CENG 465
Introduction to Bioinformatics

Spring 2008-2009

Tolga Can (Office: B-109)
e-mail: tcan@ceng.metu.edu.tr

Course Web Page:
http://www.ceng.metu.edu.tr/~tcan/ceng465_s0809/

Teaching Assistant

- Dr. Ahmet Saçan
- e-mail: ahmet@ceng.metu.edu.tr
- Office: A-206

Goals of the course

- Working at the interface of computer science and biology
  - New motivation
  - New data and new demands
  - Real impact
- Introduction to main issues in computational biology
- Opportunity to interact with algorithms, tools, data in current practice

High level overview of the course

- A general introduction
  - what problems are people working on?
  - how people solve these problems?
  - what key computational techniques are needed?
  - how much help computing has provided to biological research?

- A way of thinking -- tackling "biological problems" computationally
  - how to look at a "biological problem" from a computational point of view?
  - how to formulate a computational problem to address a biological issue?
  - how to collect statistics from biological data?
  - how to build a "computational" model?
  - how to solve a computational modeling problem?
  - how to test and evaluate a computational algorithm?

Course outline

- Motivation and introduction to biology (1 week)
- Sequence analysis (4 weeks)
  - Analyze DNA and protein sequences for clues regarding function
  - Identification of homologues
    - Pairwise sequence alignment
  - Statistical significance of sequence alignments
  - Sequence Motifs
  - Suffix trees
  - Multiple sequence alignment
- Phylogenetic trees, clustering methods (1 week)

Course outline

- Protein structures (4 weeks)
  - Analyze protein structures for clues regarding function
    - Structure alignment
  - Structure prediction (secondary, tertiary)
  - Structural motifs, active sites, docking
  - Multiple structural alignment, geometric hashing
- Microarray data analysis (2 weeks)
  - Correlations, clustering
  - Inference of function
- Gene/Protein networks, pathways (2 weeks)
  - Protein-protein, protein/DNA interactions
  - Construction and analysis of large scale networks

Grading

- Midterm exam - 25%
- Final exam - 35%
- Written assignments - 20%
- Programming assignments - 20%

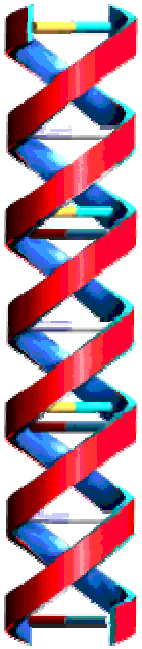
Miscellaneous

- Course webpage
  - http://www.ceng.metu.edu.tr/~tcan/ceng465_s0809/
  - Lecture slides and reading materials
  - Assignments
  - Other relevant information
- Newsgroup
  - metu.ceng.course.465
  - You should follow the newsgroup for course related announcements
  - Students from other departments should get a CENG account for this semester (Room: A-210) in order to access the newsgroup

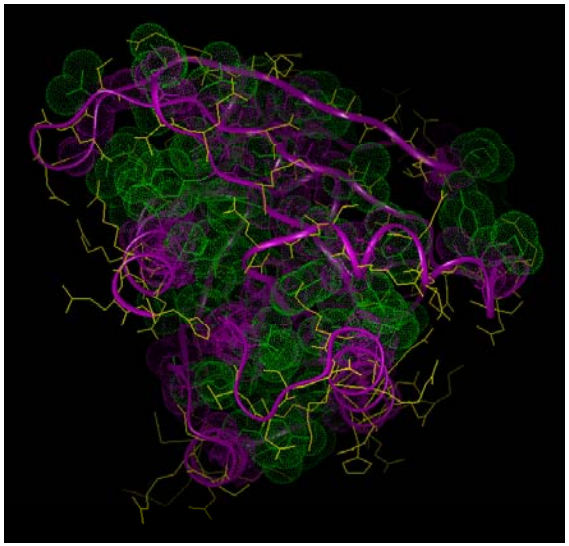
What is Bioinformatics?

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

Introductory Biology



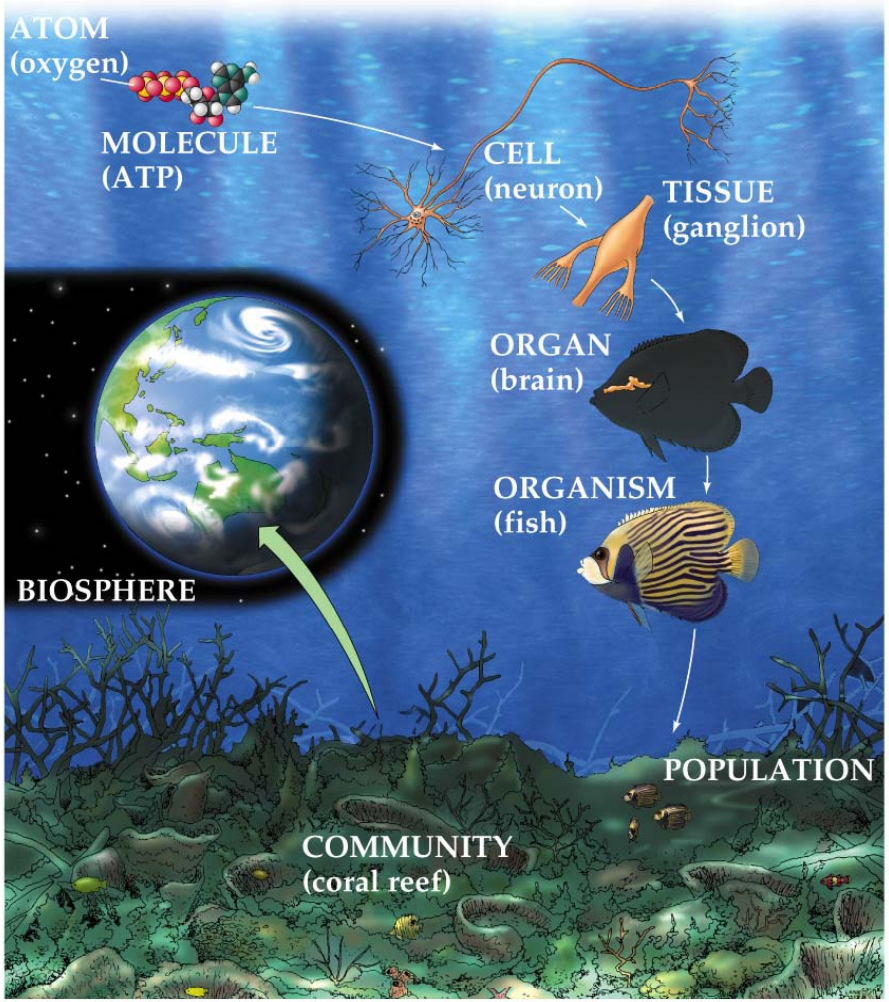
DNA
(Genotype)



