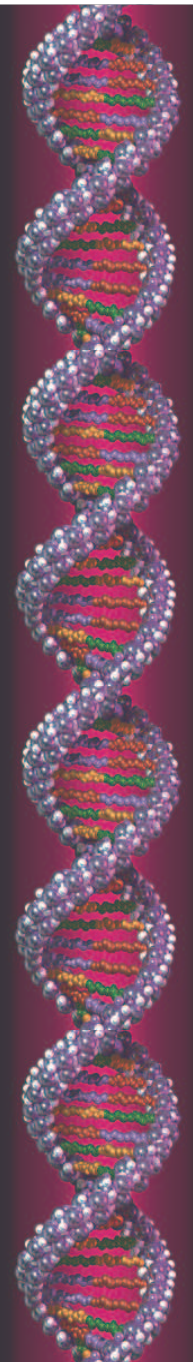
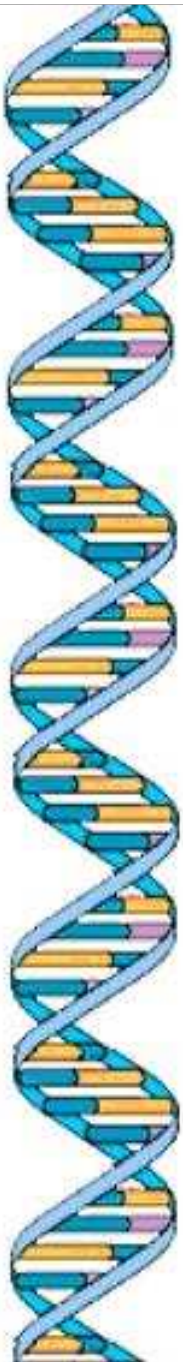


