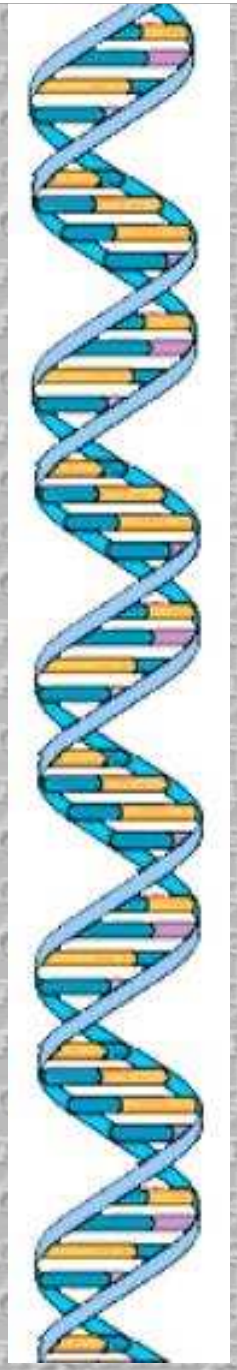
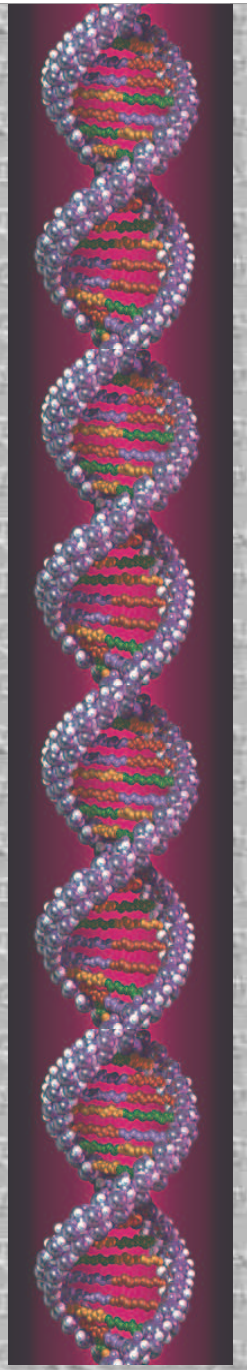