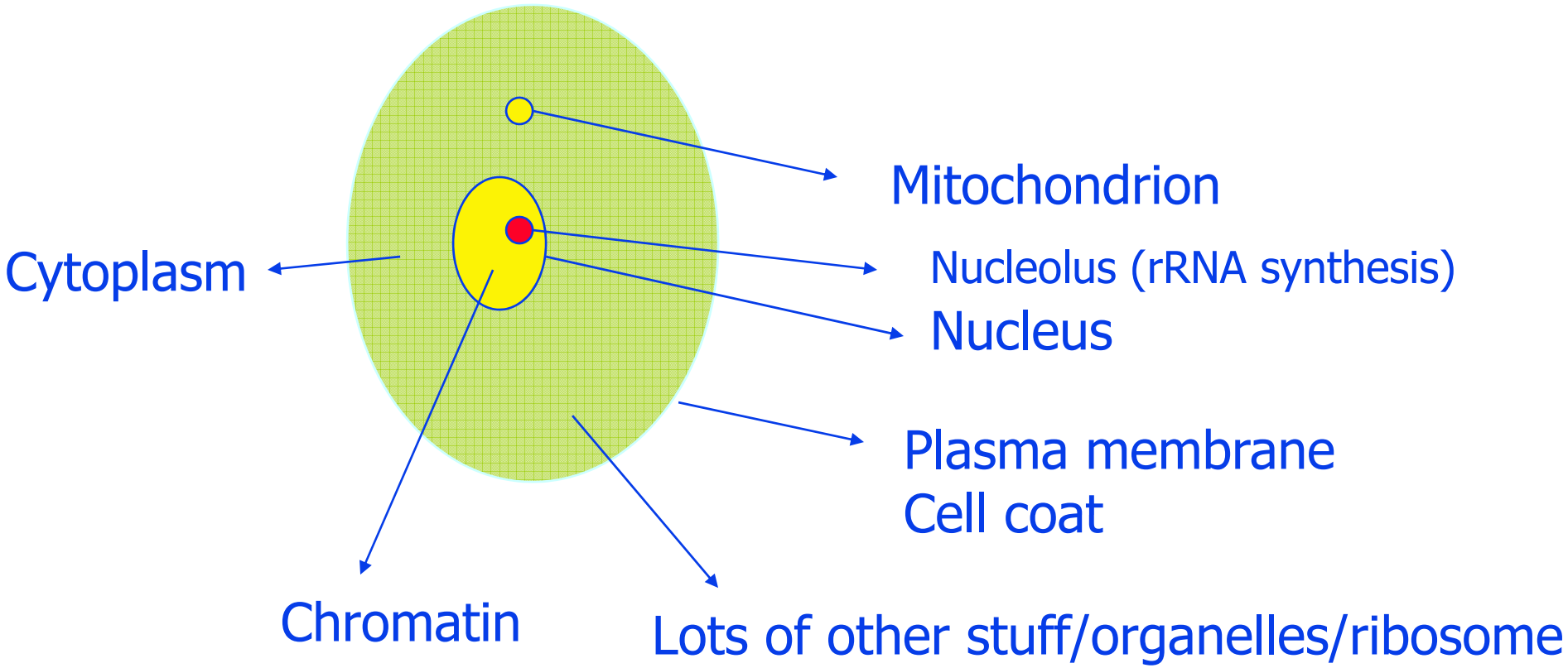
Protein

Phenotype

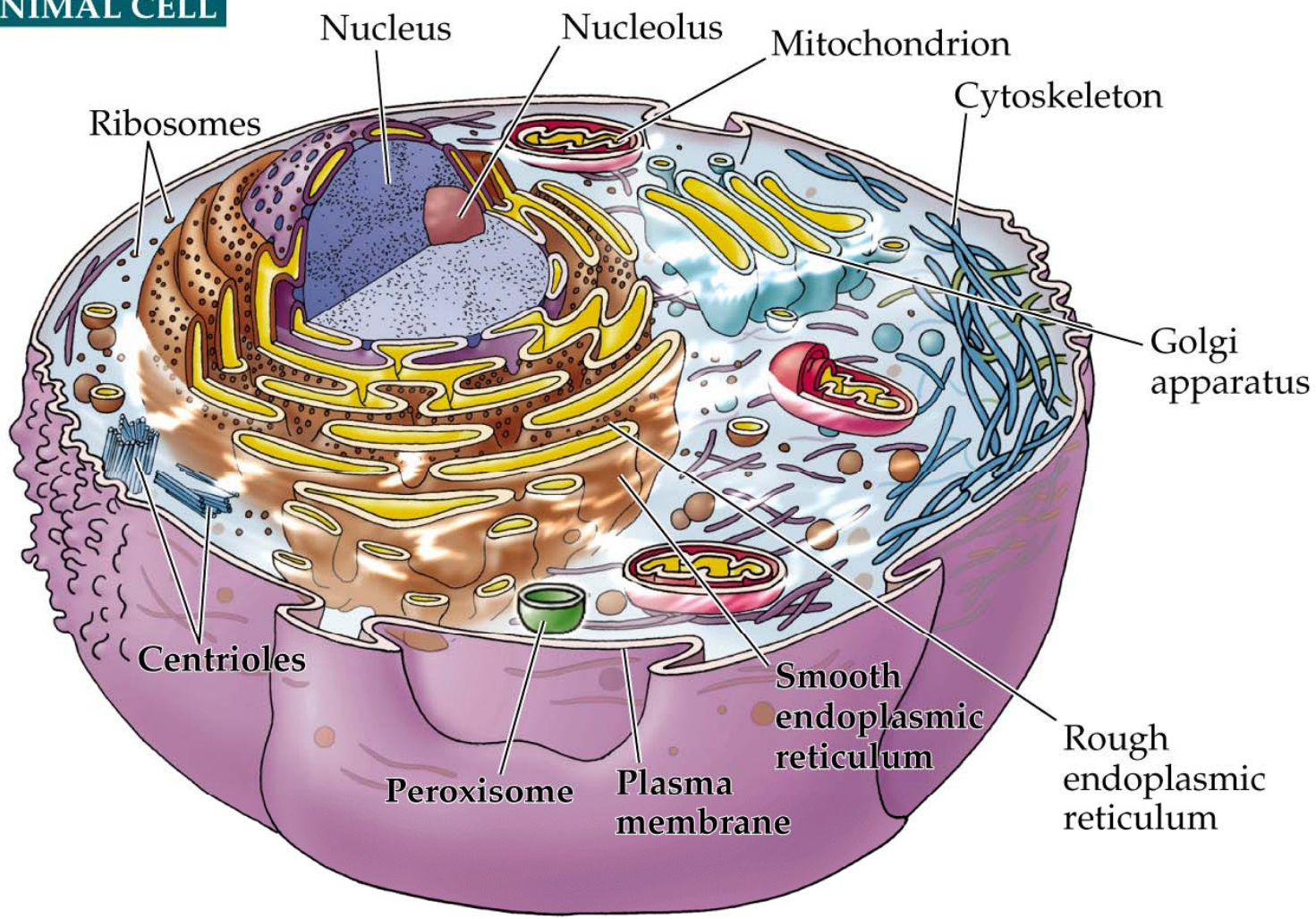
Scales of life



Animal Cell


Mitochondrion
Cytoplasm
Nucleolus (rRNA synthesis)
Nucleus
Plasma membrane
Cell coat
Chromatin
Lots of other stuff/organelles/ribosome

AN ANIMAL CELL



Two kinds of Cells

- Prokaryotes – no nucleus (bacteria)
  - Their genomes are circular
- Eukaryotes – have nucleus (animal,plants)
  - Linear genomes with multiple chromosomes in pairs. When pairing up, they look like

Middle: centromere
Top: p-arm
Bottom: q-arm

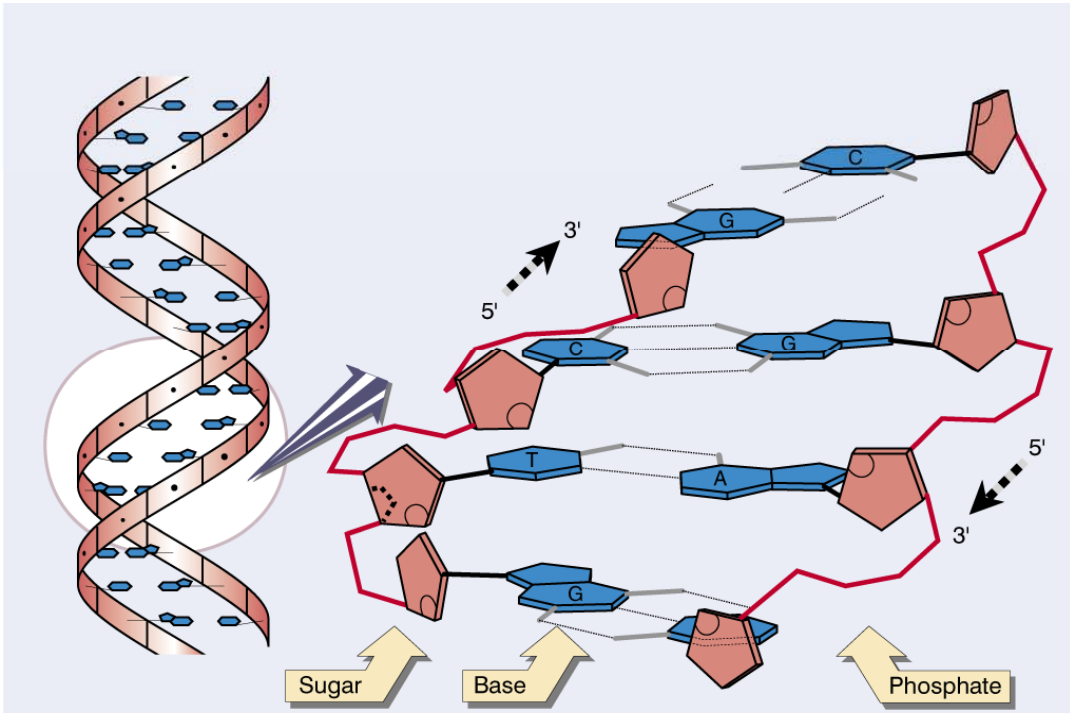
Molecular Biology Information - DNA

- Raw DNA Sequence
  - Coding or Not?
  - Parse into genes?
  - 4 bases: AGCT
  - ~1 Kb in a gene, ~2 Mb in genome
  - ~3 Gb Human

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


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

DNA structure

Figure 1.7 Flat base pairs lie perpendicular to the sugar-phosphate backbone.



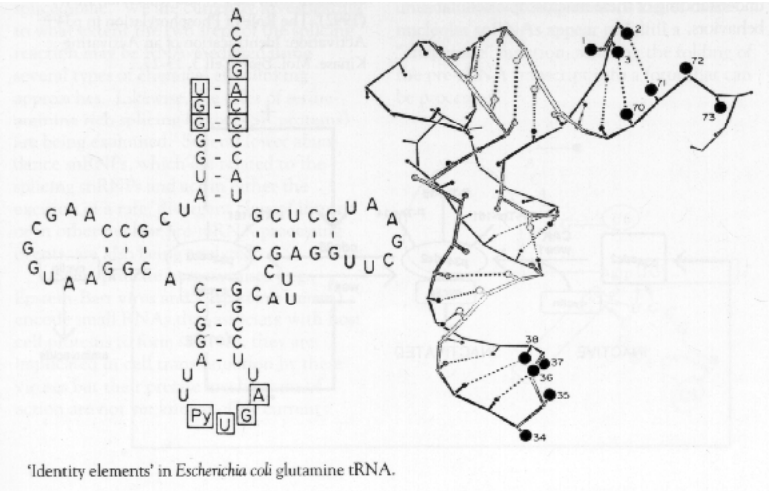
Molecular Biology Information: Protein Sequence

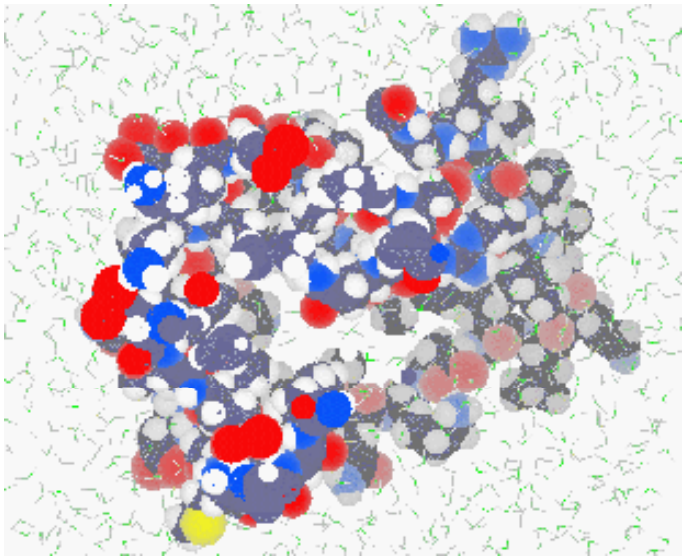
- 20 letter alphabet
  - ACDEFGHIKLMNPQRSTVWY but not BJOUXZ
- Strings of ~300 aa in an average protein (in bacteria),
  ~200 aa in a domain
- ~1M known protein sequences

