the Finiteness of the "Parts" in Biology



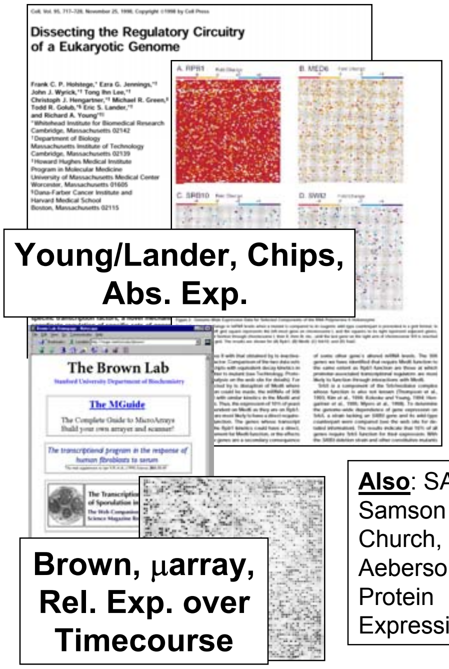


'98 spoof

↔

real thing, Apr '00

Gene Expression Datasets: the Transcriptome



Cell, Vol. 95, 717–728, November 25, 1998, Copyright ©1998 by Cell Press

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

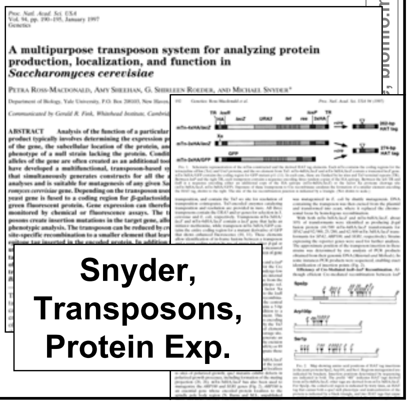
The Complete Guide to MicroArrays
Build your own arrayer and scanner!

The transcriptional program in the response of human fibroblasts to serum

The Transcription
of Sporulation in
The Web Companion
Science Magazine

Brown, μarray, 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*

Department of Biology, Yale University, P.O. Box 208103, New Haven,

Communicated by Gerald R. Fink, Whitehead Institute, Cambrid

Snyder, Transposons, Protein Exp.

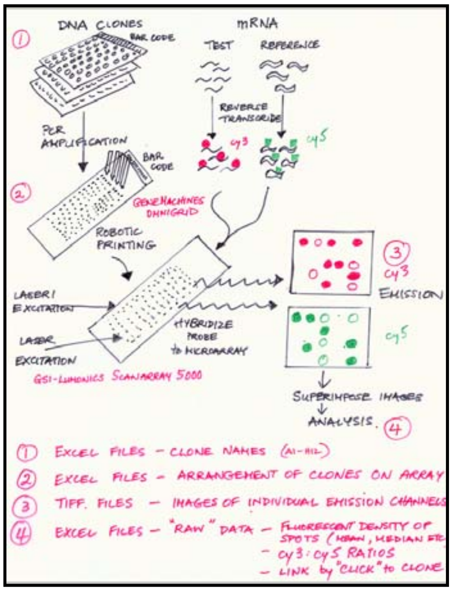
Array Data

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



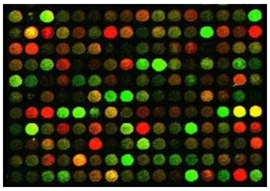
1. EXCEL FILES – CLONE NAMES (A1-H12)
2. EXCEL FILES – ARRANGEMENT OF CLONES ON ARRAY
3. TIFF. FILES – IMAGES OF INDIVIDUAL EMISSION CHANNELS
4. EXCEL FILES – "RAW" DATA – FLUORESCENT DENSITY OF SPOTS (MEAN, MEDIAN ETC
   – cy3 : cy5 RATIOS
   – LINK by "CLICK" to CLONE

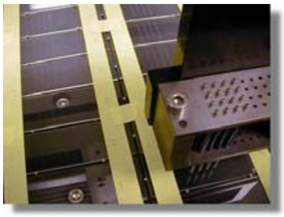
(courtesy of J Hager)

microarrays

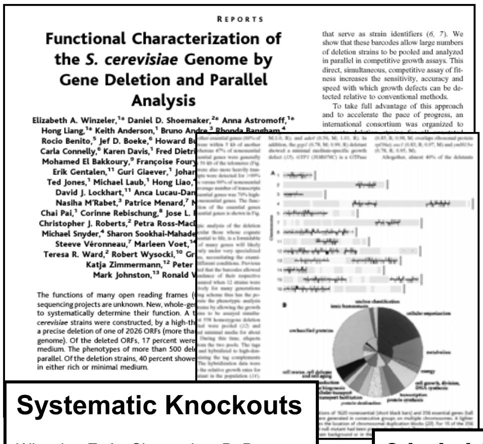
- Affymetrix
  - Oligos
    - – Don't have to know sequence

- 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 N Chai Pai,1 Corinne Rebischung,8 Jose L. Christopher J. Roberts,2 Petra Ross-Macl 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 V