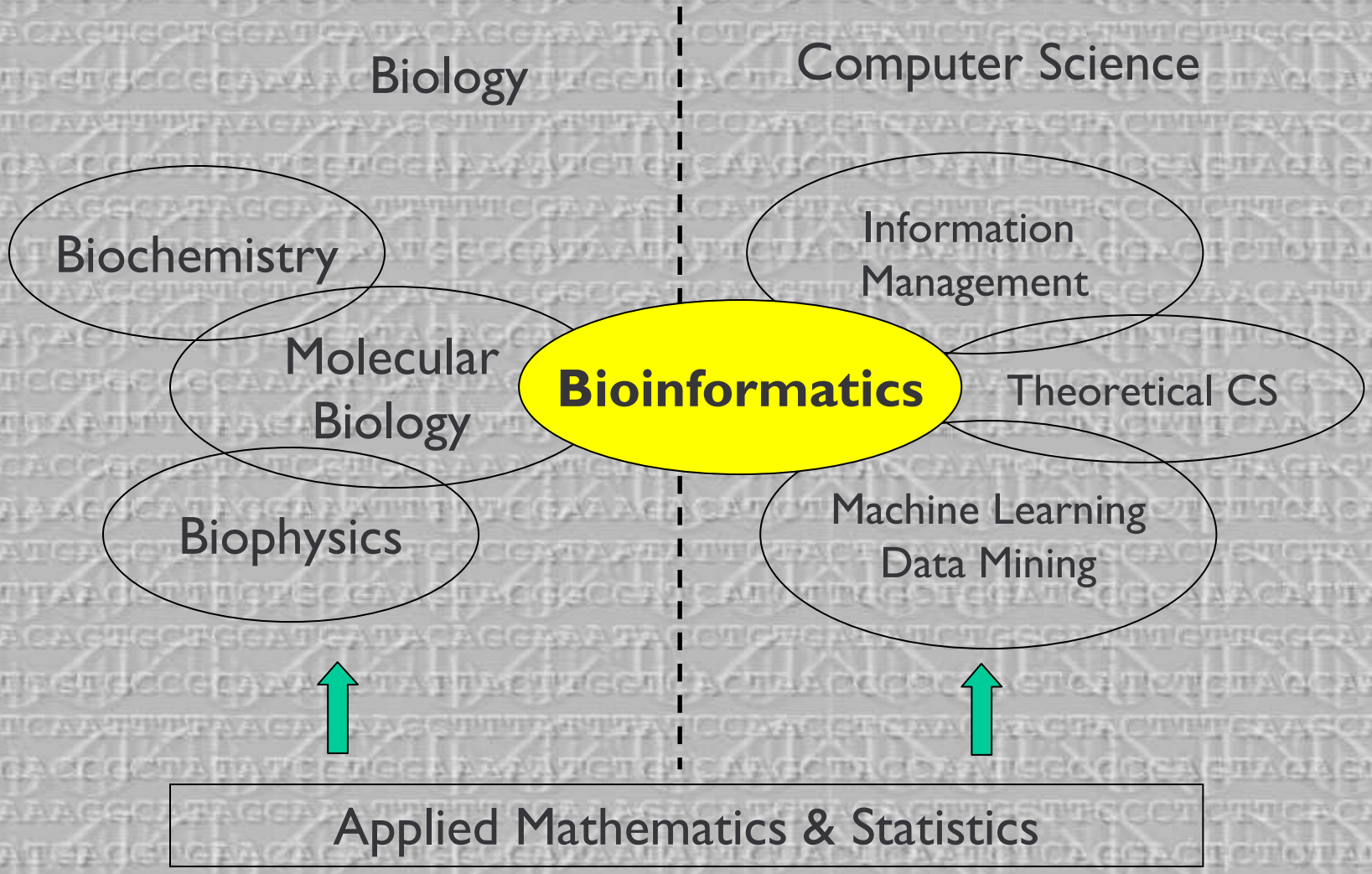
Bioinformatics



مهدی صادقی

پژوهشگاه ملی مهندسی ژنتیک و زیست فناوری

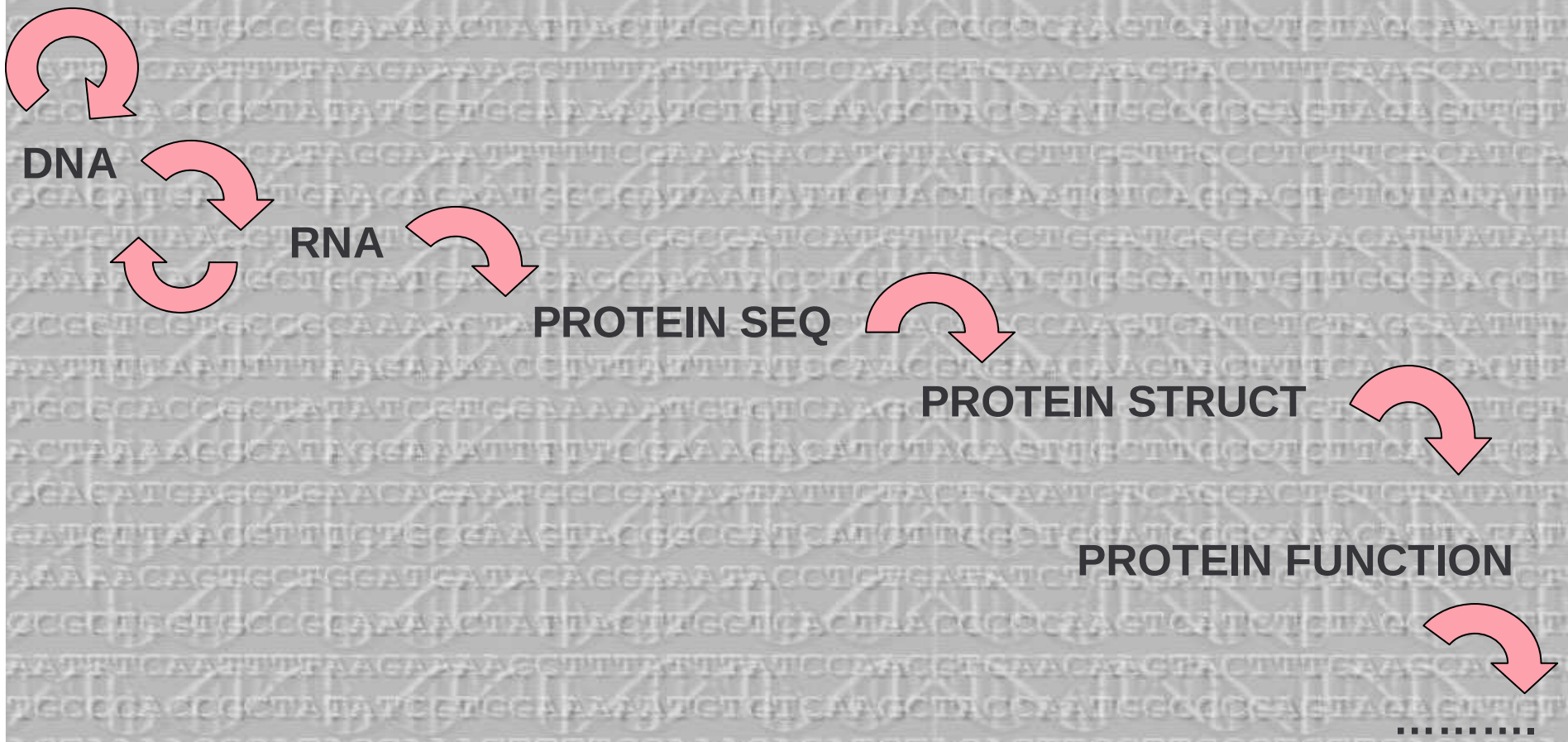
Bioinformatics is interdisciplinary



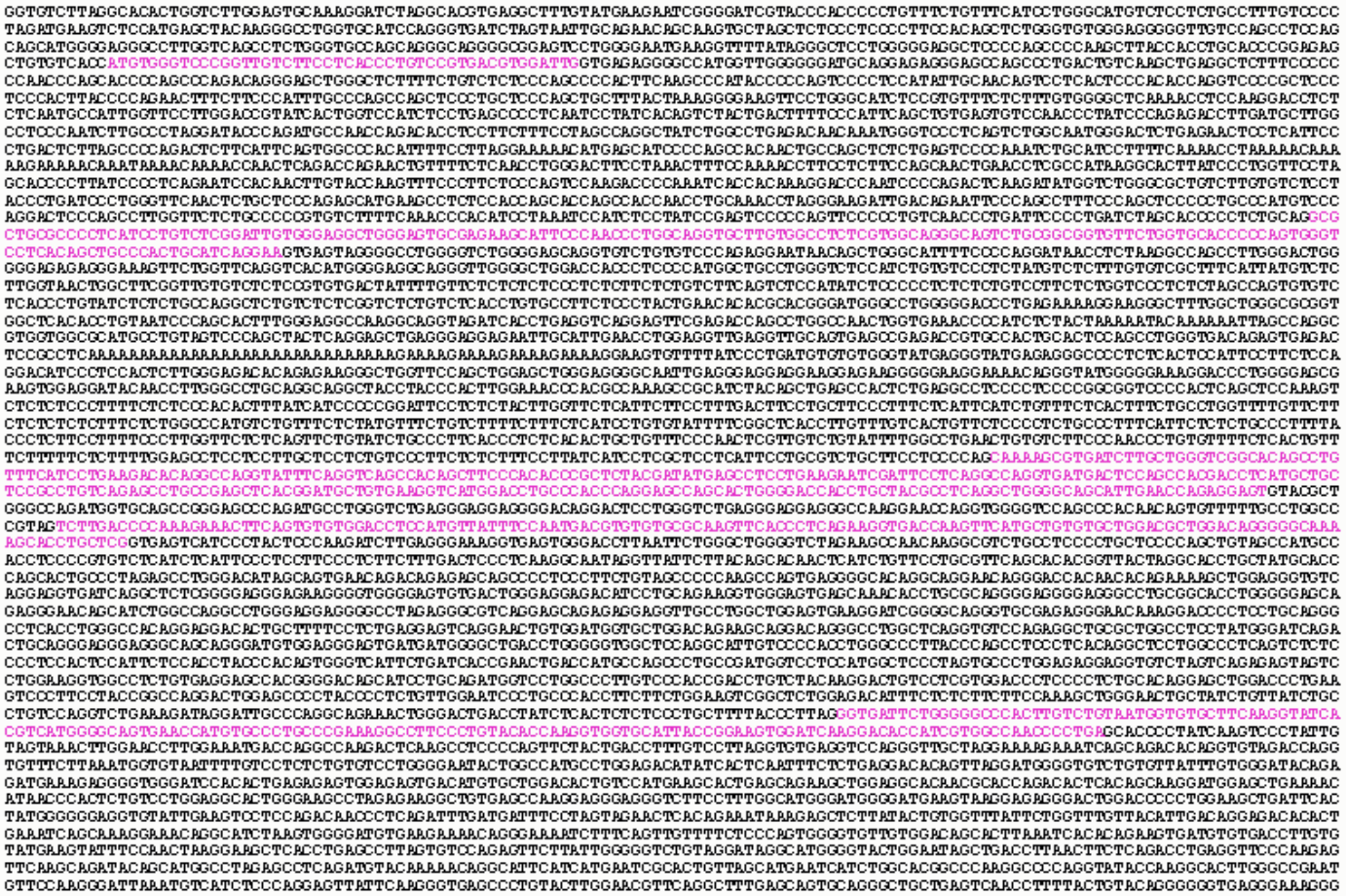
bio – informatics:

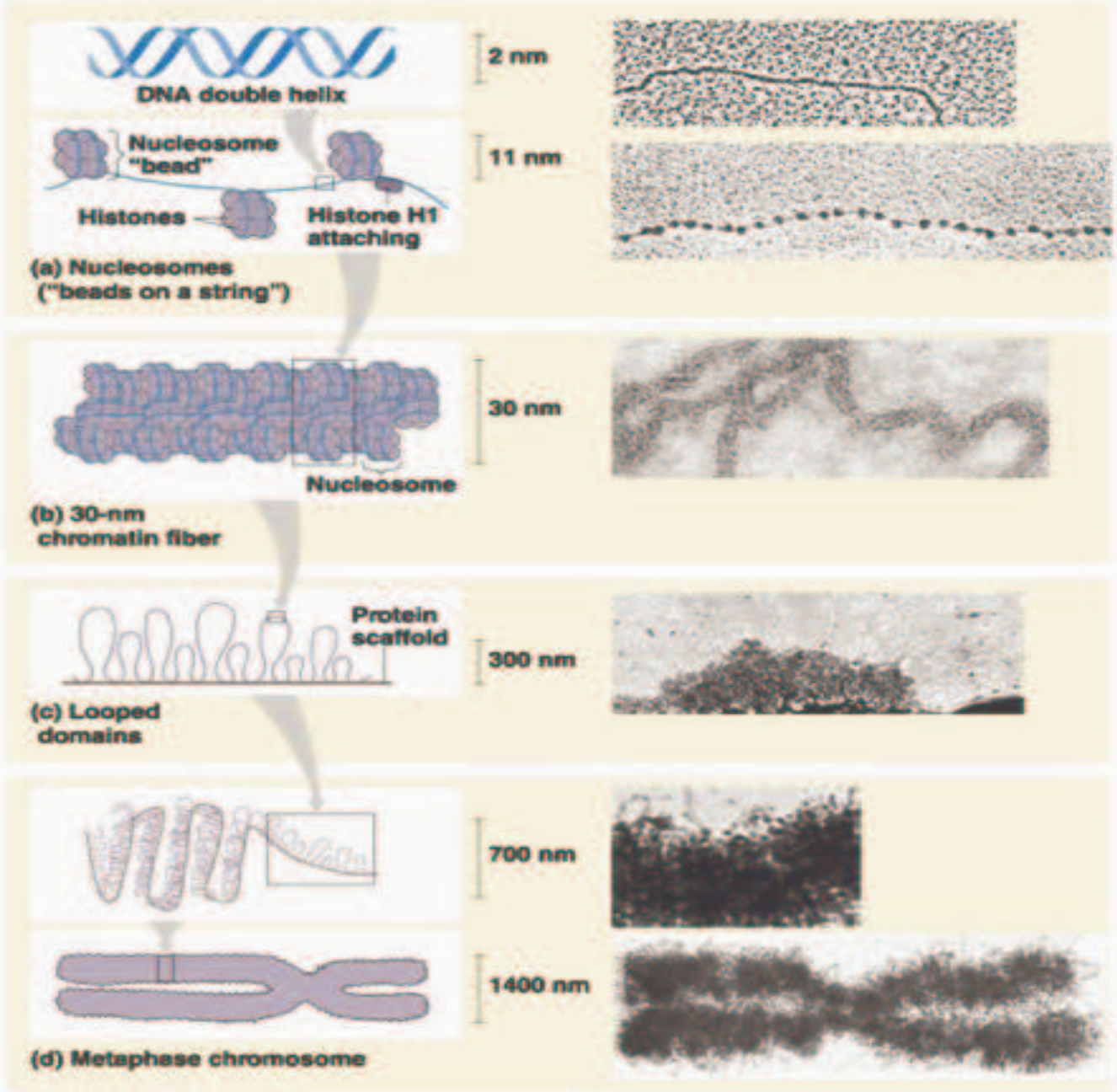
bioinformatics is conceptualizing biology in terms of molecules (in the sense of physical chemistry) and applying "***informatics techniques***" (derived from disciplines such as applied maths, computer science and statistics) to ***understand*** and ***organize*** the ***information*** associated with these molecules, on a ***large scale***. In short, bioinformatics is a management information system for molecular biology and has many ***practical applications***.