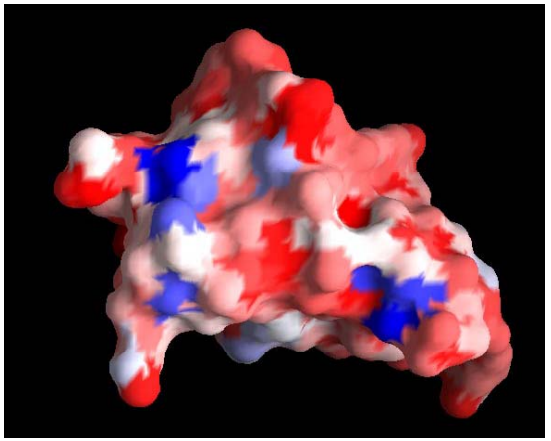
```
d1dhfa_ LNCIVAVSQNMGIGKNGDLPWPPLRNEFRYFQRMTTTSSVEGKQ-NLVIMGKKTWFSI
d8dfr__ LNSIVAVCQNMGIGKDGNLPWPPLRNEYKYFQRMTSTSHVEGKQ-NAVIMGKKTWFSI
d4dfra_ ISLIAALAVDRVIGMENAMPWN-LPADLAWFKRNTL--------NKPVIMGRHTWESI
d3dfr__ TAFLWAQDRDGLIGKDGHLPWH-LPDDLHYFRAQTV--------GKIMVVGRRTYESF

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

Molecular Biology Information: Macromolecular Structure

- DNA/RNA/Protein
  - Almost all protein

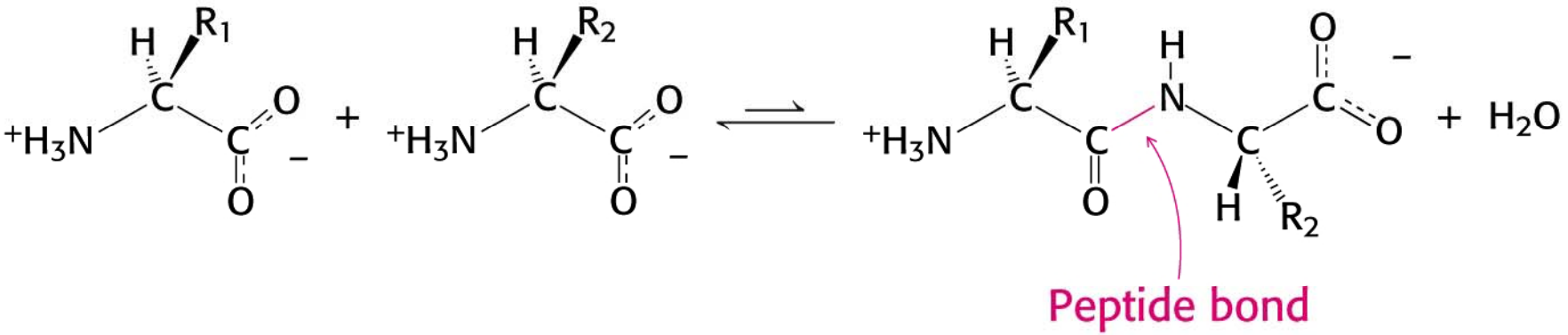






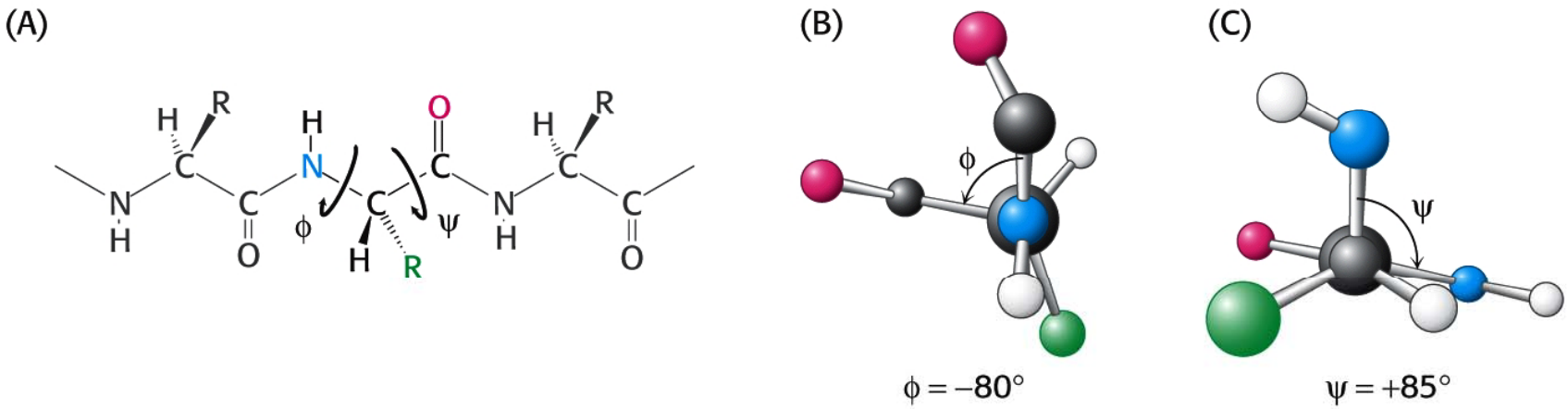
More on Macromolecular Structure

- Primary structure of proteins
  - Linear polymers linked by peptide bonds
  - Sense of direction

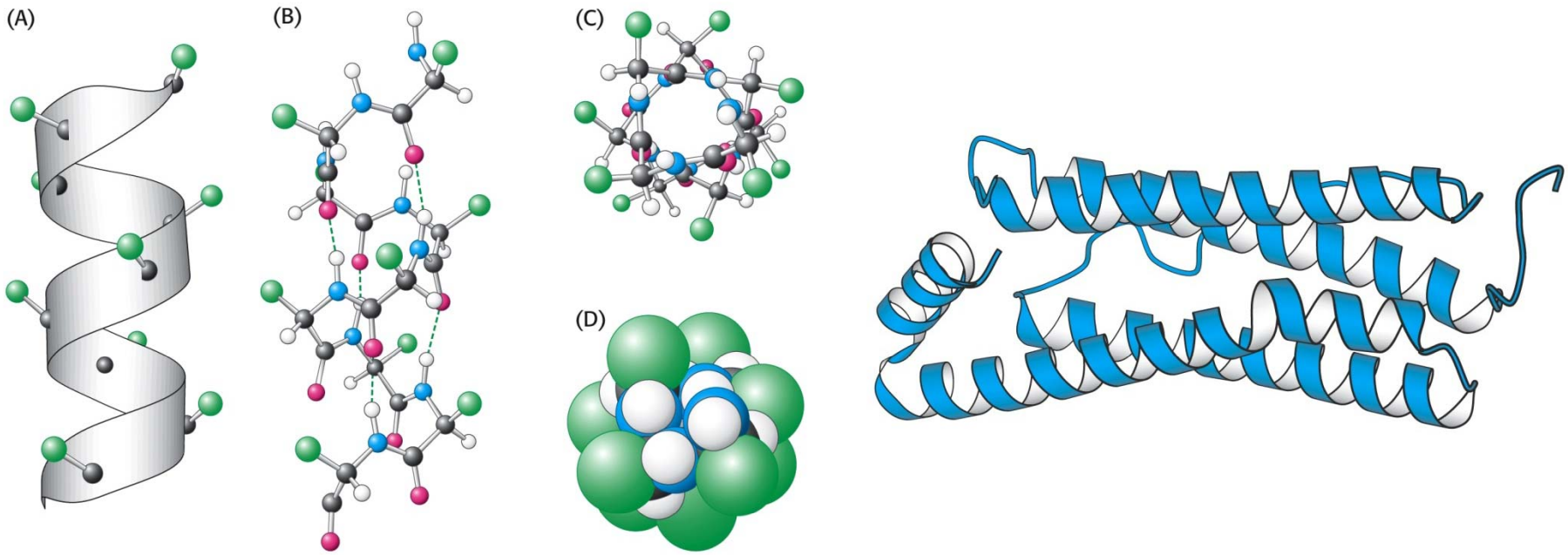


Secondary Structure

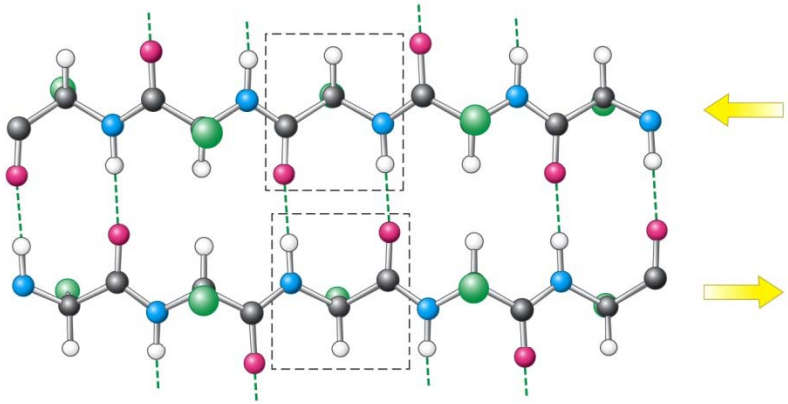
- Polypeptide chains fold into regular local structures
  - alpha helix, beta sheet, turn, loop
  - based on energy considerations
  - Ramachandran plots



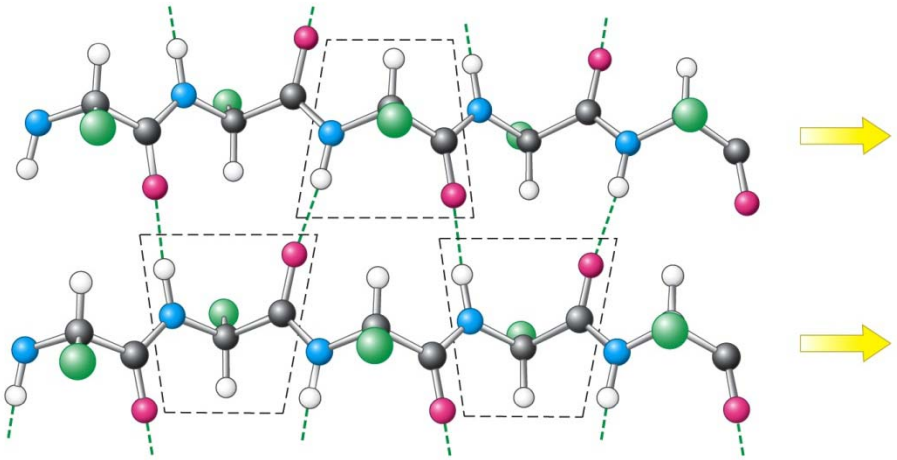
Alpha helix



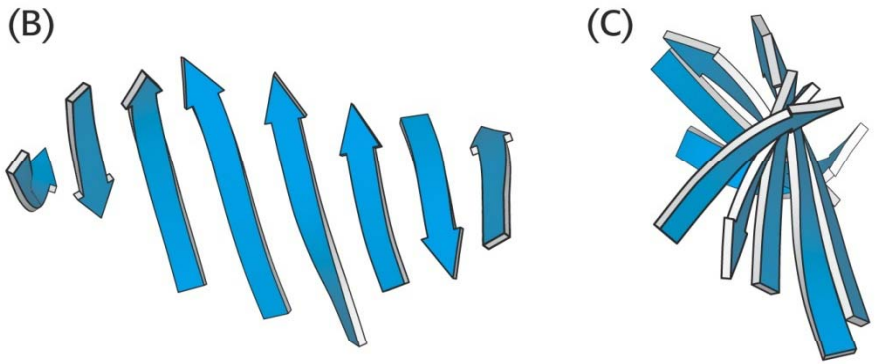
Beta sheet



anti-parallel



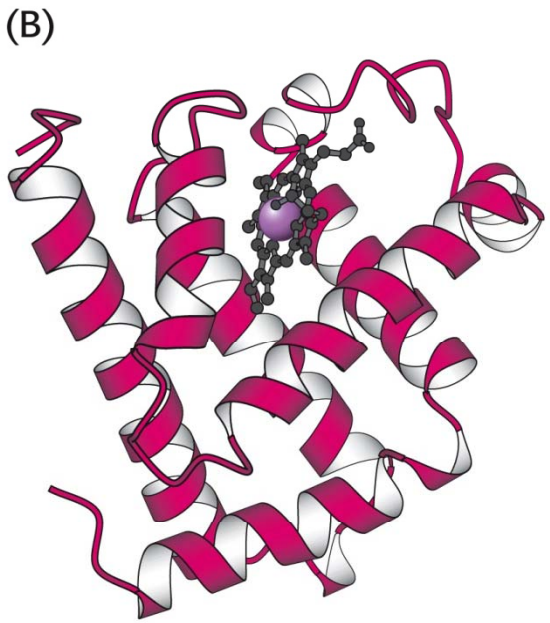
parallel



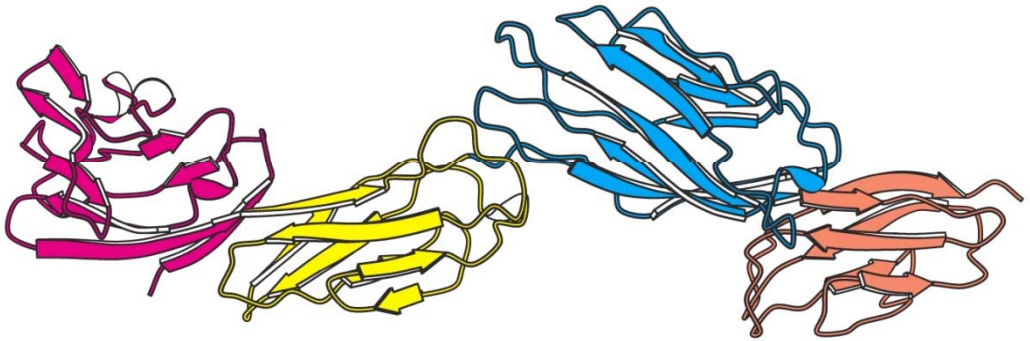
schematic

Tertiary Structure

- 3-d structure of a polypeptide sequence
  - interactions between non-local and foreign atoms
  - often separated into domains



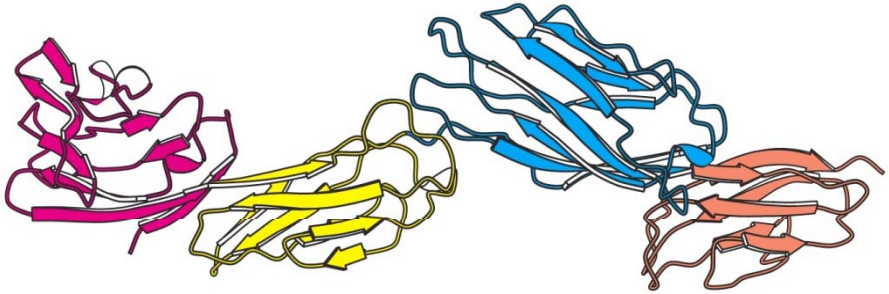
tertiary structure of myoglobin



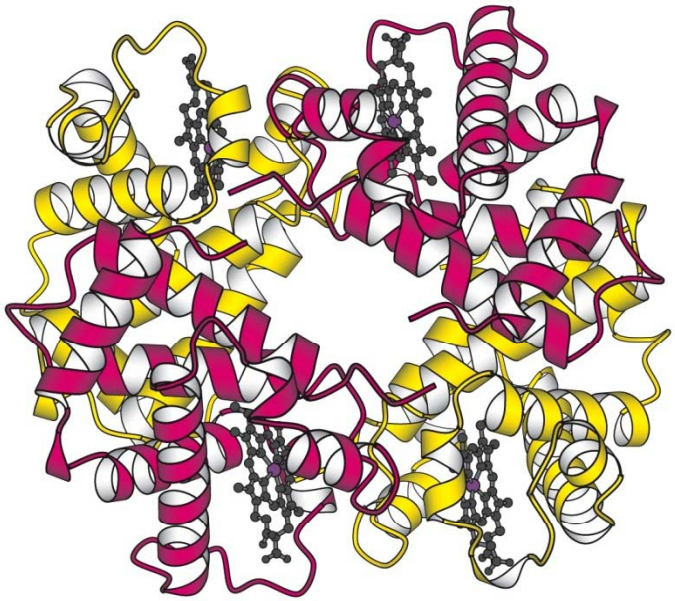
domains of CD4

Quaternary Structure

- Arrangement of protein subunits
  - dimers, tetramers



quaternary structure
of Cro



human hemoglobin
tetramer

Structure summary

- 3-d structure determined by protein sequence
- Cooperative and progressive stabilization
- Prediction remains a challenge
  - ab-initio (energy minimization)
  - knowledge-based
    - Chou-Fasman and GOR methods for SSE prediction
    - Comparative modeling and protein threading for tertiary structure prediction
- Diseases caused by misfolded proteins
  - Mad cow disease
- Classification of protein structures

Genes and Proteins

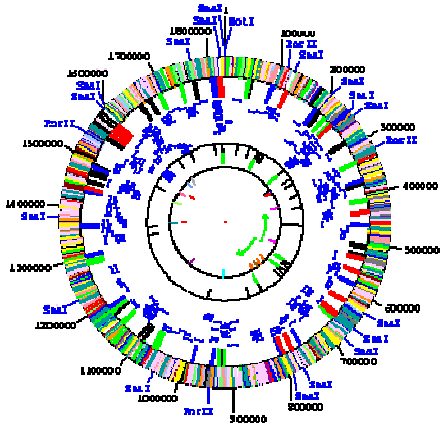
- One gene encodes one* protein.
- Like a program, it starts with start codon (e.g. ATG), then each three code one amino acid. Then a stop codon (e.g. TGA) signifies end of the gene.
- Sometimes, in the middle of a (eukaryotic) gene, there are introns that are spliced out (as junk) during transcription. Good parts are called exons. This is the task of gene finding.

A.A. Coding Table

Glycine (GLY)	GG*
Alanine(ALA)	GC*
Valine (VAL)	GT*
Leucine (LEU)	CT*
Isoleucine (ILE)	AT(*-G)
Serine (SER)	AGT, AGC
Threonine (THR)	AC*
Aspartic Acid (ASP)	GAT,GAC
Glutamic Acid(GLU)	GAA,GAG
Lysine (LYS)	AAA, AAG
Start: ATG, CTG, GTG	

Arginine (ARG)	CG*
Asparagine (ASN)	AAT, AAC
Glutamine (GLN)	CAA, CAG