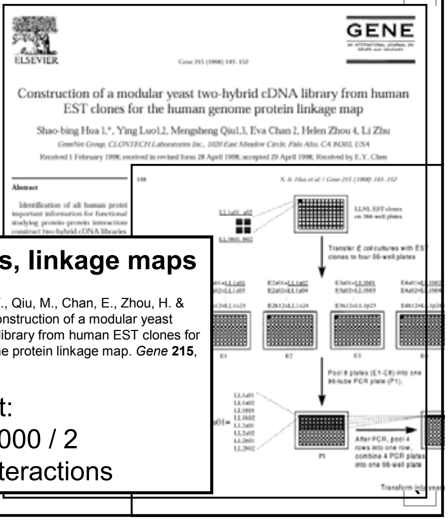
The functions of many open reading frames (sequencing projects are unknown. New, whole-ge to systematically determine their function. A t *cerevisiae* strains were constructed, by a high-th a precise deletion of one of 2026 ORFs (more tha genome). Of the deleted ORFs, 17 percent were medium. The phenotypes 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 215 (1998) 143–152

GENE

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

LLNL EST clones on 384-well plates
Transfer E. coli cultures with EST clones to four 96-well plates
Pool 8 plates (E1-E8) into one 96-tube PCR plate (P1).
After PCR, pool 4 rows into one row, combine 4 PCR plates into one 96-well plate
P1

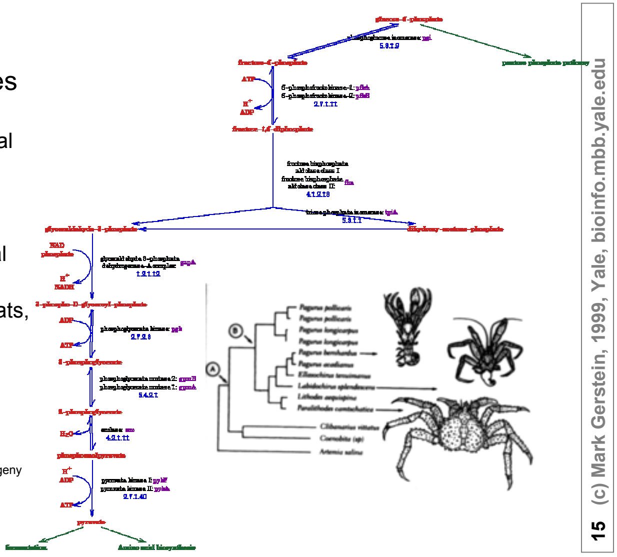
2 hybrids, linkage maps

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

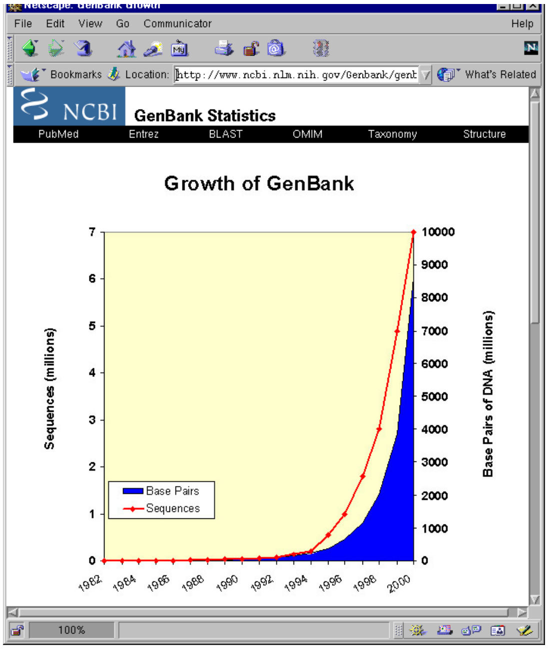
- Information to understand genomes
  - ◊ Metabolic Pathways (glycolysis), traditional biochemistry
  - ◊ Regulatory Networks
  - ◊ Whole Organisms Phylogeny, traditional zoology
  - ◊ Environments, Habitats, ecology
  - ◊ The Literature (MEDLINE)
- The Future....



(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

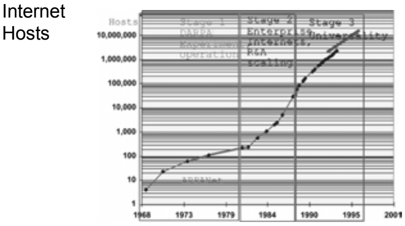
GenBank Data

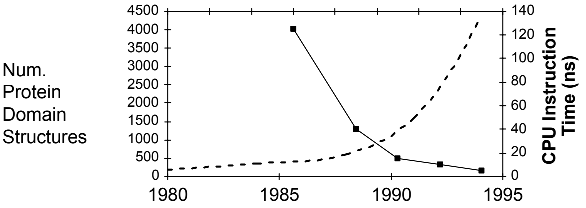
Year	Base Pairs	Sequences
1982	680338	606
1983	2274029	2427
1984	3368765	4175
1985	5204420	5700
1986	9615371	9978
1987	15514776	14584
1988	23800000	20579
1989	34762585	28791
1990	49179285	39533
1991	71947426	55627
1992	101008486	78608
1993	157152442	143492
1994	217102462	215273
1995	384939485	555694
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