Flow of information

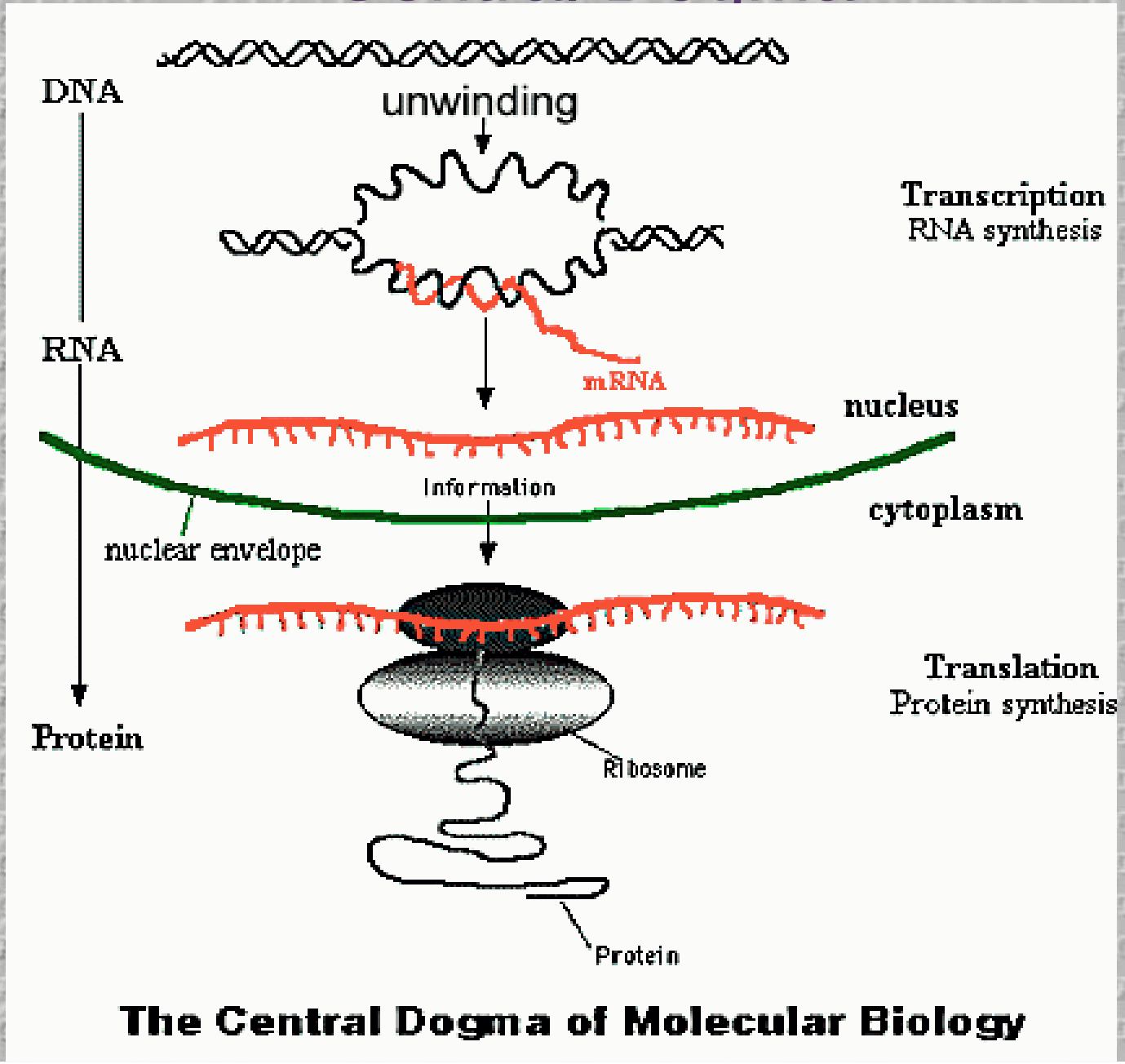


Structure of a Human Gene (PSA)





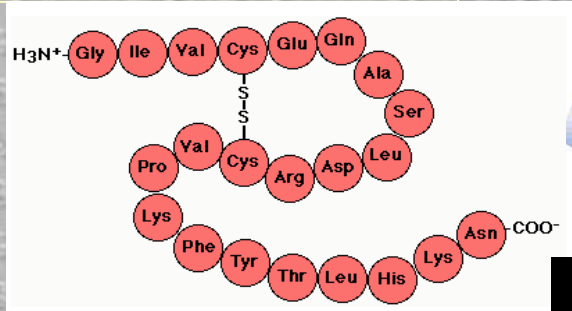
Central Dogma



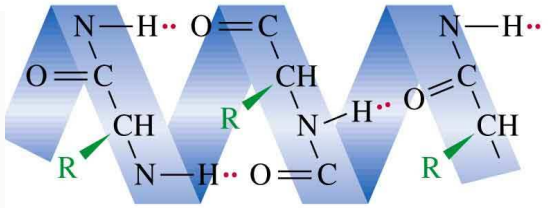
There Are Four Levels of Protein Structure

Primary: amino acid linear sequence.

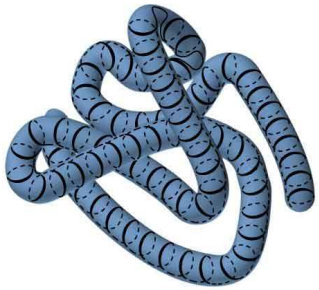
Secondary: α-helices, β-sheets and loops.



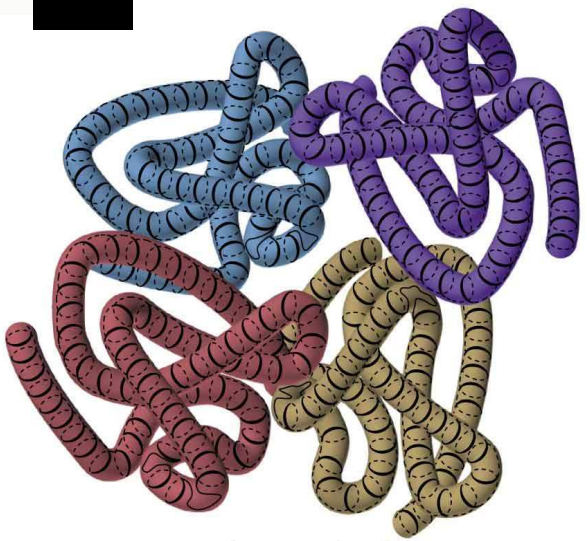
primary structure



secondary structure



tertiary structure



quaternary structure

Tertiary: the 3D shape of the fully folded polypeptide chain

Quaternary: arrangement of several polypeptide chains.

Early History

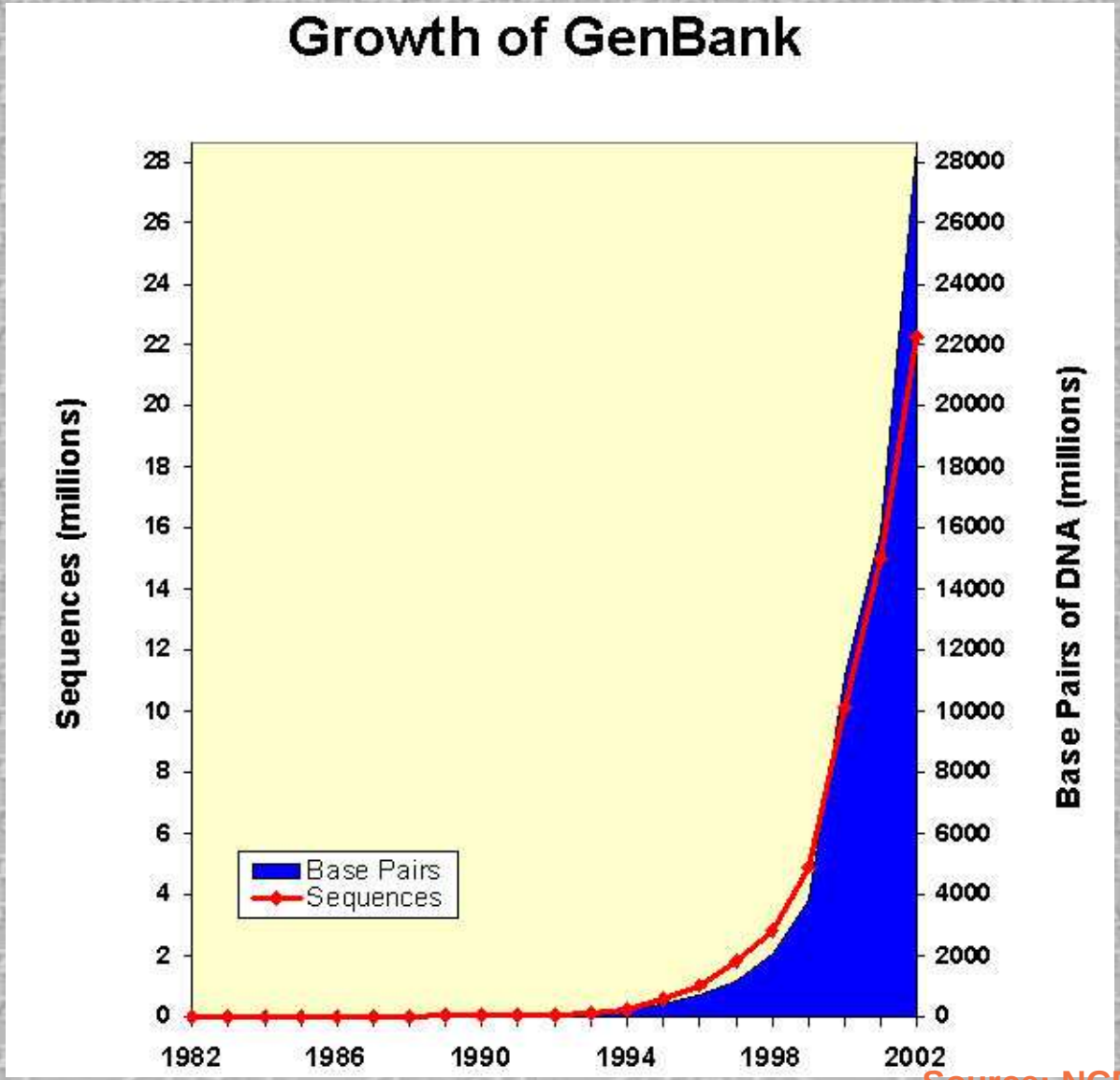
- First protein amino acid sequence database in mid 1960's.
- Researchers developing algorithms to analyze these data for individual research projects in 1960's and 1970's.
- GenBank and other public databases with freely available analysis tools in the 1980's
- Enormous growth of GenBank and PDB in 1990's

Data Explosion

- These research trends, especially the sequencing projects, have resulted in huge amounts of raw data that must be stored, manipulated, and analyzed.