Cysteine (CYS)	TGT, TGC
Methionine (MET)	ATG
Phenylalanine (PHE)	TTT,TTC
Tyrosine (TYR)	TAT, TAC
Tryptophan (TRP)	TGG
Histidine (HIS)	CAT, CAC
Proline (PRO)	CC*
Stop	TGA, TAA, TAG

Molecular Biology Information: Whole Genomes



Genome sequences 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

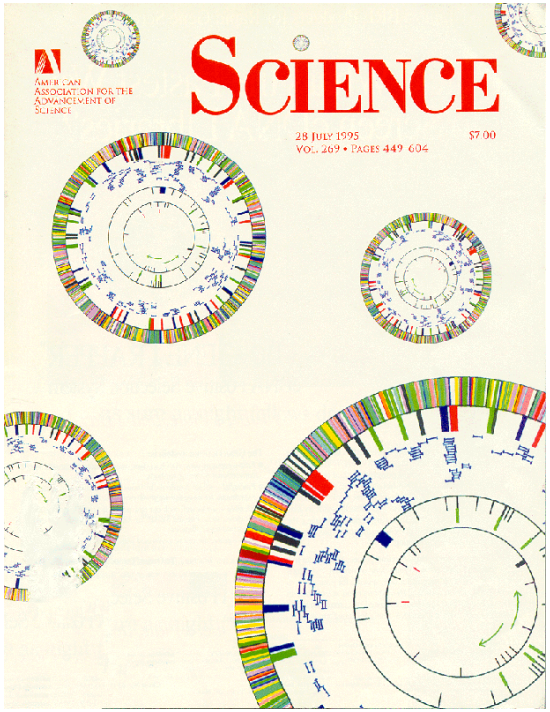
Eukaryote,
13 Mb,
~6K genes
[Nature **387**: 1]

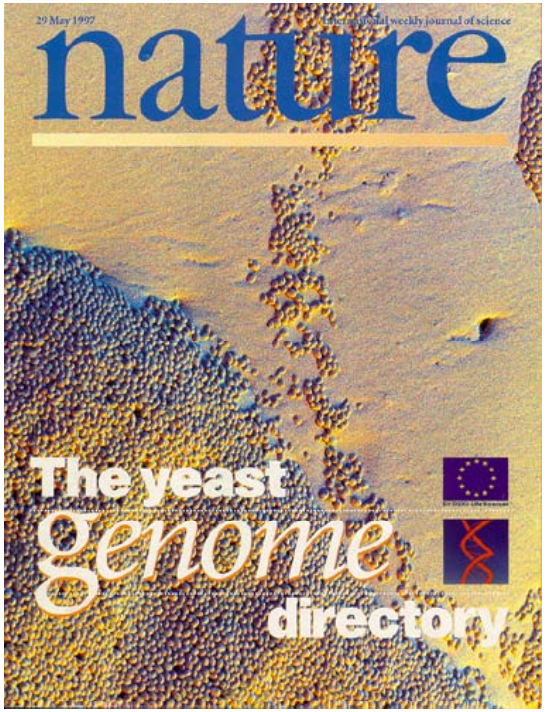
1998

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

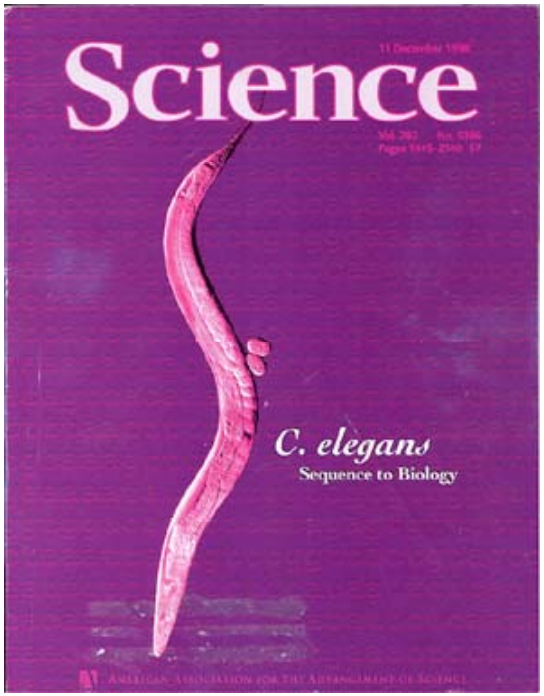
2000?

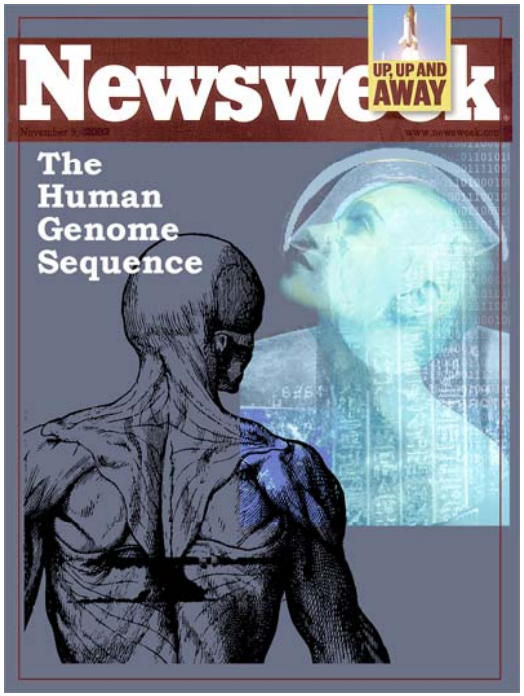
Human,
~3 Gb,
~100K
genes [???]





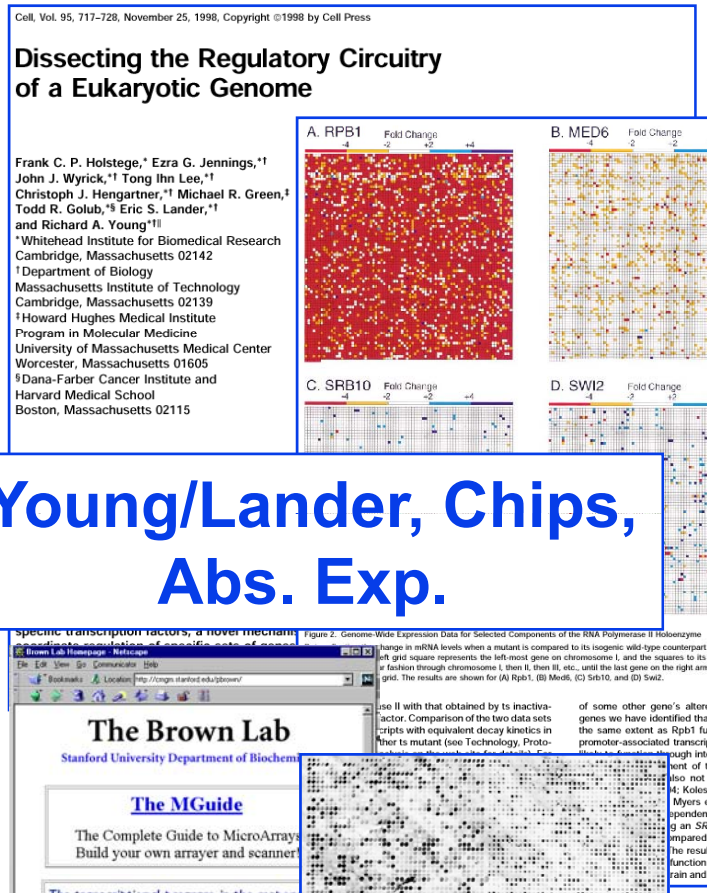
Genomes highlight the Finiteness of the "Parts" in Biology






Human Genome Project
CHEMISTRY BIOLOGY PHYSICS ETHICS ENGINEERING INFORMATICS
RESOURCES FOR THE BIOLOGY CENTURY
Impacting many disciplines
Courtesy
U.S. Department of Energy
Human Genome Program
Global Carbon Cycles
Industrial Resources • Bioremediation
Evolutionary Biology • Biofuels • Agriculture • Forensics
Molecular and Nuclear Medicine • Health Risks
YGA 99-1133R

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

A. RPB1 Fold Change B. MED6 Fold Change C. SRB10 Fold Change D. SWI2 Fold Change

Figure 2. Genome-Wide Expression Data for Selected Components of the RNA Polymerase II Holoenzyme

Young/Lander, Chips, Abs. Exp.

Brown Lab Homepage - Netscape

The Brown Lab

Stanford University Department of Biochemi

The MGuide

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

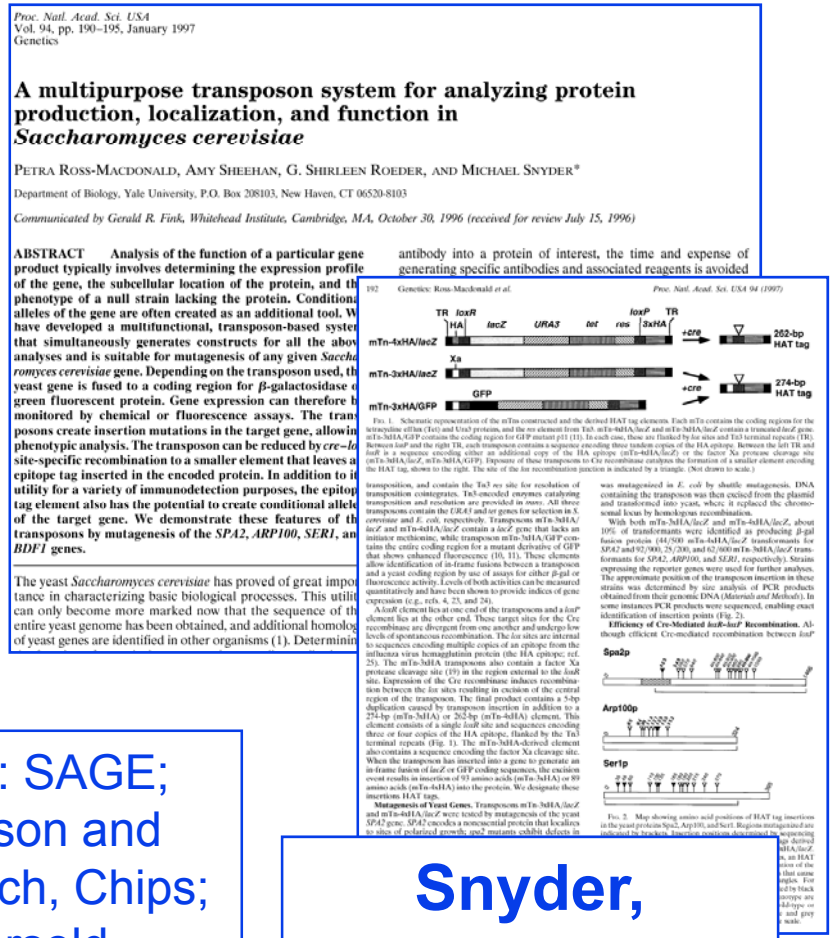
The transcriptional program in the respon... human fibroblasts to serum

The Transcriptional Program of Sporulation in Budding Ye
The Web Companion to the Science Magazine Research Article

Exploring the ... of Gene Ex...
Database

See the entir...

Brown, μarray, Rel. Exp. over Timecourse



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, CT 06520-8103

Communicated by Gerald R. Fink, Whitehead Institute, Cambridge, MA, October 30, 1996 (received for review July 15, 1996)

ABSTRACT **Analysis of the function of a particular gene product typically involves determining the expression profile of the gene, the subcellular location of the protein, and the phenotype of a null strain lacking the protein. Conditional alleles of the gene are often created as an additional tool. We have developed a multifunctional, transposon-based system that simultaneously generates constructs for all the above 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 green fluorescent protein. Gene expression can therefore be monitored by chemical or fluorescence assays. The transposons create insertion mutations in the target gene, allowing phenotypic analysis. The transposon can be reduced by *cre–loxP* site-specific recombination to a smaller element that leaves a epitope tag inserted in the encoded protein. In addition to its utility for a variety of immunodetection purposes, the epitope tag element also has the potential to create conditional alleles of the target gene. We demonstrate these features of the transposons by mutagenesis of the *SPA2*, *ARP100*, *SER1*, and *BDF1* genes.**