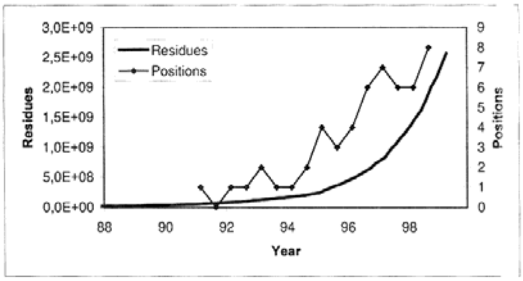
- CPU vs Disk & Net
  - ◊ As important as the increase in computer speed has been, the ability to store large amounts of information on computers is even more crucial
- Driving Force in Bioinformatics

(Internet picture adapted from D Brutlag, Stanford)

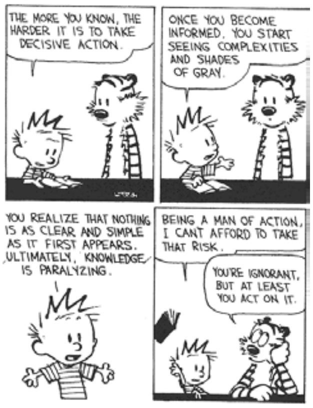




Bioinformatics is born!



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.


THE MORE YOU KNOW, THE HARDER IT IS TO TAKE DECISIVE ACTION.
ONCE YOU BECOME INFORMED, YOU START SEEING COMPLEXITIES AND SHADES OF GRAY.
YOU REALIZE THAT NOTHING IS AS CLEAR AND SIMPLE AS IT FIRST APPEARS. ULTIMATELY, KNOWLEDGE IS PARALYZING.
BEING A MAN OF ACTION, I CANT AFFORD TO TAKE THAT RISK.
YOU'RE IGNORANT, BUT AT LEAST YOU ACT ON IT.

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

(courtesy of Finn Drablos)

Weber
Cartoon

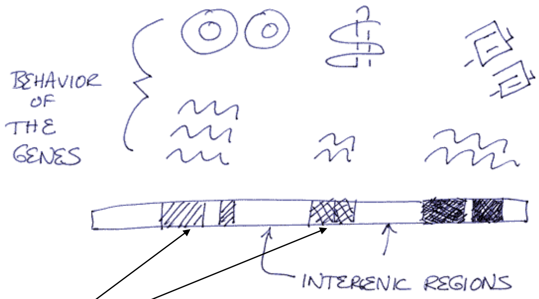
"Don't just sit there! If you've processed all the data there is, go out and find more data!"

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

Comprehensive Understanding of Gene Function on a Genomic Scale

The Next Step after the sequence:

Proteomics Expression Analysis Structural Genomics, Protein Interactions

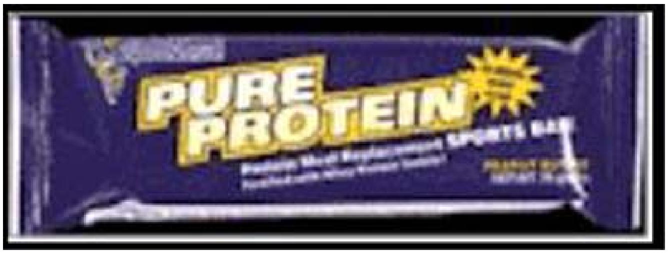


Step 1: The genome sequence and genes

Pseudogenes, Regulatory Regions, Repeats

Evolutionary Implications of Intergenic Regions as Gene Graveyard

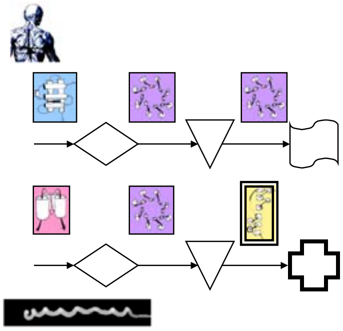
The next step: proteomics



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

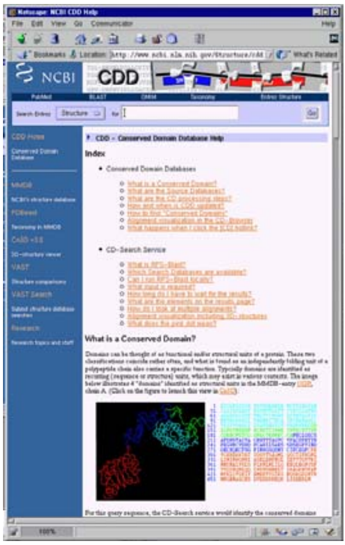


- Different Sequences Have the Same Structure
- Organism has many similar genes
- Single Gene May Have Multiple Functions
- Genes are grouped into Pathways
- Genomic Sequence Redundancy due to the Genetic Code
- **How do we find the similarities?.....**

Core

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
Cn3D v3.0
3D-structure viewer
VAST
Structure comparisons
VAST Search
Submit structure database searches
Research
Research topics and staff
100%

CDD - Conserved Domain Database Help

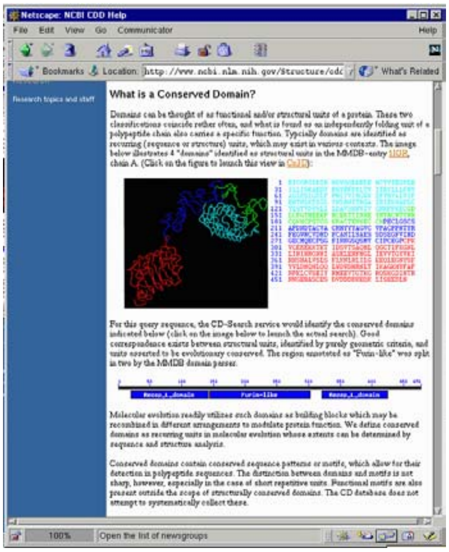
Index

- 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] button?
- 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. Typcially domains are identified as recurring (sequence or structure) units, which may exist in various contexts. The image below illustrates 4 "domains" identified as structural units in the MMDB-entry 1IGR, chain A. (Click on the figure to launch this view in Cn3D):