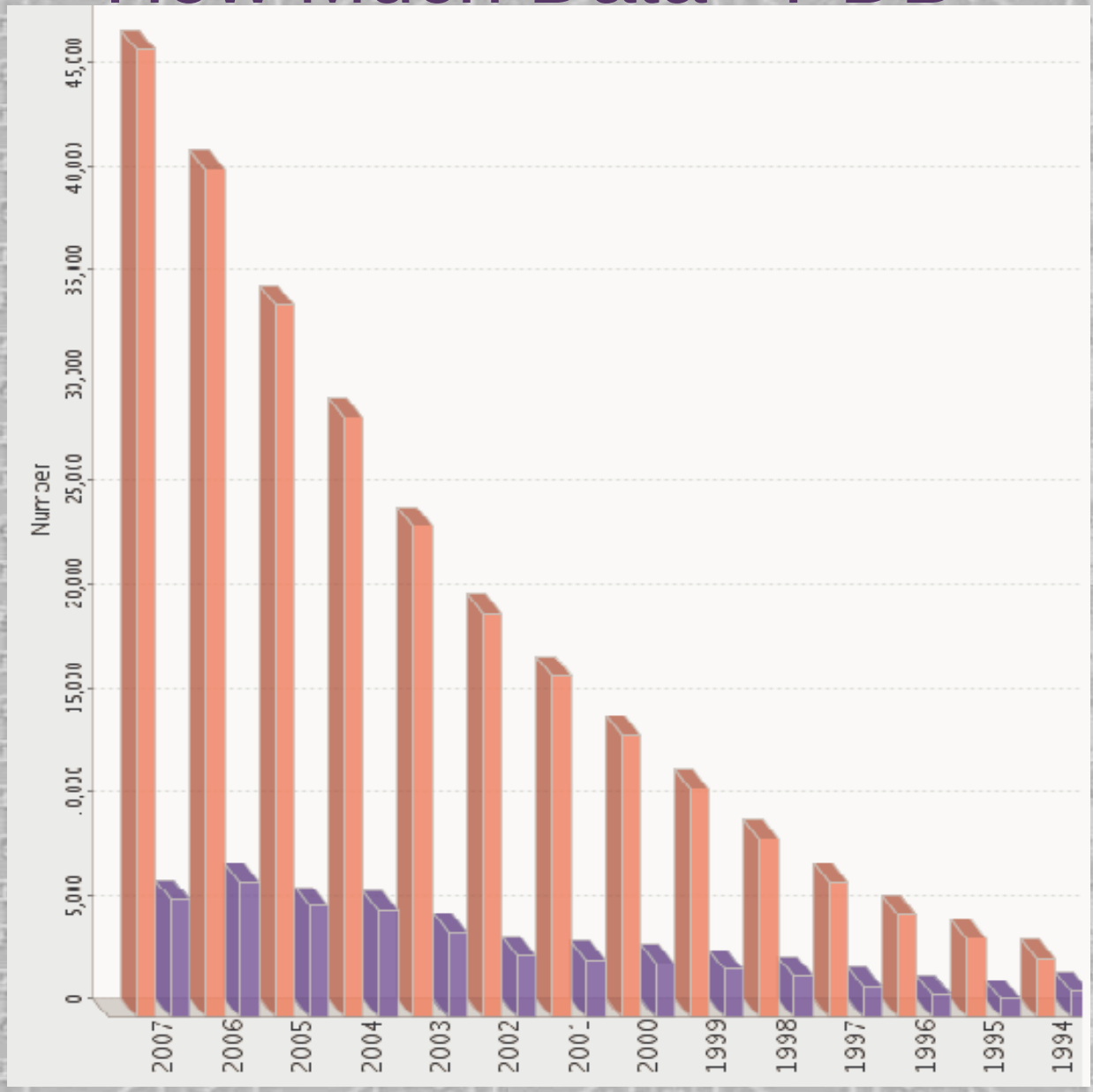
- Computer information technology has also advanced very rapidly and has provided the means for handling this data explosion.

How Much Data - GenBank



Source: NCBI

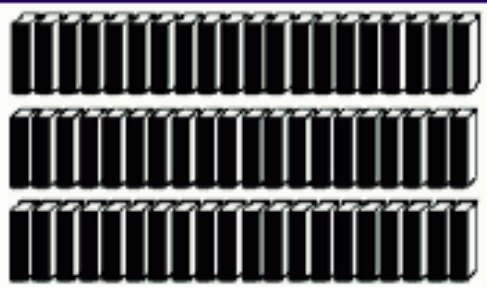
How Much Data - PDB



Source: RSCB

Genome Sizes

Human Genome
Mouse Genome



~3,000,000,000 bp

Fruit Fly Genome

~160,000,000 bp

Nematode Genome

~100,000,000 bp

Yeast Genome

~15,000,000 bp

E. coli Genome

~5,000,000 bp

>100,000 species are represented in GenBank

all species	128,941
viruses	6,137
bacteria	31,262
archaea	2,100
eukaryota	87,147

Vertebrate Genome Sequences

Mouse

Rat

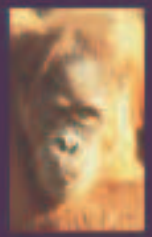
Chicken

Chimpanzee

Dog

Macaque

Monodelphis


Orangutan


Marmoset


Horse


Cow


Platypus


Xenopus

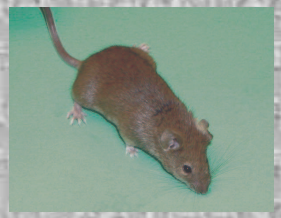

Zebrafish


Pufferfish

The most sequenced organisms in GenBank

Homo sapiens **(6.9 million entries)**

Mus musculus **(5.0 million)**

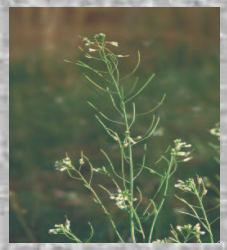


Zea mays **(896,000)**

Rattus norvegicus **(819,000)**

Gallus gallus **(567,000)**

Arabidopsis thaliana **(519,000)**



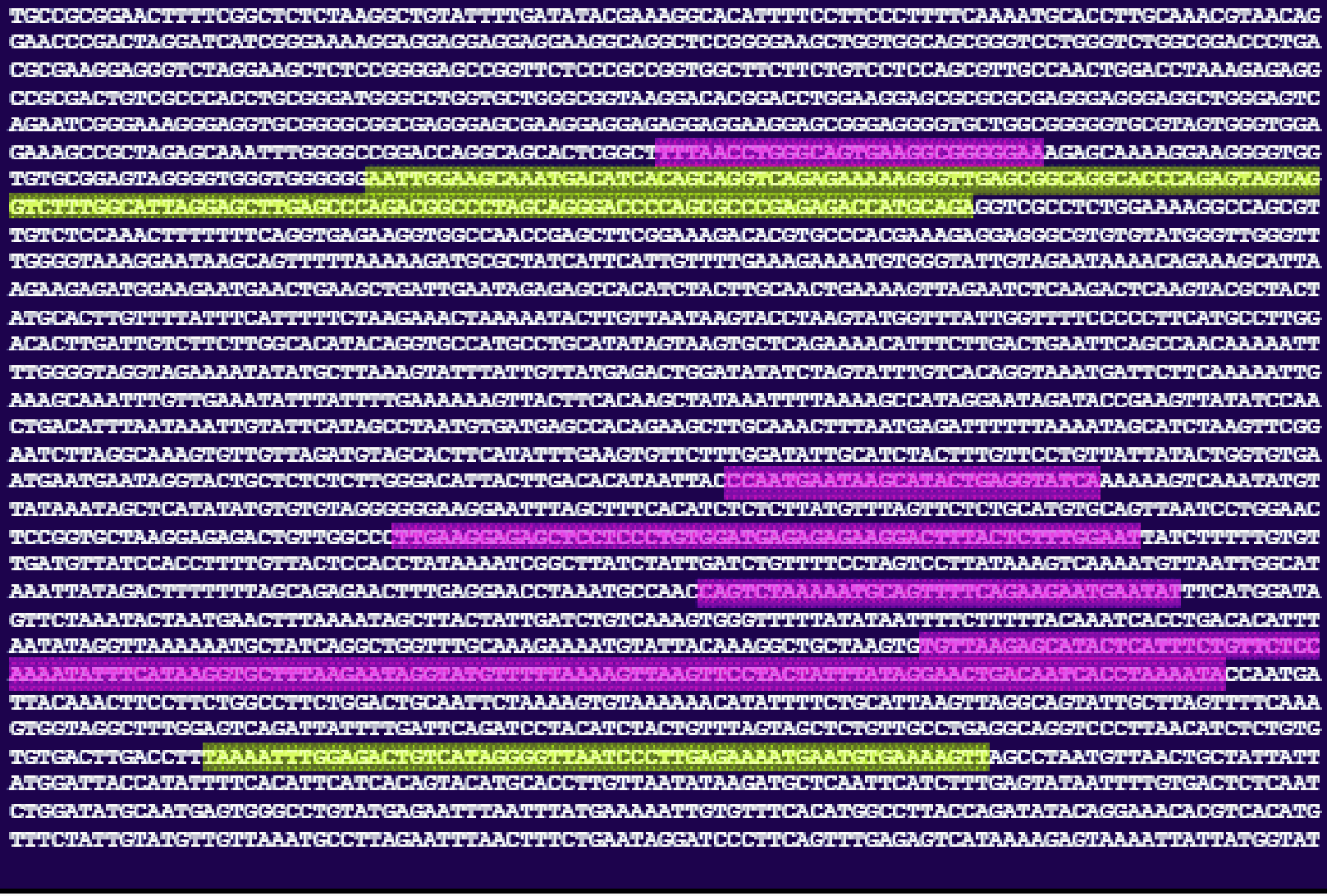
Danio rerio **(492,000)**

Drosophila melanogaster **(350,000)**

Orzya sativa **(221,000)**

Human Genome Sequencing Centers

~3,000 bp (0.0001%) of Human Genome Sequence



Aims of bioinformatics

- First

Data organization

researchers access to existing information
submit new entries

- Second

develop tools and resources that aid in the analysis of data

- Third

interpret the results in a biologically meaningful manner.