antibody into a protein of interest, the time and expense of generating specific antibodies and associated reagents is avoided

The yeast *Saccharomyces cerevisiae* has proved of great importance in characterizing basic biological processes. This utility can only become more marked now that the sequence of the entire yeast genome has been obtained, and additional homologs of yeast genes are identified in other organisms (1). Determinin...

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

Snyder, Transposons, Protein Exp.

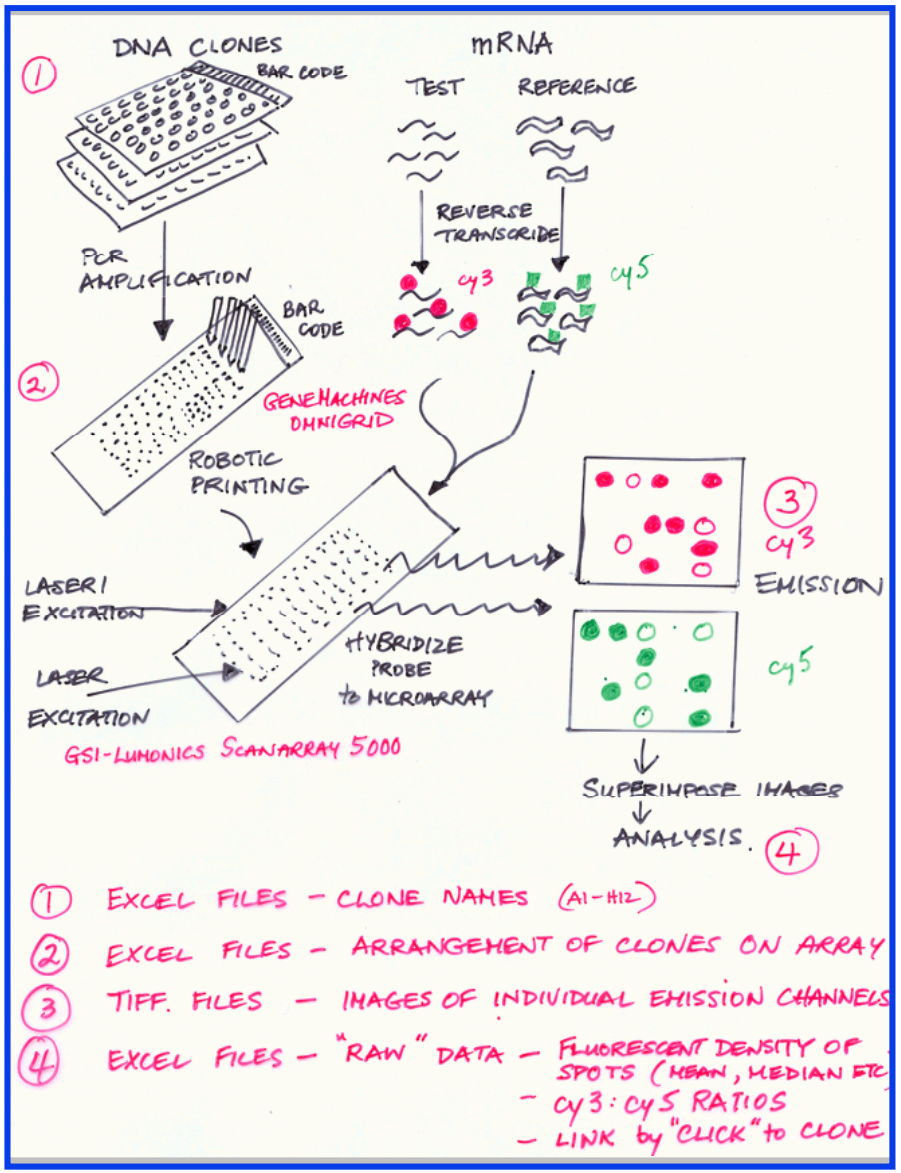
Array Data

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

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

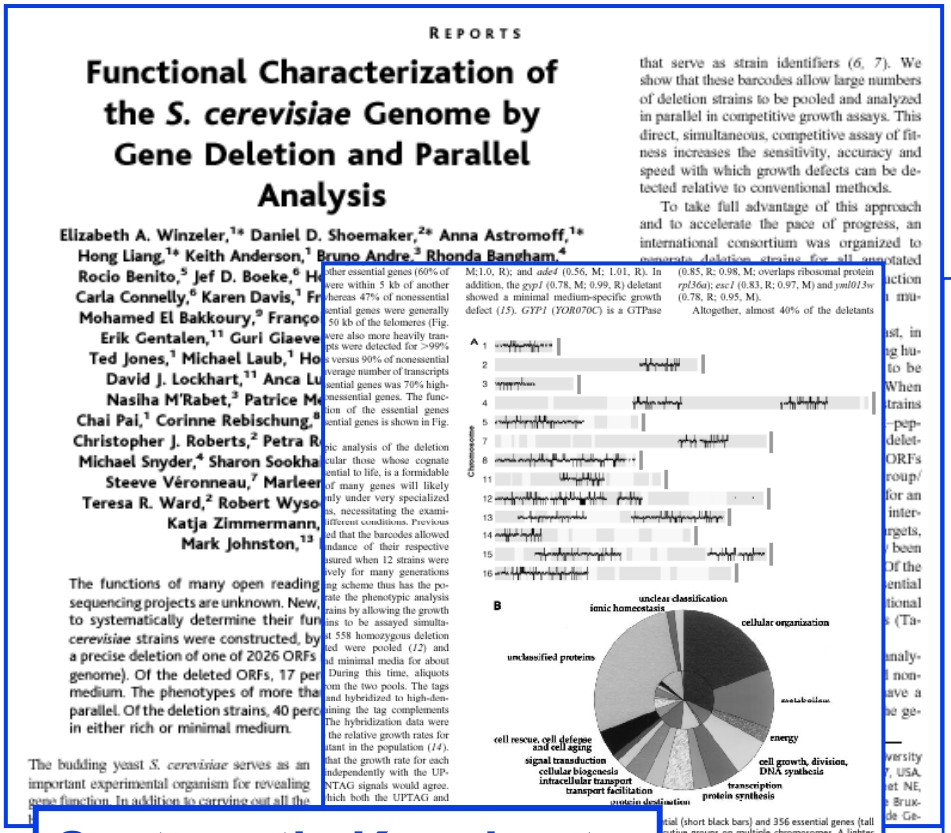
at 10 time points,
6000 x 10 = 60K floats

telling signal from background



(courtesy of J Hager)

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] H… Carla Connelly,[6] Karen Davis,[1] Fr… Mohamed El Bakkoury,[9] Franço… Erik Gentalen,[11] Guri Giaeve… Ted Jones,[1] Michael Laub,[1] Ho… David J. Lockhart,[11] Anca Lu… Nasiha M'Rabet,[3] Patrice M… Chai Pai,[1] Corinne Rebischung,[8] Christopher J. Roberts,[2] Petra R… Michael Snyder,[4] Sharon Sookha… Steeve Véronneau,[7] Marleen… Teresa R. Ward,[2] Robert Wyso… Katja Zimmermann,[1] Mark Johnston,[13]

The functions of many open reading… sequencing projects are unknown. New… to systematically determine their fu… *cerevisiae* strains were constructed, by… a precise deletion of one of 2026 ORFs … genome). Of the deleted ORFs, 17 per… medium. The phenotypes of more tha… parallel. Of the deletion strains, 40 perc… in either rich or minimal medium.

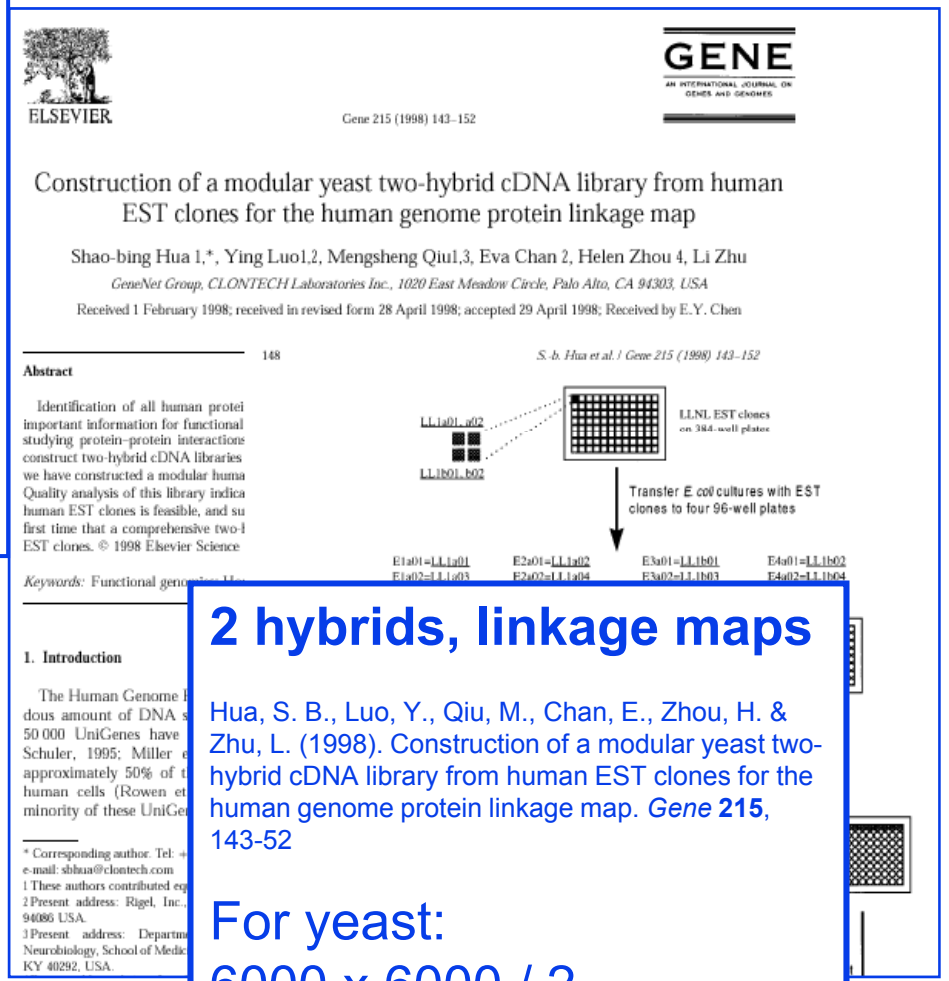
The budding yeast *S. cerevisiae* serves as an important experimental organism for revealing

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

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

Systematic Knockouts

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



ELSEVIER

GENE
AN INTERNATIONAL JOURNAL ON GENES AND GENOMES

Gene 215 (1998) 143–152

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

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

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

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

148

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

Abstract