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


Netscape: NCBI CDD Help
File Edit View Go Communicator Help
Bookmarks Location: http://www.ncbi.nlm.nih.gov/Structure/cdd What's Related
Research topics and staff
Recep_L_domain
Furin-like
Recep_L_domain
Open the list of newsgroups
100%

What is a Conserved Domain?

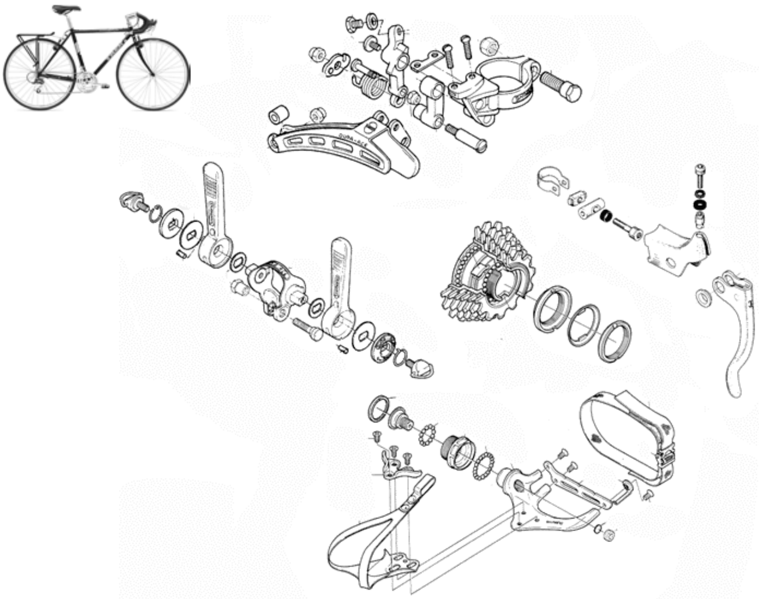
Domains can be thought of as functional and/or structural units of a protein. These two classifications coincide rather often, and what is found as an independently folding unit of a polypeptide chain also carries a specific function. Typcially domains are identified as recurring (sequence or structure)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

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

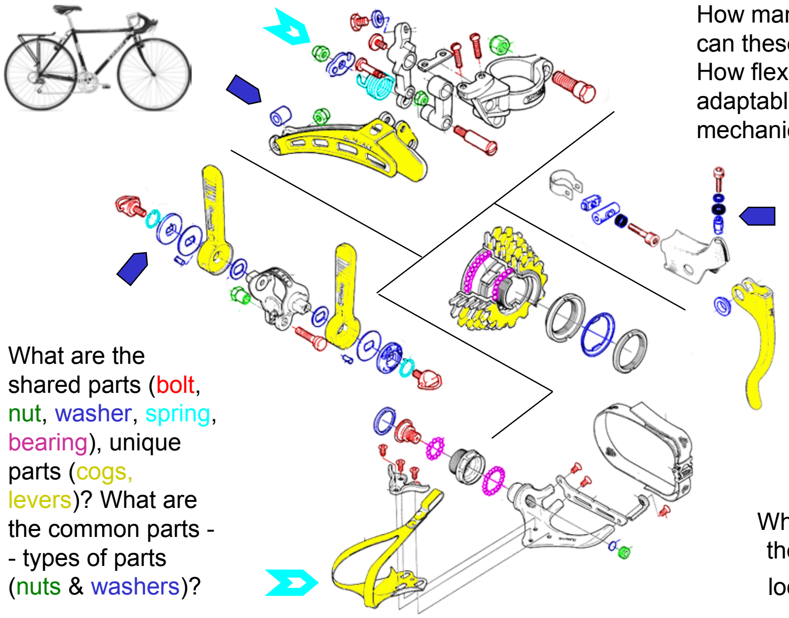
Molecular evolution readily utilizes such domains as building blocks which may be recombined in different arrangements to modulate protein function. We define conserved domains as recurring units in molecular evolution whose extents can be determined by sequence and structure analysis.

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

A Parts List Approach to Bike Maintenance



A Parts List Approach to Bike Maintenance

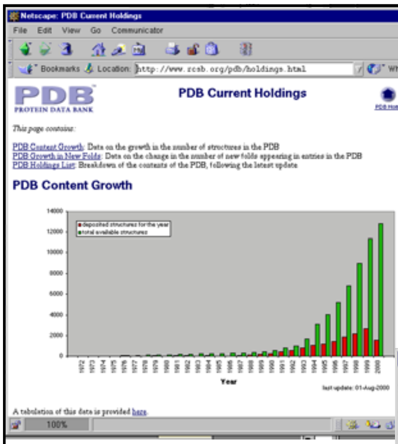


How many roles can these play? How flexible and adaptable are they mechanically?