"… *ORGANISE* the information on a *LARGE SCALE* …"

Database	URL
Protein sequence (primary)	
SWISS-PROT	www.expasy.ch/sprot/sprot-top.html
PIR-International	www.mips.biochem.mpg.de/proj/protseqdb
Protein sequence (composite)	
OWL	www.bioinf.man.ac.uk/dbbrowser/OWL
NRDB	www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Protein
Protein sequence (secondary)	
PROSITE	www.expasy.ch/prosite
PRINTS	www.bioinf.man.ac.uk/dbbrowser/PRINTS/PRINTS.html
Pfam	www.sanger.ac.uk/Pfam/
Macromolecular structures	
Protein Data Bank (PDB)	www.rcsb.org/pdb
Nucleic Acids Database (NDB)	ndbserver.rutgers.edu/
HIV Protease Database	www.ncifcrf.gov/CRYS/HIVdb/NEW_DATABASE
ReLiBase	www2.ebi.ac.uk:8081/home.html
PDBsum	www.biochem.ucl.ac.uk/bsm/pdbsum
CATH	www.biochem.ucl.ac.uk/bsm/cath
SCOP	scop.mrc-lmb.cam.ac.uk/scop
FSSP	www2.embl-ebi.ac.uk/dali/fssp
Nucleotide sequences	
GenBank	www.ncbi.nlm.nih.gov/Genbank
EMBL	www.ebi.ac.uk/embl
DDBJ	www.ddbj.nig.ac.jp/
Genome sequences	
Entrez genomes	www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome
GeneCensus	bioinfo.mbb.yale.edu/genome
COGs	www.ncbi.nlm.nih.gov/COG
Integrated databases	
InterPro	www.ebi.ac.uk/interpro
Sequence retrieval system (SRS)	www.expasy.ch/srs5
Entrez	www.ncbi.nlm.nih.gov/Entrez

"…the *INFORMATION* associated with these molecules…"

Data source	Data size	Bioinformatics topics
Raw DNA sequence	million sequences (billion bases)	Separating coding and non-coding regions Identification of introns and exons Gene product prediction Forensic analysis
Protein sequence	sequences (~300 amino acids each)	Sequence comparison algorithms Multiple sequence alignments algorithms Identification of conserved sequence motifs
Macromolecular structure	structures (~1,000 atomic coordinates each)	Secondary, tertiary structure prediction 3D structural alignment algorithms Protein geometry measurements Surface and volume shape calculations Intermolecular interactions Molecular simulations (force-field calculations, molecular movements, docking predictions)
Genomes	complete genomes (1.6 million – 3 billion bases each)	Characterisation of repeats Structural assignments to genes Phylogenetic analysis Genomic-scale censuses (characterisation of protein content, metabolic pathways) Linkage analysis relating specific genes to diseases
Gene expression	largest: ~20 time point measurements for ~6,000 genes	Correlating expression patterns Mapping expression data to sequence, structural and biochemical data
Other data		
Literature	million citations	Digital libraries for automated bibliographical searches Knowledge databases of data from literature
Metabolic pathways		Pathway simulations

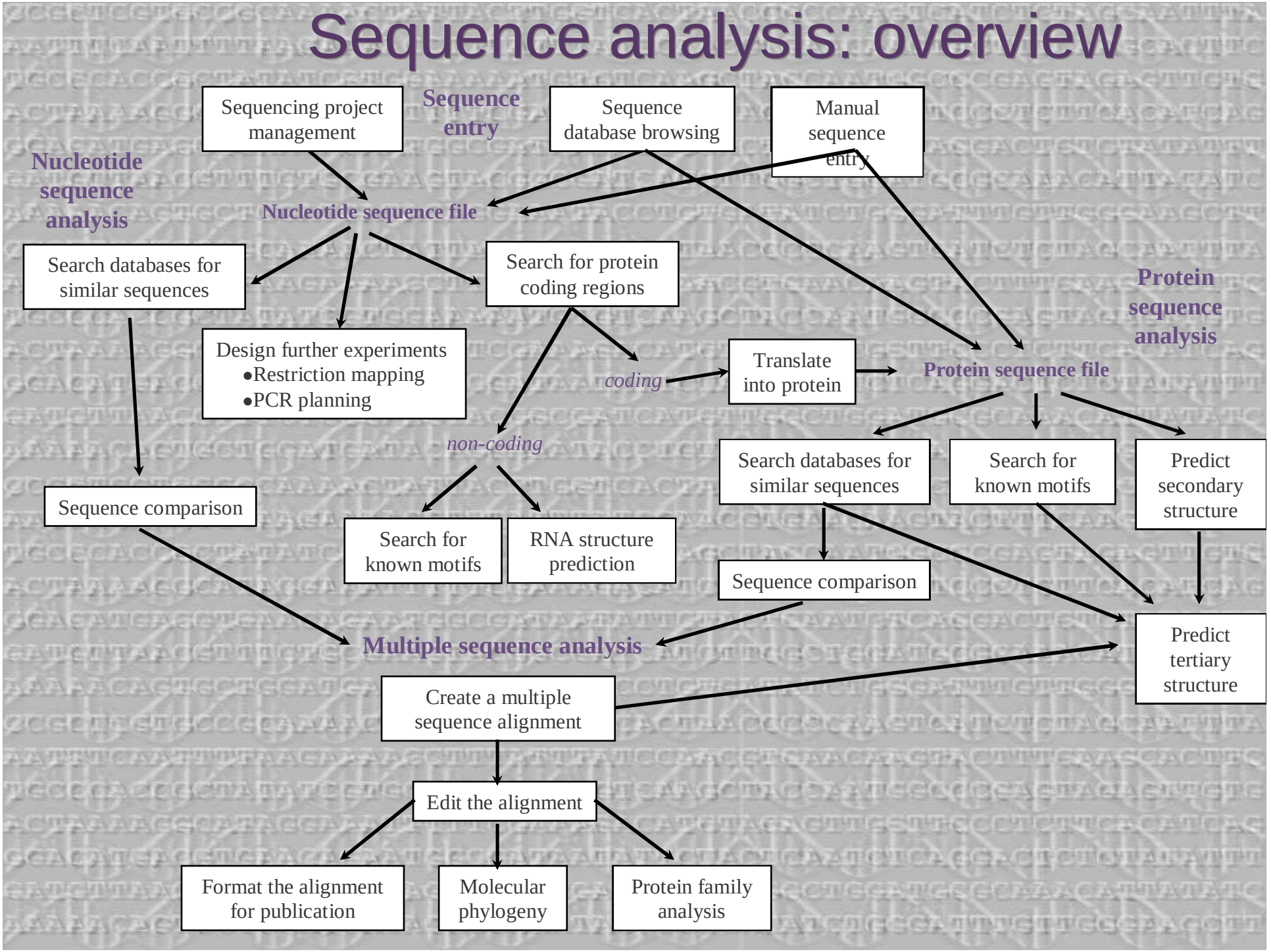
General Types of "….Informatics techniques….."

- **Databases**
  - Building, Querying
  - Object DB
- **Text String Comparison**
  - Text Search
  - 1D Alignment
  - Significance Statistics
- **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

Bioinformatcs Tools and Services

- Databases: text, sequence, structure
- Database annotation text searches
- Sequence similarity search tools
- Sequence and structure analysis tools
- 3D structure visualization tools
- Structure prediction tools
- Phylogenetic analysis tools
- Metabolic analysis tools

Sequence analysis: overview

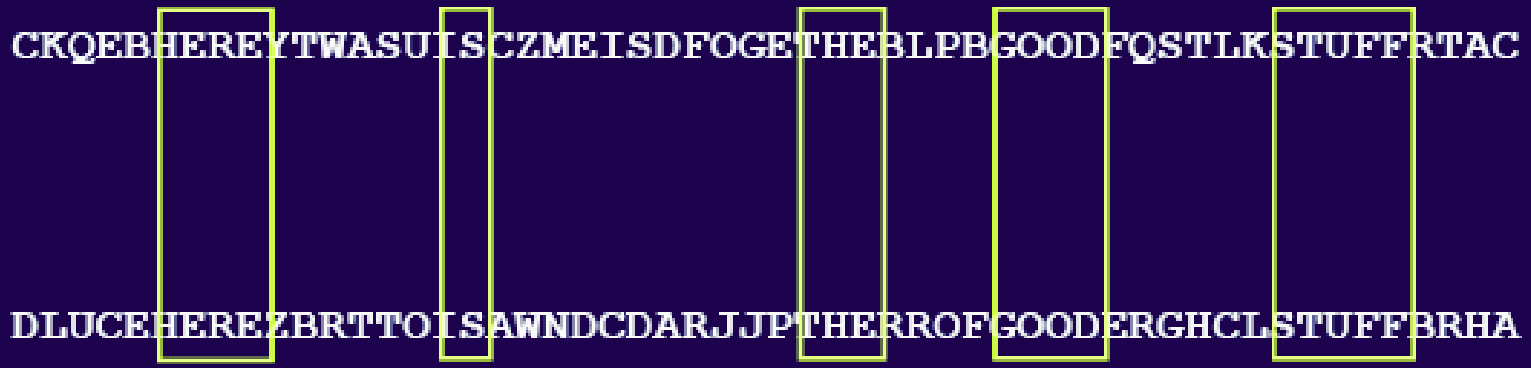


Sequence comparison:

Gene sequences can be aligned to see similarities between gene from different sources

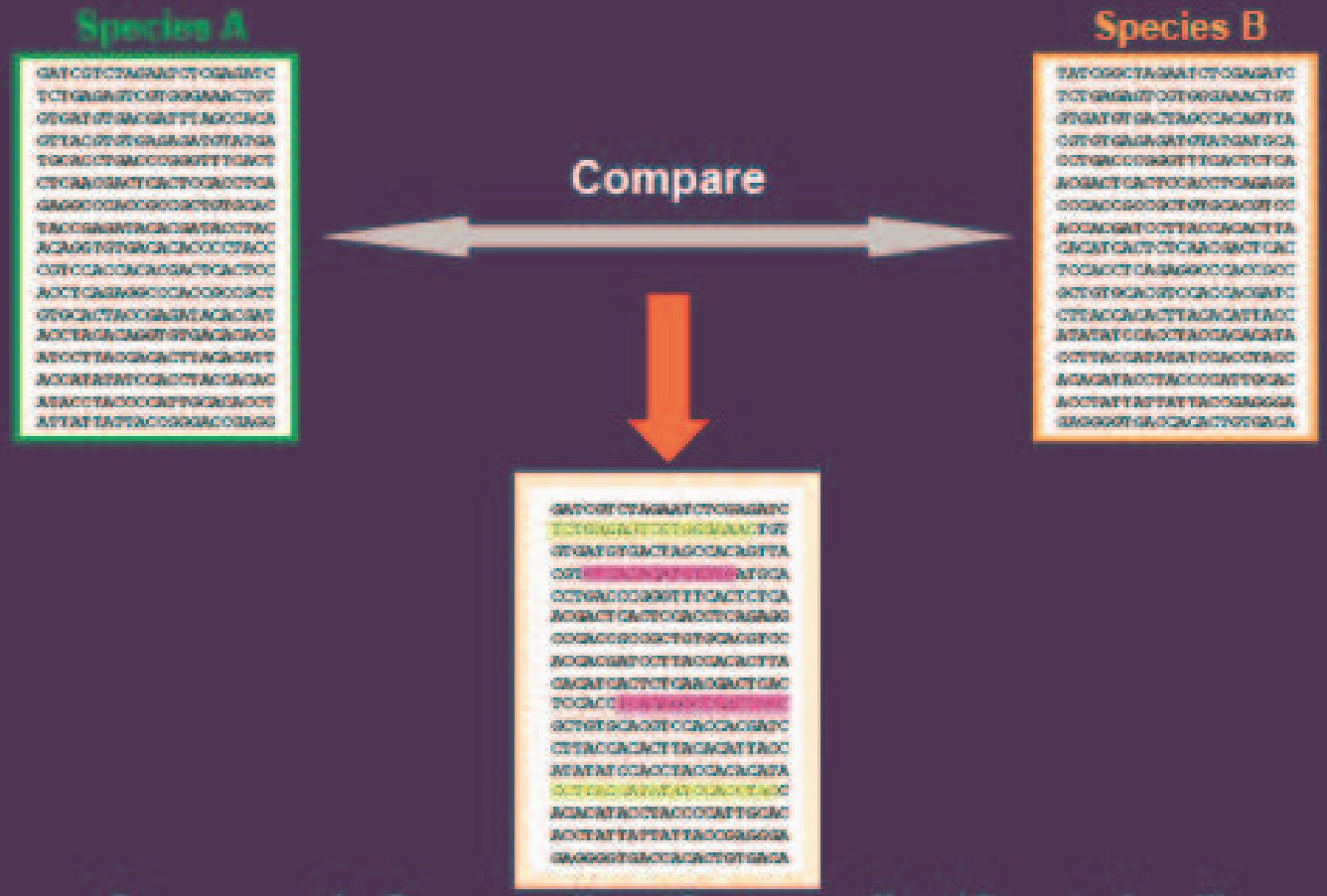
```
768 TT....TGTGTGCATTTAAGGGTGATAGTGTATTTGCTCTTTAAGAGCTG 813
    ||    ||   ||  |  | ||| |  |||| |||||   |||  |||
 87 TTGACAGGTACCCAACTGTGTGTGCTGATGTA.TTGCTGGCCAAGGACTG 135
               .         .         .         .         .
814 AGTGTTTGAGCCTCTGTTTGTGTGTAATTGAGTGTGCATGTGTGGGAGTG 863
    |  | |                 |  |||||| |   ||||  | || |   |
136 AAGGATC.............TCAGTAATTAATCATGCACCTATGTGGCGG 172
              .         .         .         .         .
864 AAATTGTGGAATGTGTATGCTCATAGCACTGAGTGAAAATAAAAGATTGT 913
    ||| | ||| || || |||   |  |||||||||  ||   |||||| |
173 AAA.TATGGGATATGCATGTCGA...CACTGAGTG..AAGGCAAGATTAT 216
```

Comparing Genomes is Like Cryptography



Comparative Sequence Analysis

Using the 'Experiments of Evolution' to Decode the Human Genome



Database similarity searching:

The BLAST program has been written to allow rapid comparison of a new gene sequence with the 100s of 1000s of gene sequences in data bases

```
Sequences producing significant alignments:                      (bits)  Value

gnl|PID|e252316  (Z74911) ORF YOR003w [Saccharomyces cerevisiae]    112  7e-26
gi|603258  (U18795) Prb1p: vacuolar protease B [Saccharomyces ce...  106  5e-24
gnl|PID|e264388  (X59720) YCR045c, len:491 [Saccharomyces cerevi...   69  7e-13
gnl|PID|e239708  (Z71514) ORF YNL238w [Saccharomyces cerevisiae]     30  0.66
gnl|PID|e239572  (Z71603) ORF YNL327w [Saccharomyces cerevisiae]     29  1.1
gnl|PID|e239737  (Z71554) ORF YNL278w [Saccharomyces cerevisiae]     29  1.5

 gnl|PID|e252316 (Z74911) ORF YOR003w [Saccharomyces cerevisiae]
           Length = 478

 Score =  112 bits (278), Expect = 7e-26
 Identities = 85/259 (32%), Positives = 117/259 (44%), Gaps = 32/259 (12%)

Query: 2   QSVPWGISRVQAPAAHNRG---------LTGSGVKVAVLDTGIST-HPDLNIRGG-ASFV 50
           +  PWG+ RV       G           G GV   VLDTGI  T H D   R    + +
Sbjct: 174 EEAPWGLHRVSHREKPKYGQDLEYLYEDAAGKGVTSYVLDTGIDTEHEDFEGRAEWGAVI 233

Query: 51  PGEPSTQDGNGHGTHVAGTIAALNNSIGVLGVAPSAELYXXXXXXXXXXXXXXXQGLE 110
           P     D NGHGTH AG I + +     GVA + ++                +G+E
Sbjct: 234 PANDEASDLNGHGTHCAGIIGSKH-----FGVAKNTKIVAVKVLRSNGEGTVSDVIKGIE 288
```

Multiple sequence alignment:

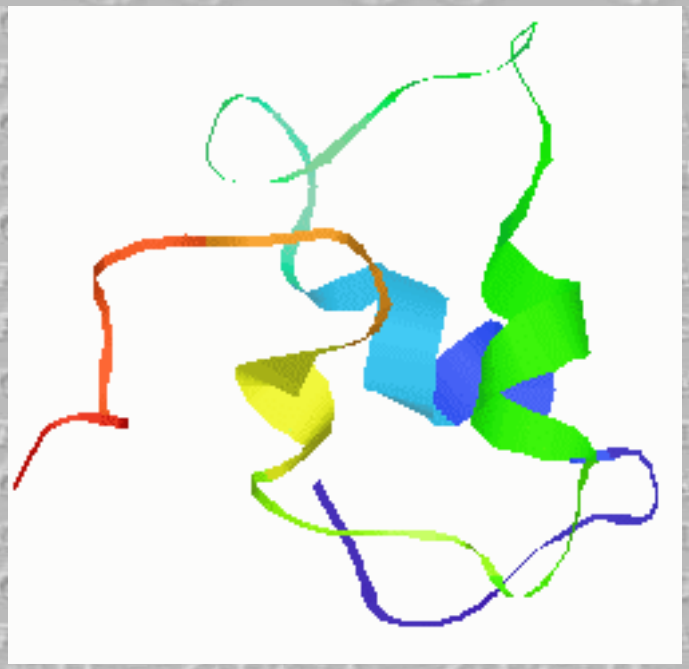
Sequences of proteins from different organisms can be aligned to see similarities and differences