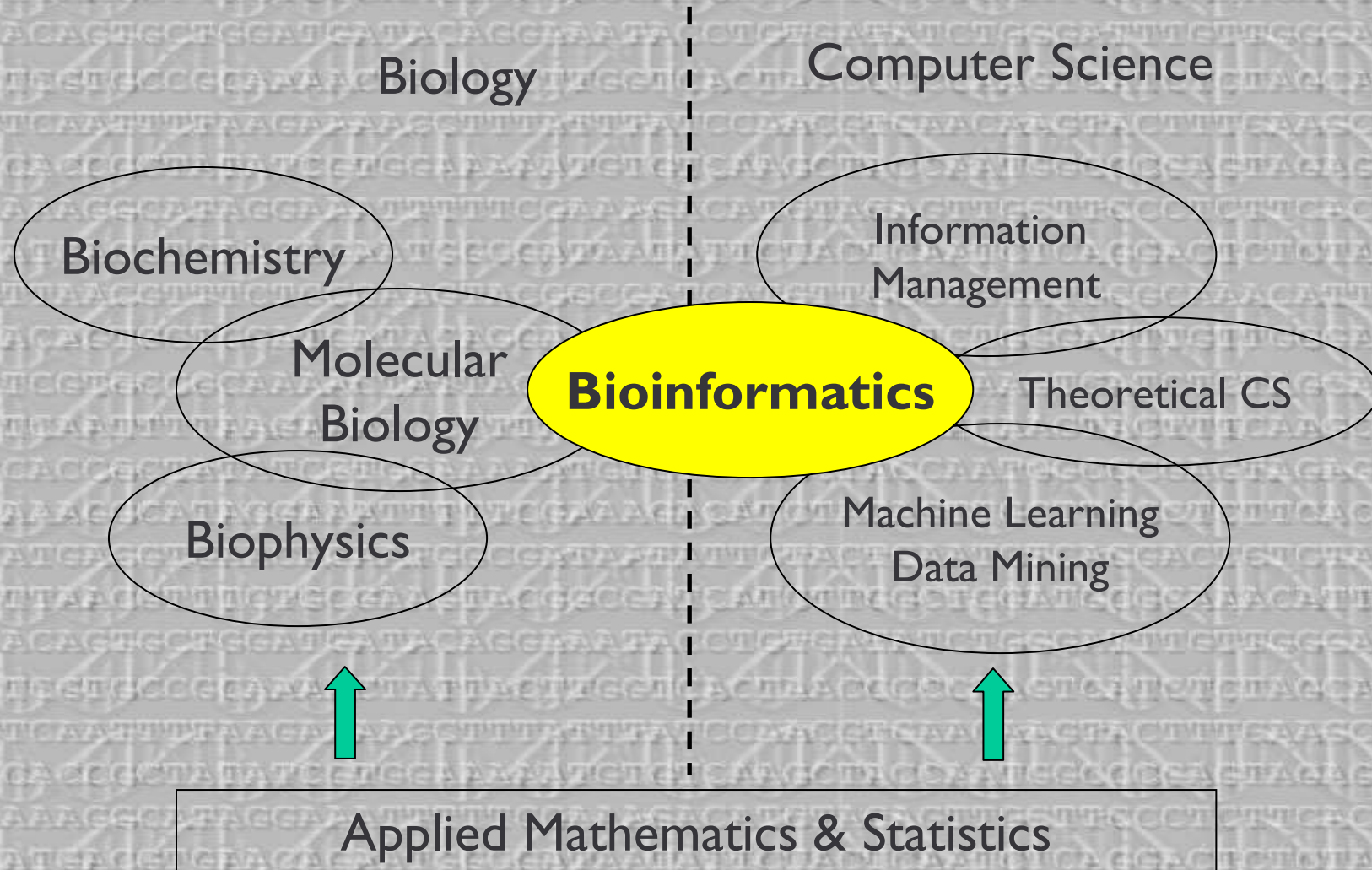
Bioinformatics

مهدی صادقی

پژوهشگاه ملی مهندسی ژنتیک و زیست فناوری



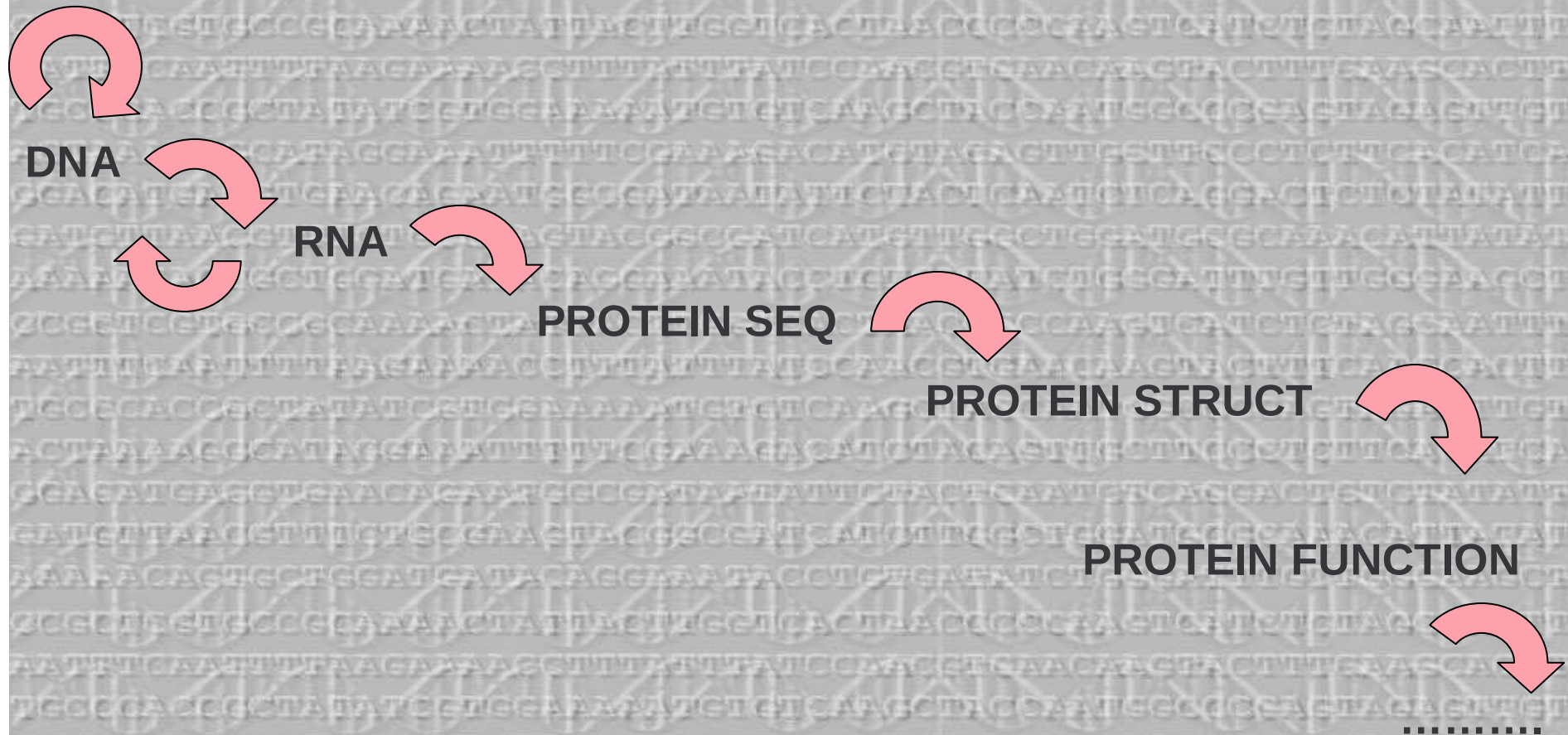
Bioinformatics is interdisciplinary



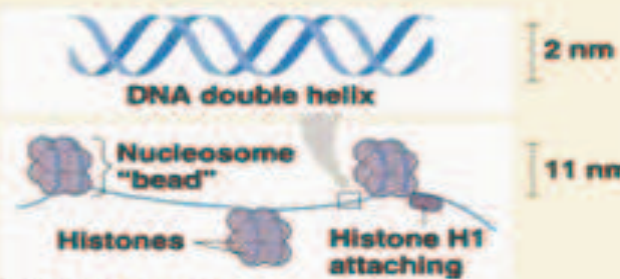
bio – informatics:

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*.

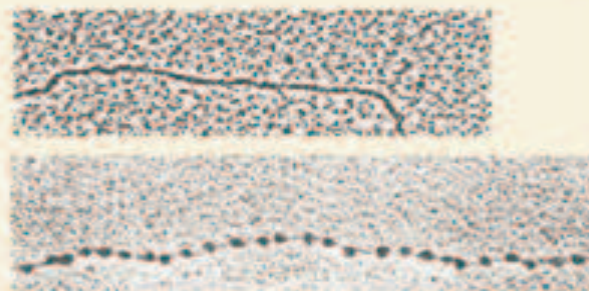
Flow of information



[illegible]



(a) Nucleosomes ("beads on a string")