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 Fundamentally New (Fold) Parts Not Increasing that Fast



Netscape: PDB Current Holdings
File Edit View Go Communicator Help
Bookmarks Location: http://www.rcsb.org/pdb/holdings.html What's Related
PDB
PROTEIN DATA BANK
PDB Home Contact us
100%

PDB Current Holdings

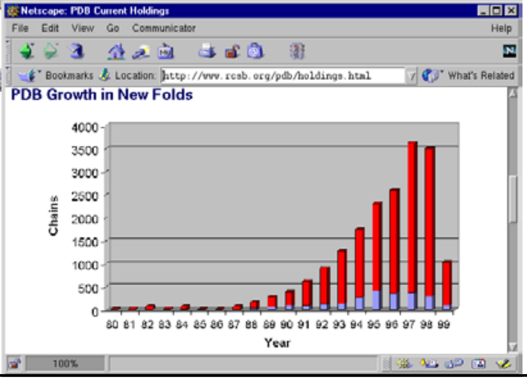
This page contains:

PDB Content Growth: Data on the growth in the number of structures in the PDB
PDB Growth in New Folds: Data on the change in the number of new folds appearing in entries in the PDB
PDB Holdings List: Breakdown of the contents of the PDB, following the latest update

PDB Content Growth

14000
12000
10000
8000
6000
4000
2000
0
deposited structures for the year
total available structures
Year
last update: 01-Aug-2000

A tabulation of this data is provided here.



Netscape: PDB Current Holdings
File Edit View Go Communicator Help
Bookmarks Location: http://www.rcsb.org/pdb/holdings.html What's Related
100%

PDB Growth in New Folds

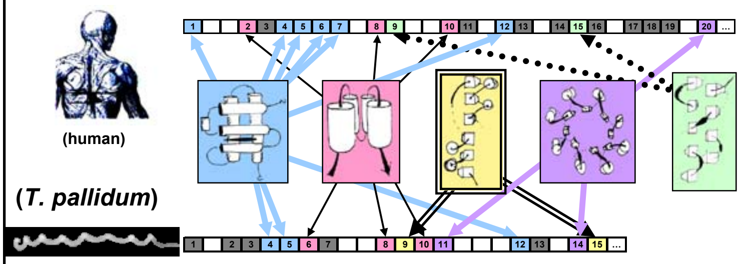
4000
3500
3000
2500
2000
1500
1000
500
0
Chains
80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99
Year

Total in Databank 🟩

New Submissions 🟥

New Folds 🟦

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



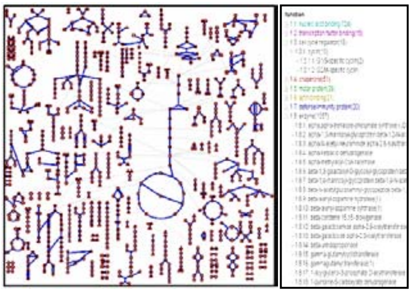
~100000 genes

~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?

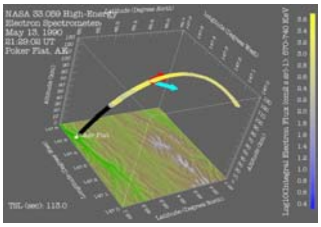
- *(Molecular)* **Bio** - `informatics`
- One idea for a definition?
  Bioinformatics is conceptualizing **biology in terms of molecules** (in the sense of physical-chemistry) and then applying **"informatics" techniques** (derived from disciplines such as applied math, CS, and statistics) to understand 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

- 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 (Visision, recognition)
- Physical Simulation
  - ◊ Newtonian Mechanics
  - ◊ Electrostatics
  - ◊ Numerical Algorithms
  - ◊ Simulation

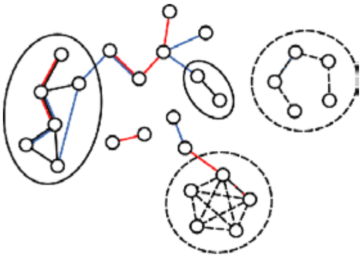
Bioinformatics as New Paradigm for Scientific Computing

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



Core

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



Statistical Physics vs. Classical Physics

Bioinformatics, Genomic Surveys

Vs.

Chemical Understanding, Mechanism, Molecular Biology

End of class 2002,09.09

(Bioinfo-1)

[next class joins intro & seqs.]