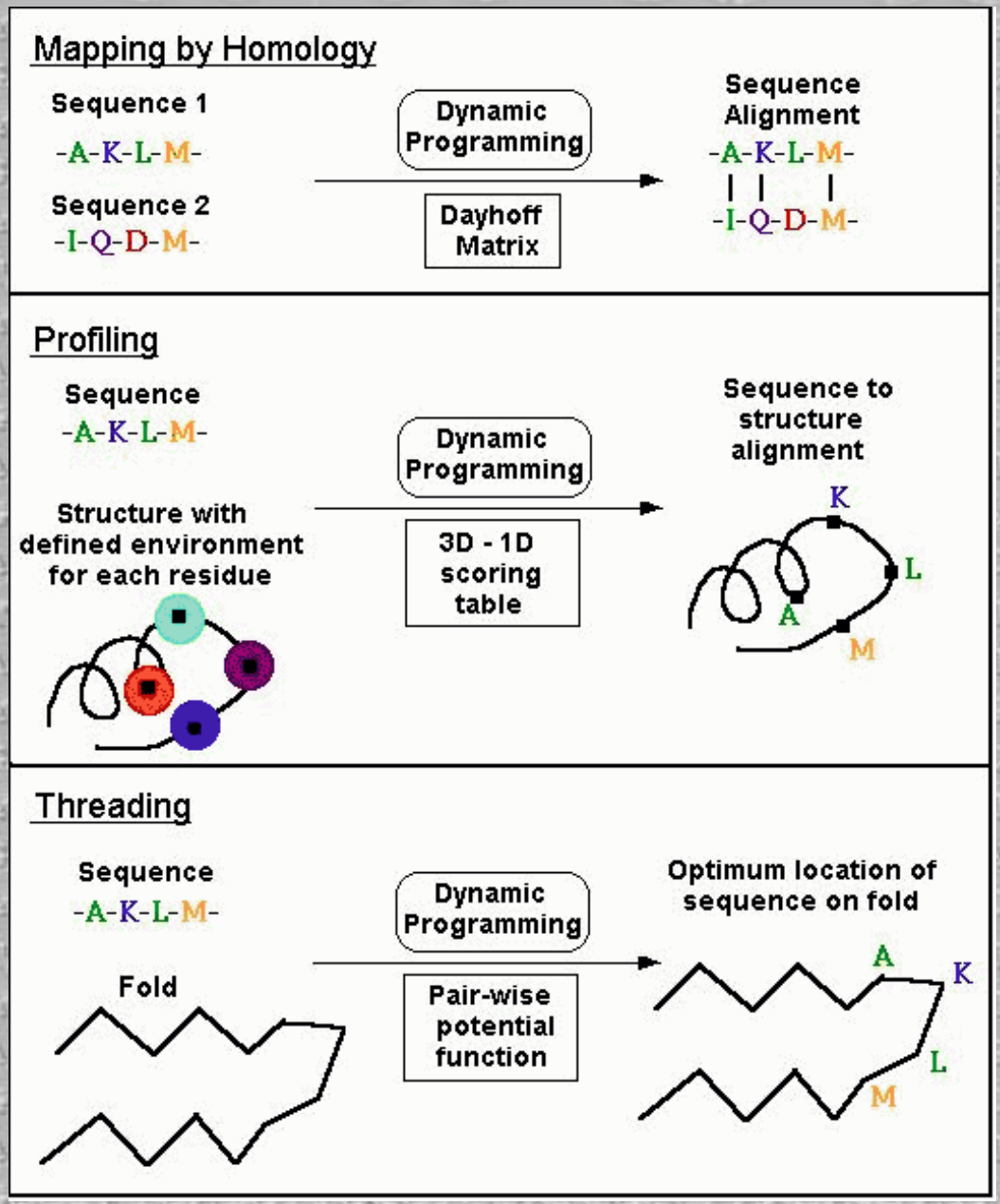
Phylogeny inference:

Analysis of sequences allows evolutionary relationships to be determined

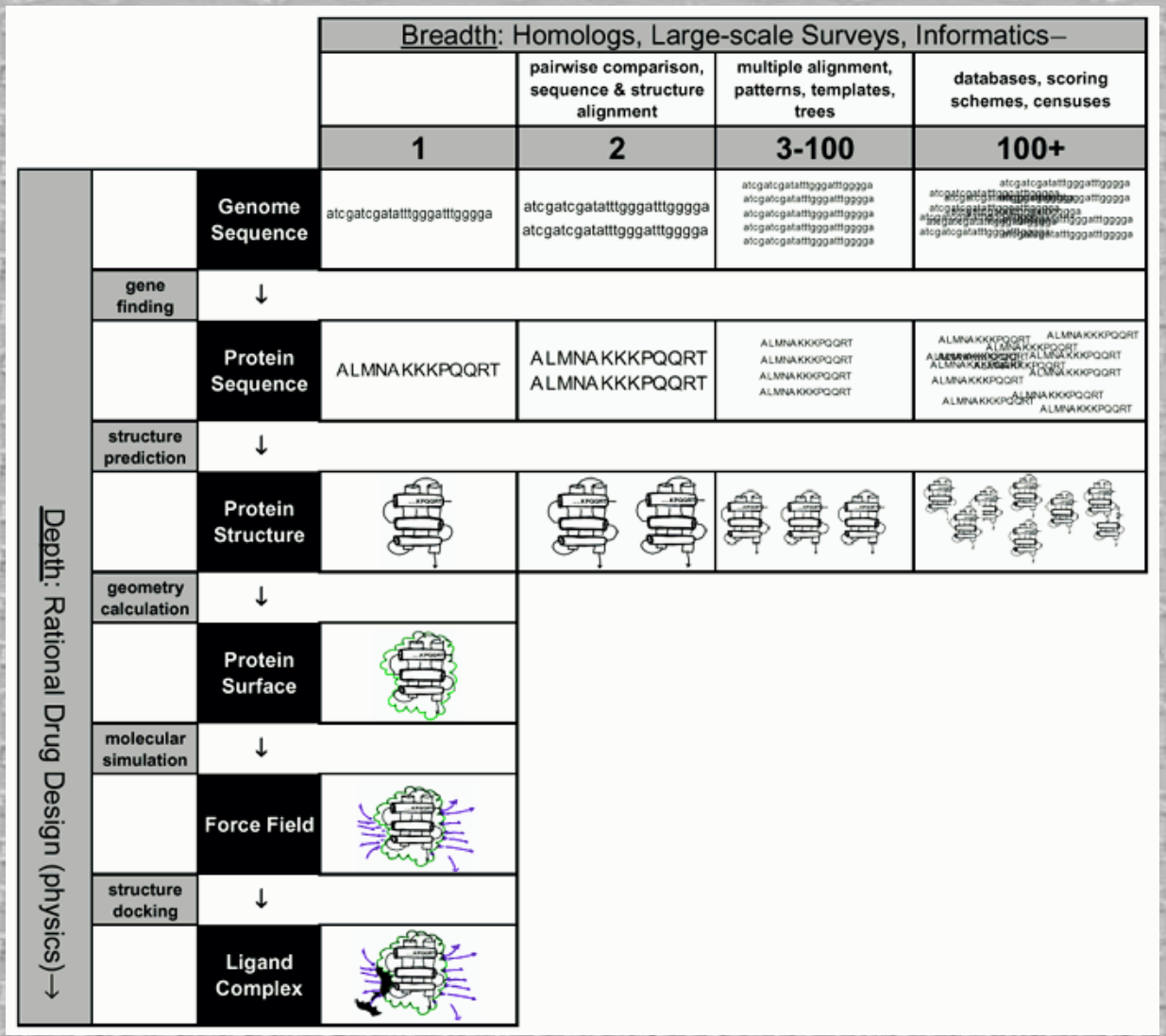
Protein Structure :

the 3-D structure of proteins is used to understand protein function and design new drugs



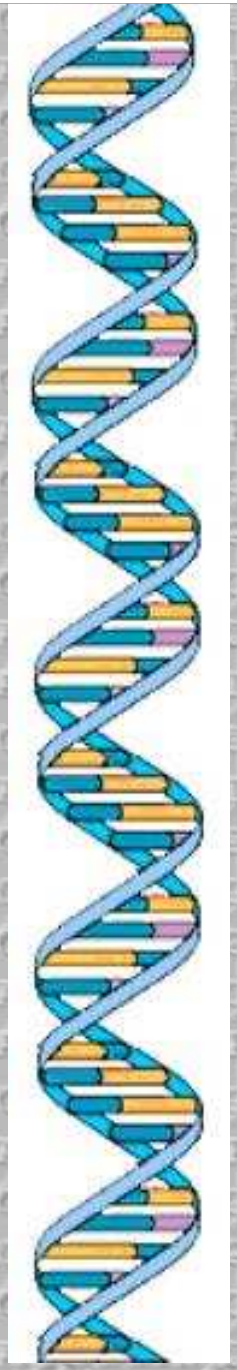


"…*UNDERSTAND* and organise the information…"



Challenges in bioinformatics

- Explosion of information
  - Need for faster, automated analysis to process large amounts of data
  - Need for integration between different types of information (sequences, literature, annotations, protein levels, RNA levels etc…)
  - Need for “smarter” software to identify interesting relationships in very large data sets
- Lack of “bioinformaticians”
  - Software needs to be easier to access, use and understand
  - Biologists need to learn about the software, its limitations, and how to interpret its results