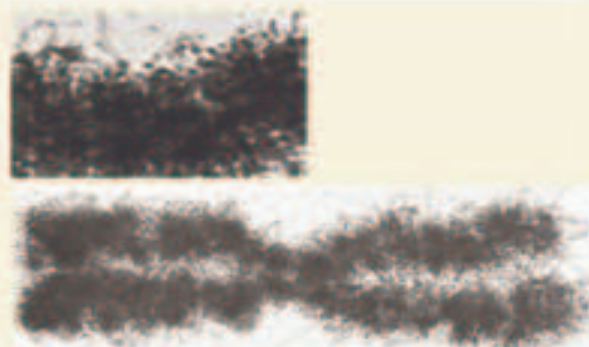
(b) 30-nm chromatin fiber



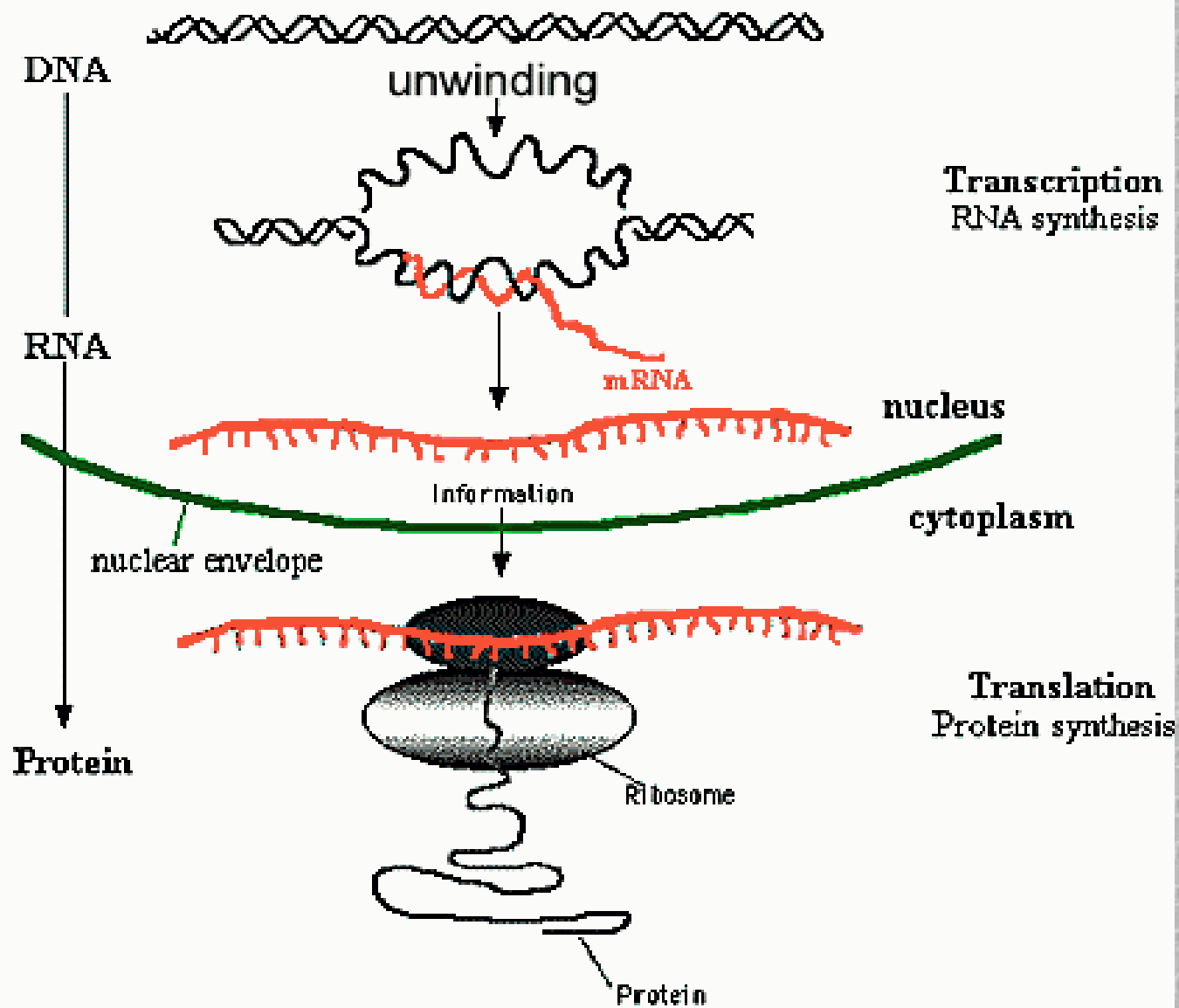
(c) Looped domains



(d) Metaphase chromosome



Central Dogma

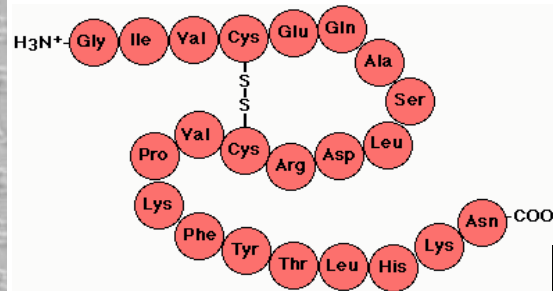


The Central Dogma of Molecular Biology

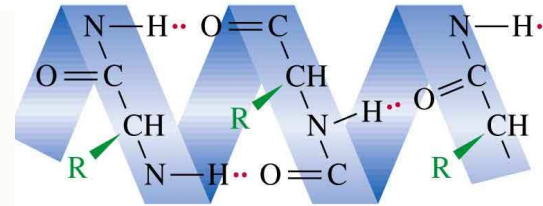
There Are Four Levels of Protein Structure

Primary: amino acid linear sequence.