Identification of all human protei… important information for functional… studying protein–protein interactions… construct two-hybrid cDNA libraries… we have constructed a modular huma… Quality analysis of this library indica… human EST clones is feasible, and su… first time that a comprehensive two-h… EST clones. © 1998 Elsevier Science

Keywords: Functional genomics; Hu…

1. Introduction

The Human Genome … dous amount of DNA … 50 000 UniGenes have … Schuler, 1995; Miller … approximately 50% of … human cells (Rowen et … minority of these UniGe…

* Corresponding author. Tel: +… e-mail: sbhua@clontech.com

1 These authors contributed eq…

2 Present address: Rigel, Inc., … 94086 USA.

3 Present address: Departm… Neurobiology, School of Medic… KY 40292, USA.

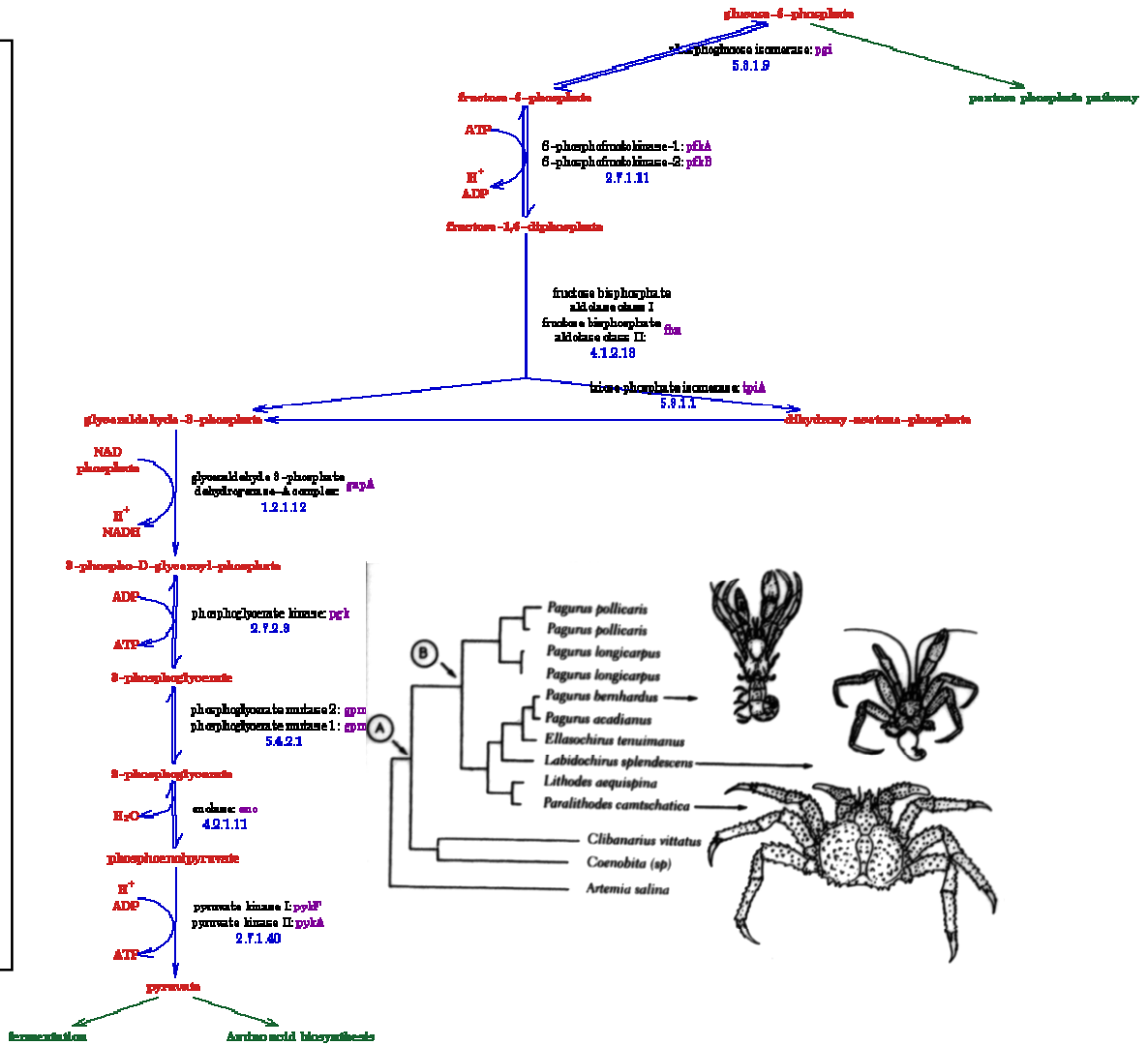
2 hybrids, linkage maps

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

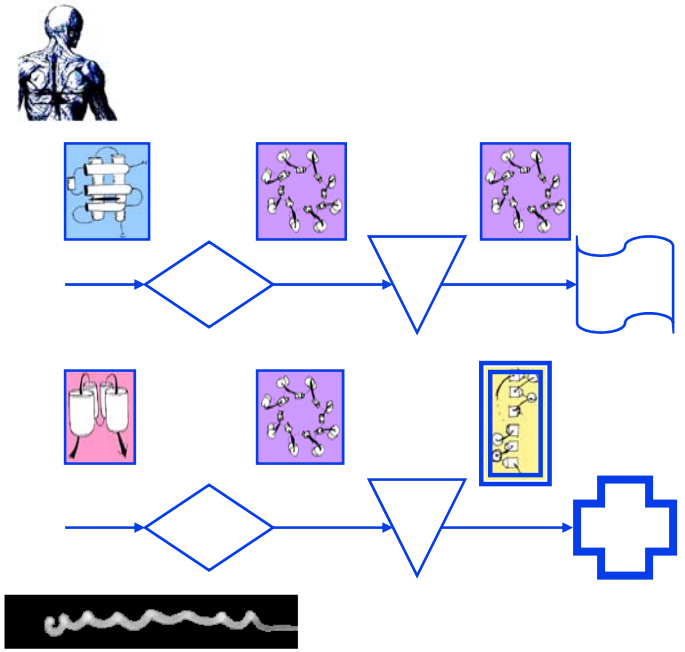
For yeast:
6000 x 6000 / 2
~ 18M interactions

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....



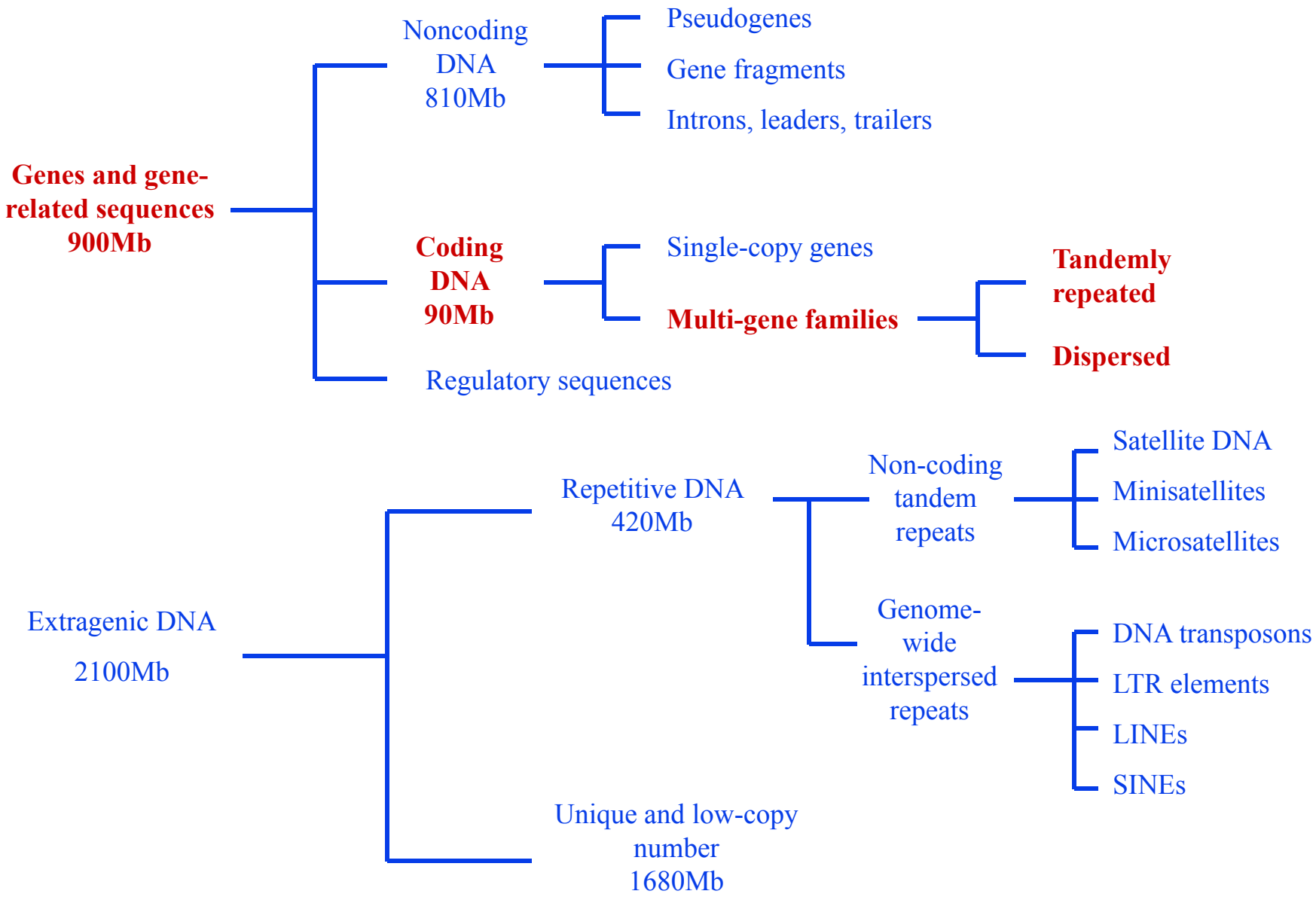
Organizing
Molecular Biology Information:
Redundancy and Multiplicity



- 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 ↔ ….

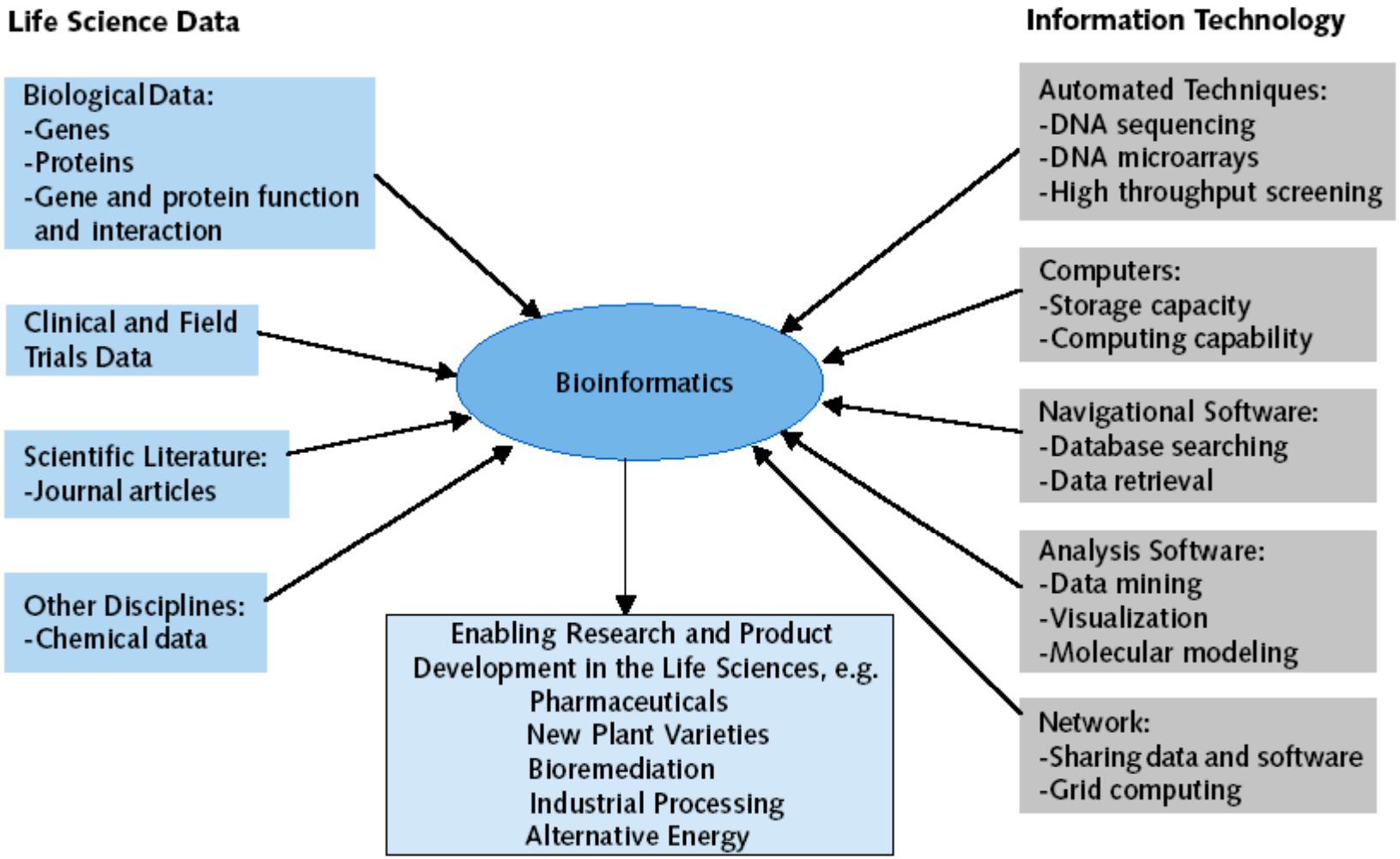
Human genome



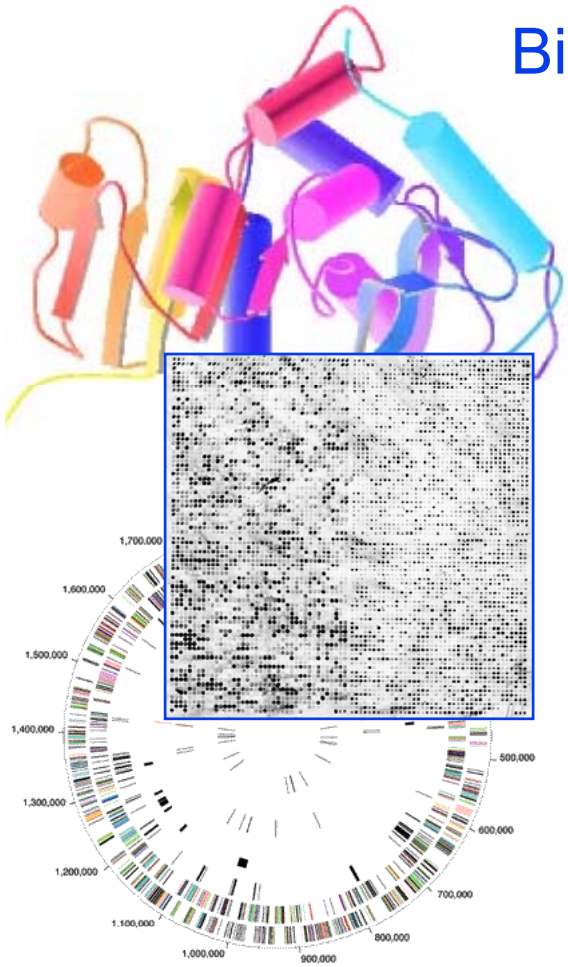
Where to get data?

- GenBank
  - http://www.ncbi.nlm.nih.gov
- Protein Databases
  - SWISS-PROT: http://www.expasy.ch/sprot
  - PDB: http://www.pdb.bnl.gov/
- And many others

Figure 6.1. Bioinformatics Uses Information Technology to Manage and Analyze Information Generated by the Life Sciences



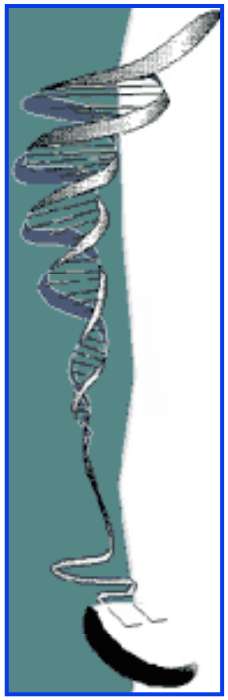
Bioinformatics: A simple view



Biological Data

+

Computer Calculations



Application domains

Table 6.2. Number of Survey Respondents Indicating Bioinformatics Research Activities by Application, 2002

Application	Number of firms in application	Conduct bioinformatics research
Human Health	780	247
Animal Health	144	37
Agricultural & Aquacultural/Marine	128	41
Marine & Terrestrial Microbial	41	19
Industrial and Agricultural-Derived Processing	132	45
Environmental Remediation and Natural Resource Recovery	41	12
Other	160	30

Bio-defense

Note: The total number of firms that responded to the biotechnology survey was 1,031, and 304 of these firms indicated that they had some activity in bioinformatics. The number of firms by biotechnology application does not add up to the total number of firms that responded to the survey because firms were classified in an application if they indicated it as either a "primary" or "secondary" focus.

Source: Survey data from *Critical Technology Assessment of Biotechnology in U.S. Industry*, U.S. Department of Commerce, Technology Administration and Bureau of Industry and Security, August 2002.

Kinds of activities

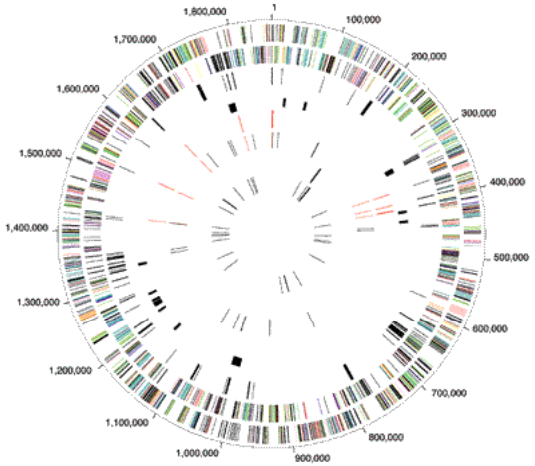
	Conduct research on/in	Approved, marketed, or in production		Total