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

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

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

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

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

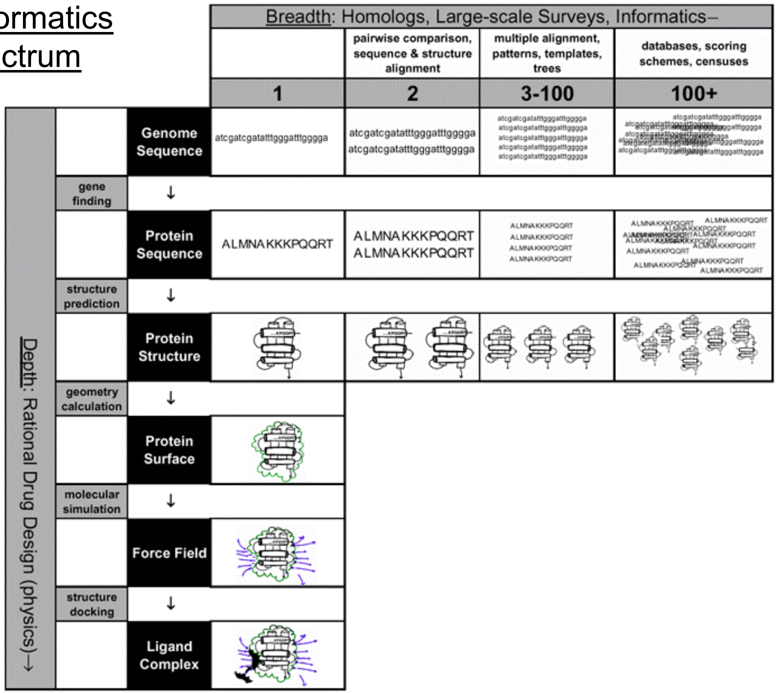
Topics -- Genomics

- 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 Spectrum



Breadth: Homologs, Large-scale Surveys, Informatics–

pairwise comparison, sequence & structure alignment

multiple alignment, patterns, templates, trees

databases, scoring schemes, censuses

1

2

3-100

100+

Depth: Rational Drug Design (physics)→

Genome Sequence

gene finding ↓

Protein Sequence

structure prediction ↓

Protein Structure

geometry calculation ↓

Protein Surface

molecular simulation ↓

Force Field

structure docking ↓

Ligand Complex

atcgatcgatatttgggatttgggga
atcgatcgatatttgggatttgggga
atcgatcgatatttgggatttgggga
ALMNAKKKPQQRT
ALMNAKKKPQQRT
ALMNAKKKPQQRT

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?

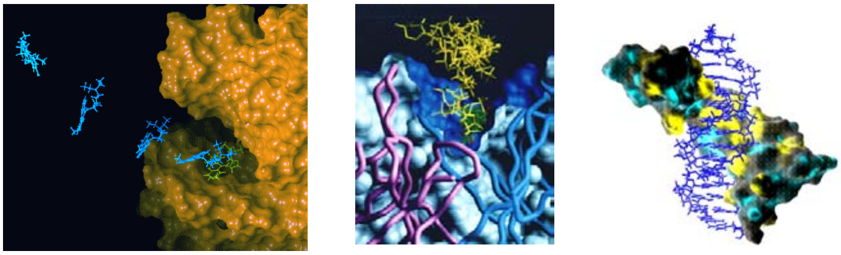
- *(Molecular)* **Bio** - **`informatics`**
- One idea for a definition?
  Bioinformatics is conceptualizing **biology in terms of molecules** (in the sense of physical-chemistry) and then applying **"informatics" techniques** (derived from disciplines such as applied math, CS, and statistics) to understand 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

Core

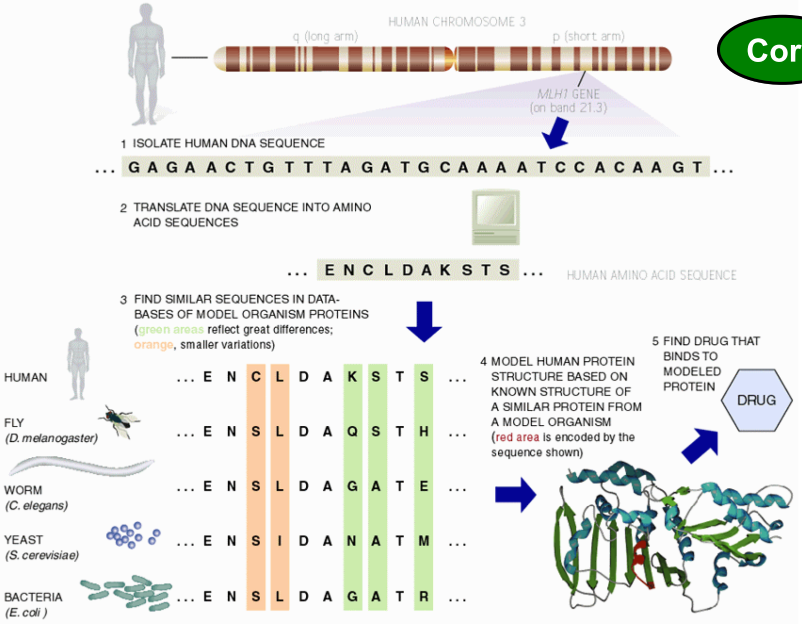
- Understanding How Structures Bind Other Molecules (Function)
- Designing Inhibitors
- Docking, Structure Modeling