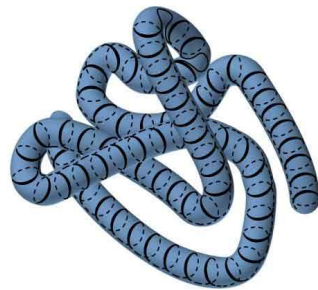
Secondary: α -helices, β -sheets and loops.



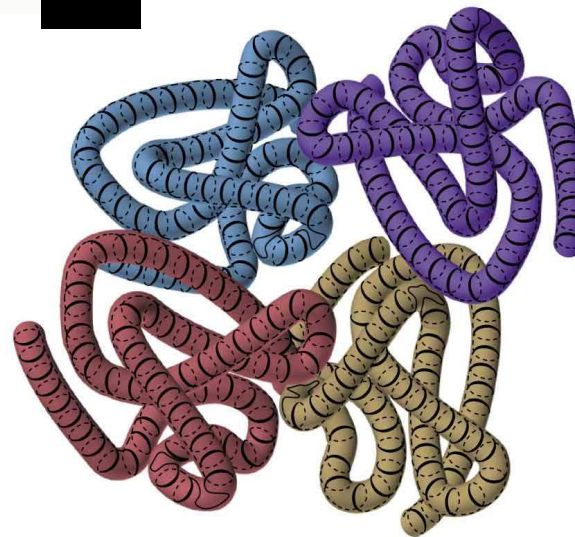
primary structure



secondary structure



tertiary structure



quaternary structure

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

Quaternary: arrangement of several polypeptide chains.

Early History

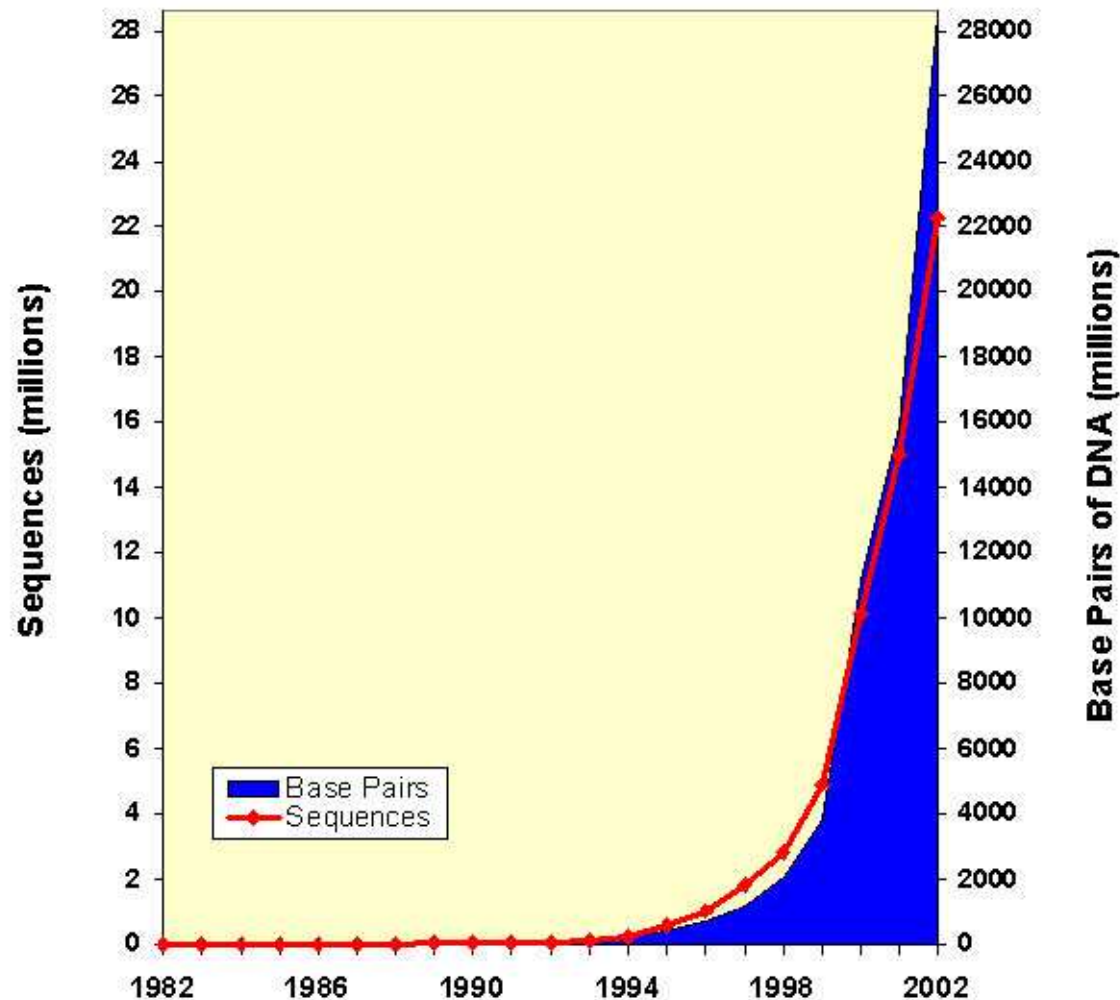
- First protein amino acid sequence database in mid 1960's.
- Researchers developing algorithms to analyze these data for individual research projects in 1960's and 1970's.
- GenBank and other public databases with freely available analysis tools in the 1980's
- Enormous growth of GenBank and PDB in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