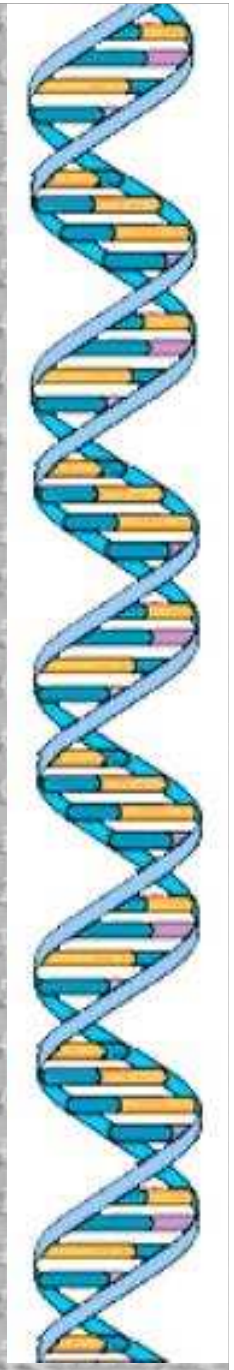
in silico Biology

in vivo, in vitro, …….. in silico

Bioinformatics is not just using a computer to store data, or speed up biology

With bioinformatics, you do biological

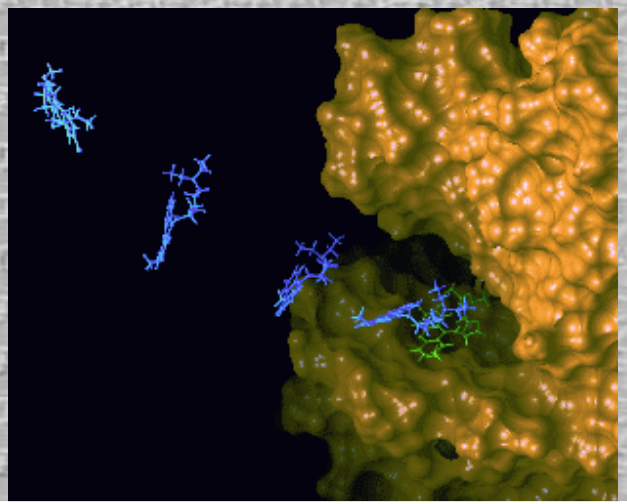
hypothesis testing on a computer

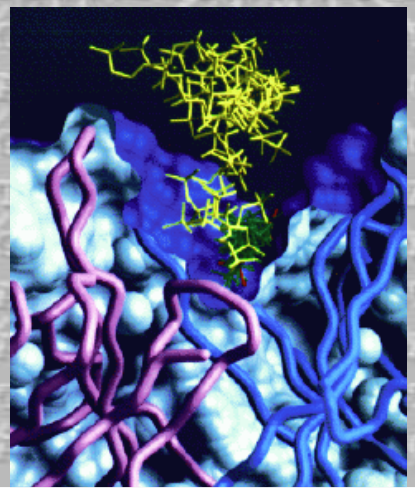


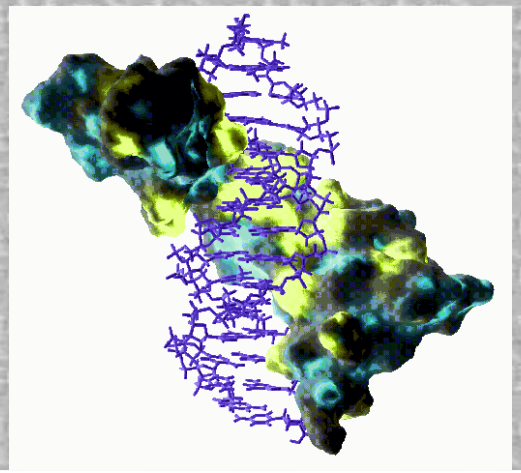
Major Application I: Designing Drugs

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

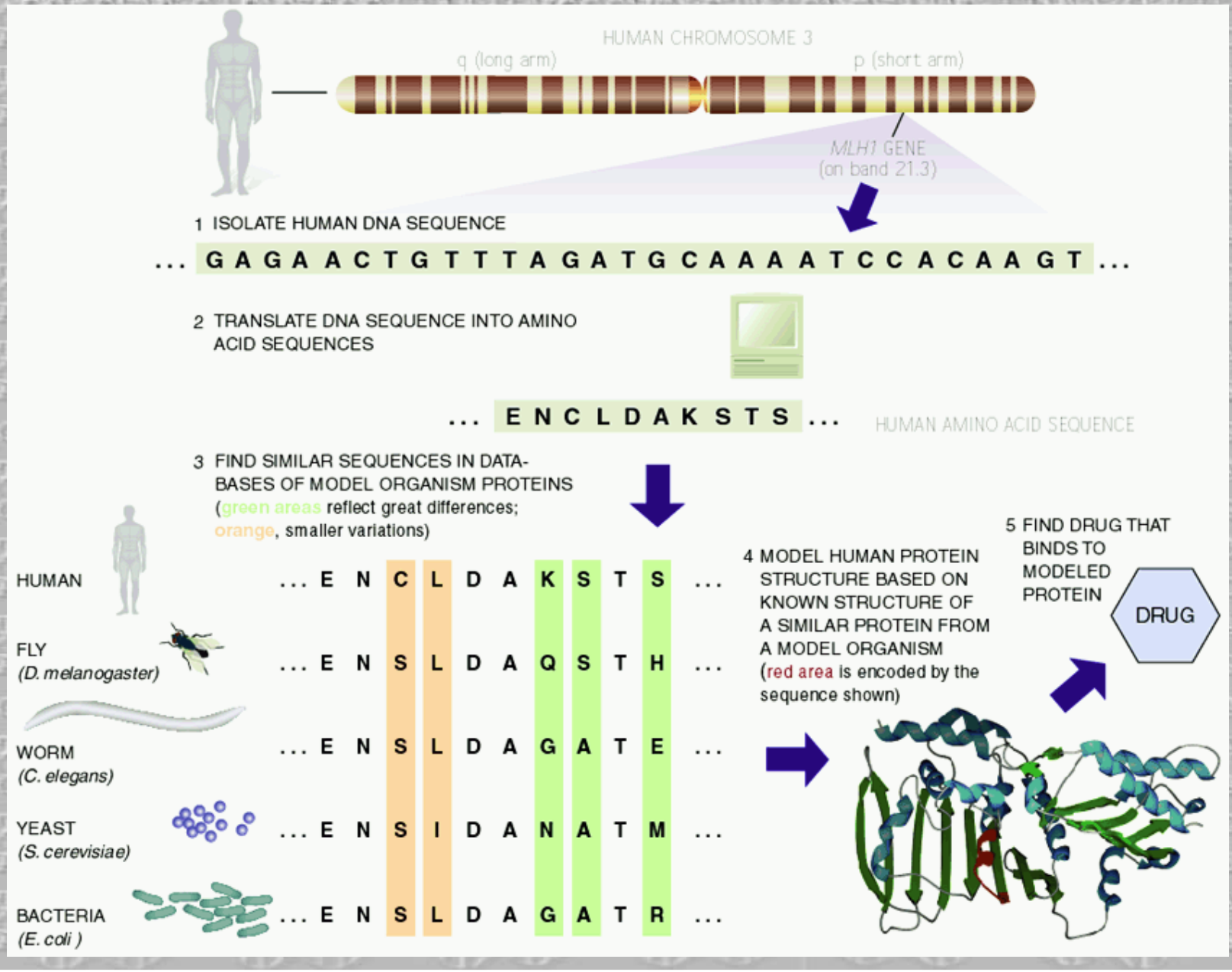
(From left to right, figures adapted from Olsen Group Docking Page at Scripps, Dyson NMR Group Web page at Scripps, and from Computational Chemistry Page at Cornell Theory Center).



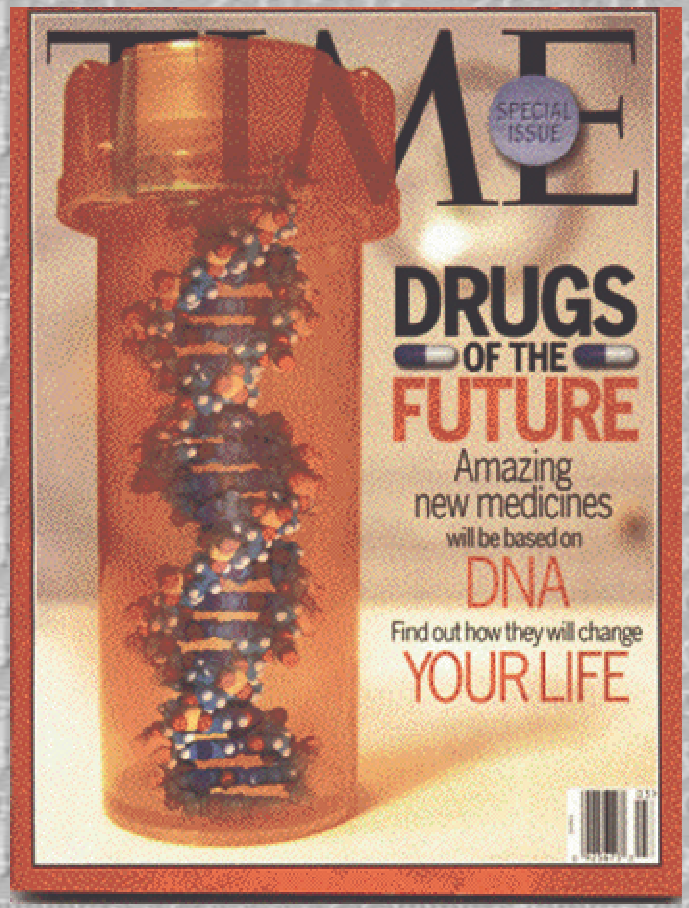




Major Application II: Finding Homologs



Pharmacogenomics



Everybody is different

The Right Drug

To The Right Patient

For The Right Disease

At The Right Time

Pharmaceutical Motivation

Speed up drug discovery pipeline

Drug Discovery 10,000 candidates
10 preclinical
5 clinical
1 New Drug

The goal is to weed out bad candidates quickly and cheaply. Approximate cost for a new drug = $150-800 Million

Some Challenges for Bioinformatics

- Precise, predictive model of transcription initiation and termination: ability to predict where and when transcription will occur in a genome
- Precise, predictive model of RNA splicing/alternative splicing: ability to predict the splicing pattern of any primary transcript in any tissue
- Precise, quantitative models of signal transduction pathways: ability to predict cellular responses to external stimuli
- Determining effective protein:DNA, protein:RNA and protein:protein recognition codes
- Accurate *ab initio* protein structure prediction

Some Challenges for Bioinformatics

- ✓ Rational design of small molecule inhibitors of proteins
- ✓ Mechanistic understanding of protein evolution: understanding exactly how new protein functions evolve
- ✓ Mechanistic understanding of speciation: molecular details of how speciation occurs
- ✓ Continued development of effective gene ontologies - systematic ways to describe the functions of any gene or protein
- ✓ Education: development of appropriate bioinformatics curricula for secondary, undergraduate and graduate education