(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

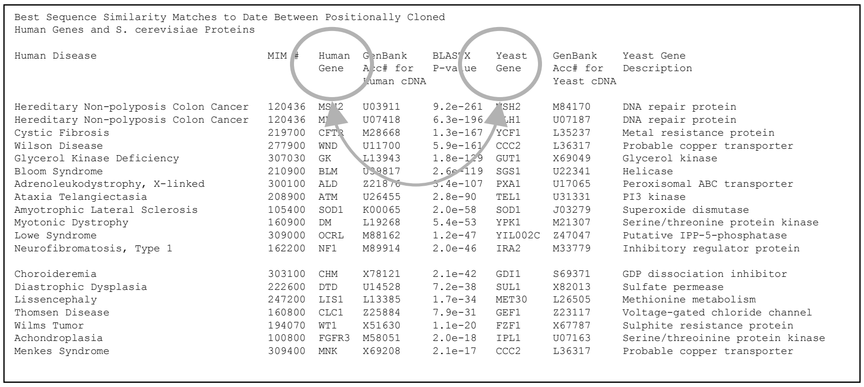
Core


HUMAN CHROMOSOME 3
q (long arm)
p (short arm)
MLH1 GENE
(on band 21.3)
1 ISOLATE HUMAN DNA SEQUENCE
... G A G A A C T G T T T A G A T G C A A A A T C C A C A A G T ...
2 TRANSLATE DNA SEQUENCE INTO AMINO ACID SEQUENCES
... E N C L D A K S T S ...
HUMAN AMINO ACID SEQUENCE
3 FIND SIMILAR SEQUENCES IN DATABASES OF MODEL ORGANISM PROTEINS (green areas reflect great differences; orange, smaller variations)
HUMAN
... E N C L D A K S T S ...
FLY (D. melanogaster)
... E N S L D A Q S T H ...
WORM (C. elegans)
... E N S L D A G A T E ...
YEAST (S. cerevisiae)
... E N S I D A N A T M ...
BACTERIA (E. coli)
... E N S L D A G A T R ...
4 MODEL HUMAN PROTEIN STRUCTURE BASED ON KNOWN STRUCTURE OF A SIMILAR PROTEIN FROM A MODEL ORGANISM (red area is encoded by the sequence shown)
5 FIND DRUG THAT BINDS TO MODELED PROTEIN
DRUG

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.)

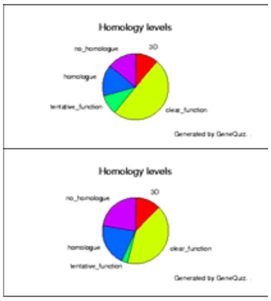
- Cross-Referencing, one thing to another thing
- Sequence Comparison and Scoring
- Analogous Problems for Structure Comparison
- Comparison has two parts:
  - (1) Optimally **Aligning** 2 entities to get a Comparison **Score**
  - (2) Assessing **Significance** of this score in a given **Context**

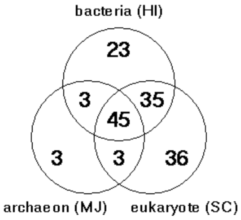
- **Integrated Presentation**
  - ◊ Align Sequences
  - ◊ Align Structures
  - ◊ Score in a Uniform Framework

Major Application I|I:
Overall Genome Characterization

Core

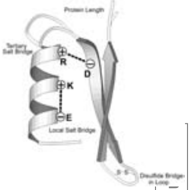
- Overall Occurrence of a Certain Feature in the Genome
  - ◊ e.g. how many kinases in Yeast
- Compare Organisms and Tissues
  - ◊ Expression levels in Cancerous vs Normal Tissues
- Databases, Statistics



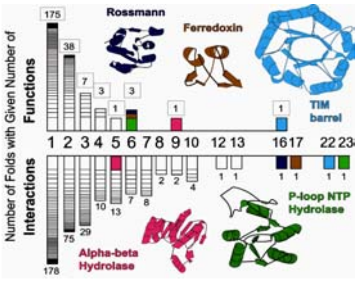


(Clock figures, yeast v. Synechocystis,
adapted from GeneQuiz Web Page, Sander Group, EBI)

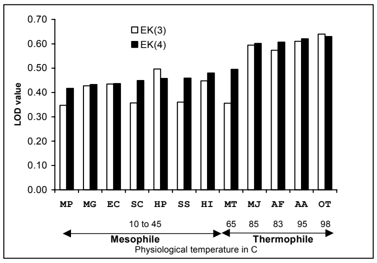
What do you get from large-scale datamining? Global statistics on the population of proteins



EX-1: Occurrence of functions per fold & interactions per fold over all genomes

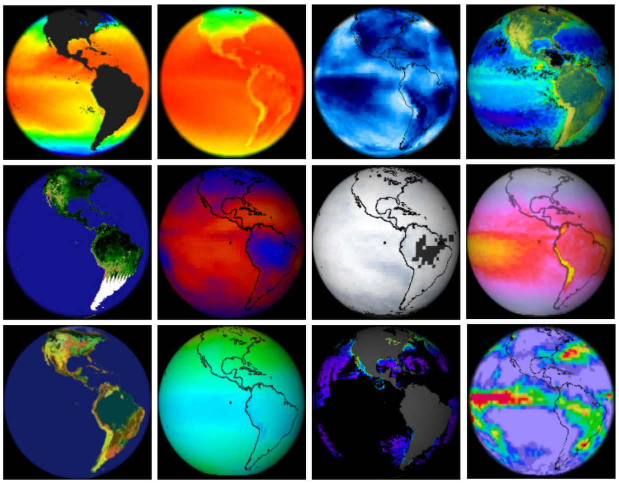


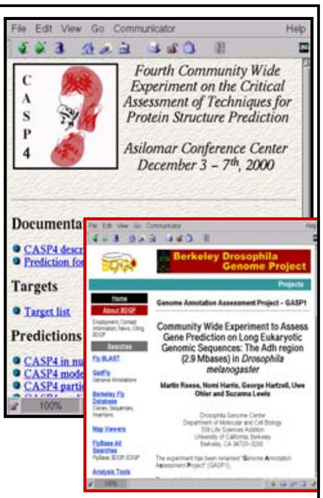
EX-2: Occurrence of 1-4 salt bridges in genomes of thermophiles v mesophiles