		Product(s)	Process(es)	
DNA-based				
Bioinformatics	29	2	1	30
Genomics, pharmacogenetics	29	3	2	30
DNA sequencing/synthesis/ amplification, genetic engineering	39	5	3	43
Biochemistry/Immunology				
Drug design & delivery	33	4	2	38
Synthesis/sequencing of proteins and peptides	27	3	1	30
Combinatorial chemistry, 3-D molecular modeling	18	1	0	19

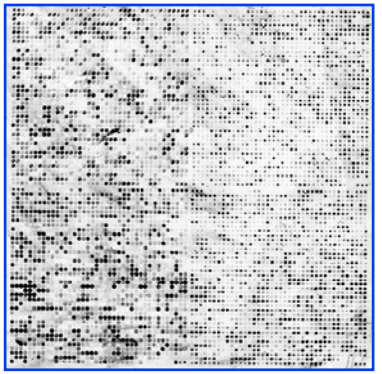
Note: The total number of responses to the biotechnology activity question was 1021. Percents do not add up to 100 percent because firms can have more than one activity.

Source: Survey data from *Critical Technology Assessment of Biotechnology in U.S. Industry*, U.S. Department of Commerce, Technology Administration and Bureau of Industry and Security, August 2002.

Motivation

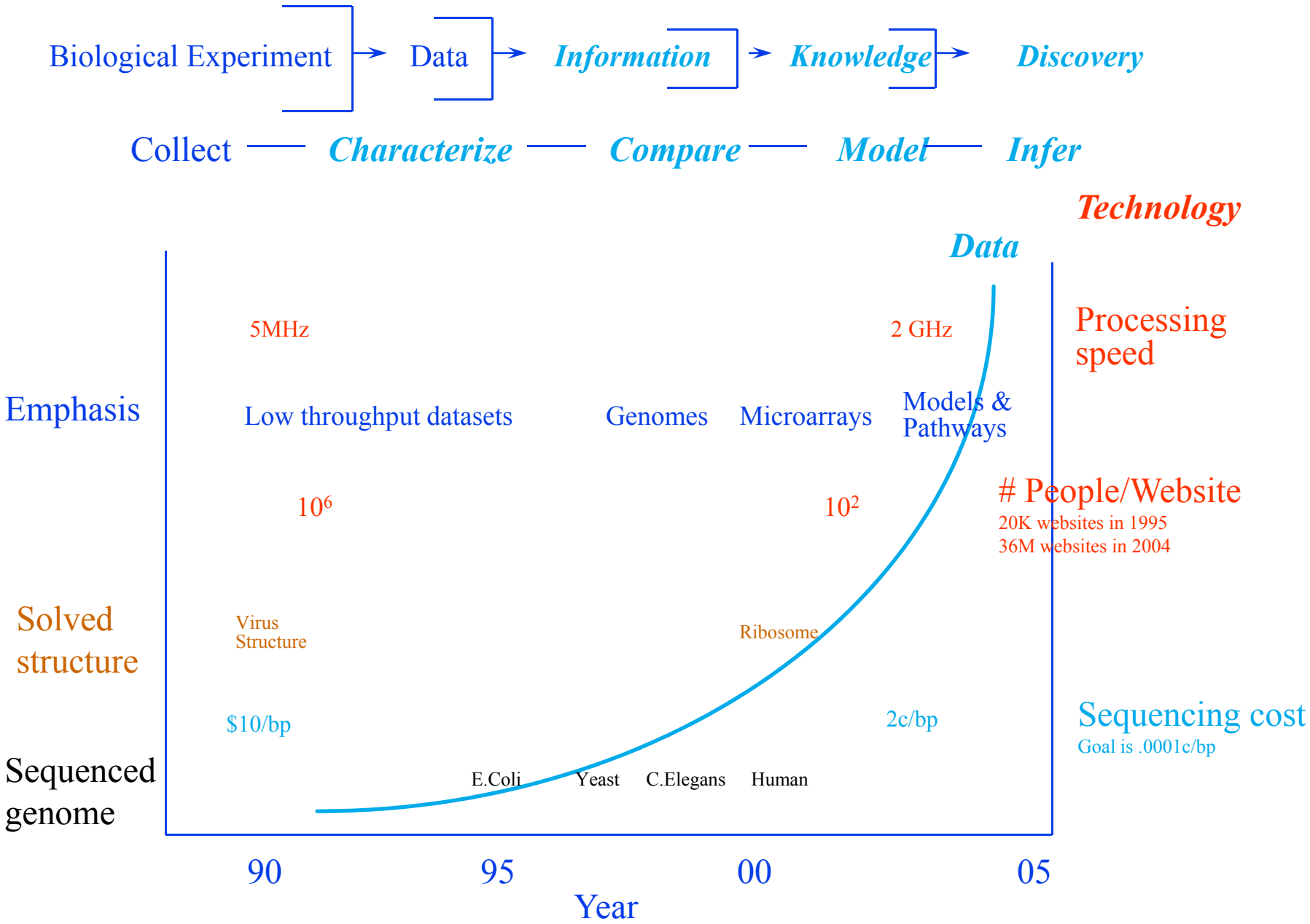
- Diversity and size of information
  - Sequences, 3-D structures, microarrays, protein interaction networks, *in silico* models, bio-images





- Understand the relationship
  - Similar to complex software design

Bioinformatics - A Revolution



Computing *versus* Biology

- *what computer science is to molecular biology is like what mathematics has been to physics .....*

  -- Larry Hunter, ISMB'94

- *molecular biology is (becoming) an information science .....*

  -- Leroy Hood, RECOMB'00

- *bioinformatics ... is the research domain focused on linking the behavior of biomolecules, biological pathways, cells, organisms, and populations to the information encoded in the genomes* --Temple Smith, Current Topics in Computational Molecular Biology

Computing *versus* Biology

looking into the future

- *Like physics, where general rules and laws are taught at the start, biology will surely be presented to future generations of students as a set of basic systems ..... duplicated and adapted to a very wide range of cellular and organismic functions, following basic evolutionary principles constrained by Earth's geological history.* --Temple Smith, Current Topics in Computational Molecular Biology

Scalability challenges

- Recent issue of NAR devoted to data collections contains 719 databases
  - Sequence
    - Genomes (more than 150), ESTs, Promoters, transcription factor binding sites, repeats, ..
  - Structure
    - Domains, motifs, classifications, ..
  - Others
    - Microarrays, subcellular localization, ontologies, pathways, SNPs, ..

Challenges of working in bioinformatics

- Need to feel comfortable in interdisciplinary area
- Depend on others for primary data
- Need to address important biological *and* computer science problems

Skill set

- Artificial intelligence
- Machine learning
- Statistics & probability
- Algorithms
- Databases
- Programming

Bioinformatics Topics
Genome Sequence

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