Integrative Genomic Surveys of Many Proteins from Many Perspectives

vs




File Edit View Go Communicator Help
C
A
S
P
4
100%
File Edit View Go Communicator Help
BDGP
Projects
Home
About BDGP
Searches

Fourth Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction

Asilomar Conference Center
December 3 – 7th, 2000

Documenta

- CASP4 descr
- Prediction fo

Targets

- Target list

Predictions

- CASP4 in nu
- CASP4 mode
- CASP4 parti

Berkeley Drosophila Genome Project

Employment, Contact Information, News, Citing BDGP

Fly BLAST

GadFly
Genome Annotations

Berkeley Fly Database
Clones, Sequences, Insertions

Map Viewers

FlyBase All Searches
FlyBase, BDGP, EDGP

Analysis Tools

Genome Annotation Assessment Project - GASP1

Community Wide Experiment to Assess Gene Prediction on Long Eukaryotic Genomic Sequences: The Adh region (2.9 Mbases) in *Drosophila melanogaster*

Martin Reese, Nomi Harris, George Hartzell, Uwe Ohler and Suzanna Lewis

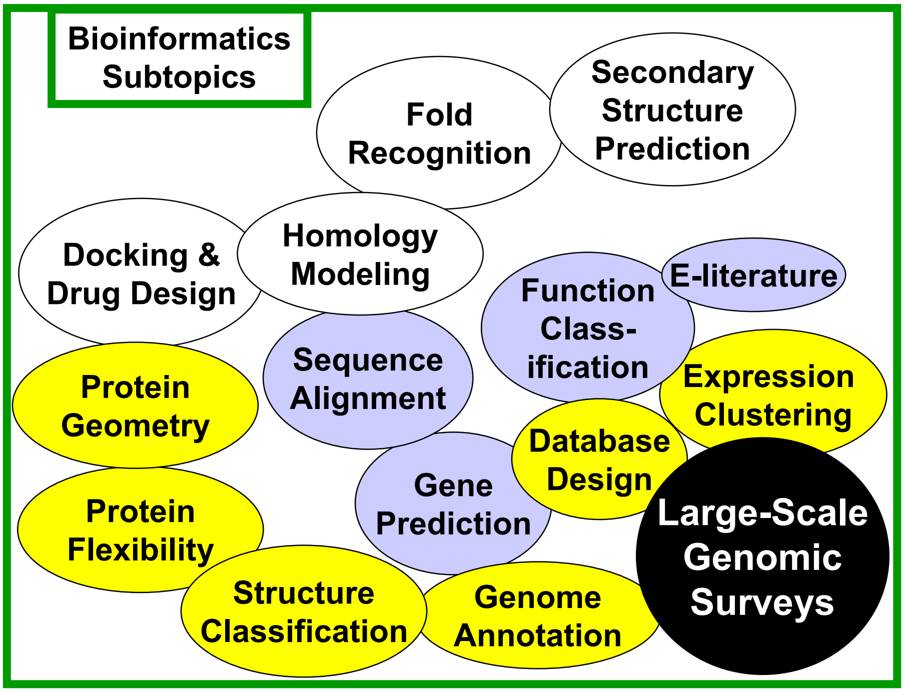
Drosophila Genome Center
Department of Molecular and Cell Biology
539 Life Sciences Addition
University of California, Berkeley
Berkeley, CA 94720-3200

The experiment has been renamed "Genome Annotation Assessment Project" (GASP1).

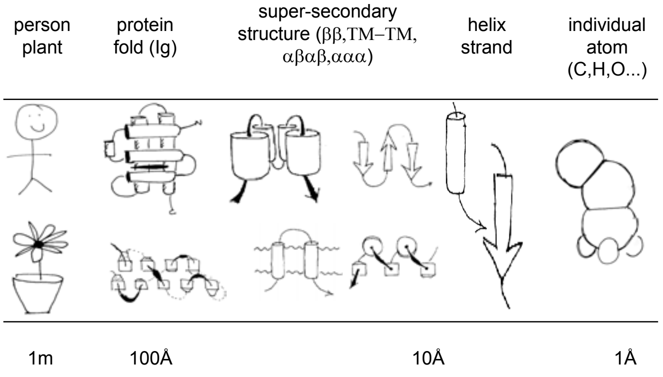
"Prediction" Bioinformatics

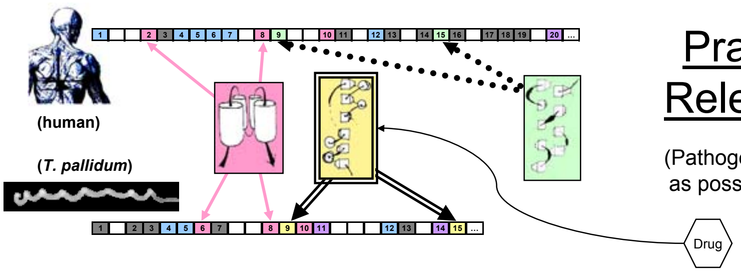
(focused on individual genes and structures)

Bioinformatics Subtopics



At What Structural Resolution Are Organisms Different?





Practical Relevance

(Pathogen only folds as possible targets)