Bioinformatics Topics
Protein Sequence

- Sequence Alignment
  - non-exact string matching, gaps
  - How to align two strings optimally via Dynamic Programming
  - Local vs Global Alignment
  - Suboptimal Alignment
  - Hashing to increase speed (BLAST, FASTA)
  - Amino acid substitution scoring matrices
- Multiple Alignment and Consensus Patterns
  - How to align more than one sequence and then fuse the result in a consensus representation
  - Transitive Comparisons
  - HMMs, Profiles
  - Motifs

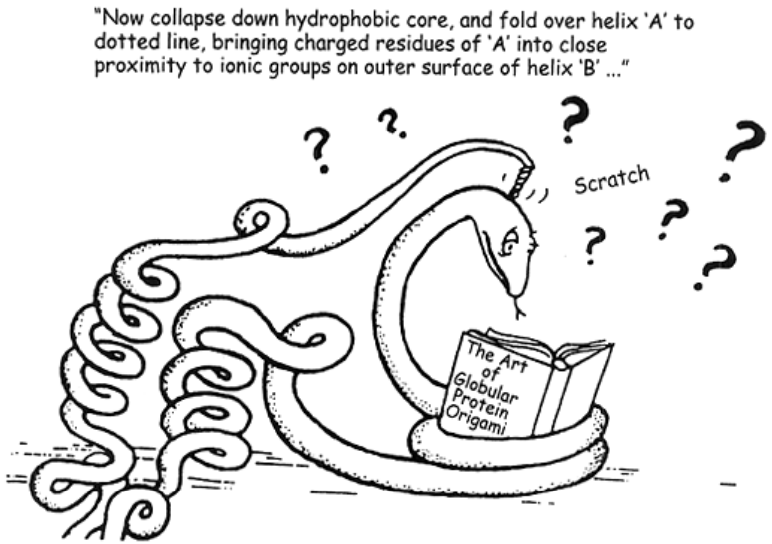
- Scoring schemes and Matching statistics
  - How to tell if a given alignment or match is statistically significant
  - A P-value (or an e-value)?
  - Score Distributions (extreme val. dist.)
  - Low Complexity Sequences
- Evolutionary Issues
  - Rates of mutation and change

Computationally challenging problems

- More sensitive pairwise alignment
  - Dynamic programming is O(mn)
    - m is the length of the query
    - n is the length of the database
- Scalable multiple alignment
  - Dynamic programming is exponential in number of sequences
  - Currently feasible for around 10 protein sequences of length around 1000
- Shotgun alignment
  - Current techniques will take over 200 days on a single machine to align the mouse genome

Bioinformatics Topics
Sequence / Structure

- Secondary Structure "Prediction"
  - via Propensities
  - Neural Networks, Genetic Alg.
  - Simple Statistics
  - TM-helix finding
  - Assessing Secondary Structure Prediction
- Structure Prediction: Protein and RNA



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

Computationally challenging problems

- Alignment against a database
  - Single comparison usually takes seconds.
  - Comparison against a database takes hours.
  - All-against-all comparison takes weeks.
- Multiple structure alignment and motifs
- Combined sequence and structure comparison
- Secondary and tertiary structure prediction

Topics -- Databases

- Relational Database Concepts and how they interface with Biological Information
  - Keys, Foreign Keys
  - SQL, OODBMS, views, forms, transactions, reports, indexes
  - Joining Tables, Normalization
    - Natural Join as "where" selection on cross product
    - Array Referencing (perl/dbm)
  - Forms and Reports
  - Cross-tabulation
- 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
- Visualization of Large Amounts of Information
- 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

General Types of "Informatics" techniques in Bioinformatics

- Databases
  - Building, querying
  - Schema design
  - Heterogeneous, distributed
- Similarity search
  - Sequence, structure
  - Significance statistics

- Finding Patterns
  - AI / Machine Learning
  - Clustering
  - Data mining
- Modeling & simulation
- Programming
  - Perl
  - Java/C/C++/..