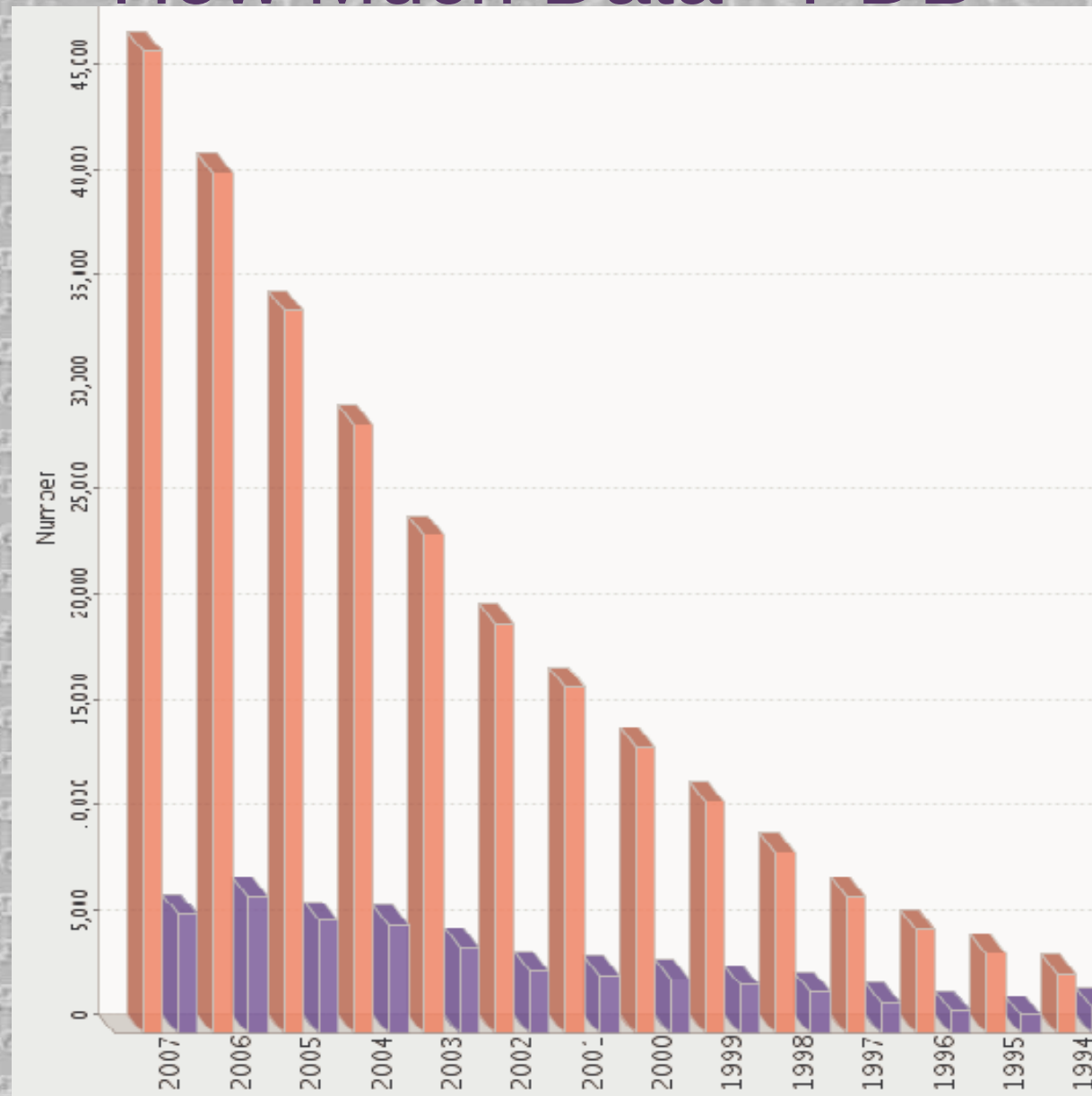
How Much Data - GenBank

Growth of GenBank



Source: NCBI

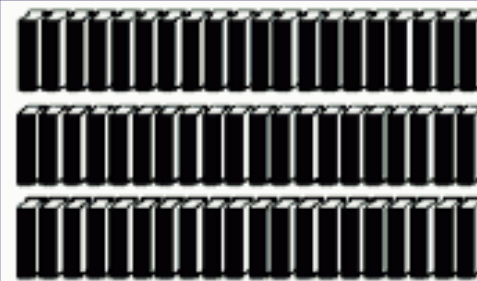
How Much Data - PDB



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

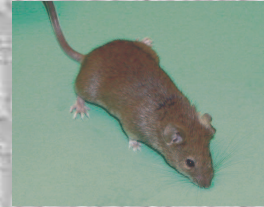


Pufferfish

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)



Oryza sativa (221,000)



Human Genome Sequencing Centers



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

[illegible]

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 PIR-International	www.expasy.ch/sprot/sprot-top.html www.mips.biochem.mpg.de/proj/protseqdb
Protein sequence (composite) OWL NRDB	www.bioinf.man.ac.uk/dbbrowser/OWL www.ncbi.nlm.nih.gov/entrez/queriv.fcgi?db=Protein
Protein sequence (secondary) PROSITE PRINTS Pfam	www.expasy.ch/prosite www.bioinf.man.ac.uk/dbbrowser/PRINTS/PRINTS.html www.sanger.ac.uk/Pfam/
Macromolecular structures Protein Data Bank (PDB) Nucleic Acids Database (NDB) HIV Protease Database ReLiBase PDBsum CATH SCOP FSSP	www.rcsb.org/pdb ndbserver.rutgers.edu/ www.ncifcrf.gov/CRYSDb/NEW_DATABASE www2.ebi.ac.uk:8081/home.html www.biochem.ucl.ac.uk/bsm/pdbsum www.biochem.ucl.ac.uk/bsm/cath scop.mrc-lmb.cam.ac.uk/scop www2.embl-ebi.ac.uk/dali/fssp
Nucleotide sequences GenBank EMBL DDBJ	www.ncbi.nlm.nih.gov/Genbank www.ebi.ac.uk/embl www.ddbj.nig.ac.jp/
Genome sequences Entrez genomes GeneCensus COGs	www.ncbi.nlm.nih.gov/entrez/queriv.fcgi?db=Genome bioinfo.mbb.yale.edu/genome www.ncbi.nlm.nih.gov/COG
Integrated databases InterPro Sequence retrieval system (SRS) Entrez	www.ebi.ac.uk/interpro www.expasy.ch/srs5 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
Genomes	complete genomes (1.6 million – 3 billion bases each)	Molecular simulations (force-field calculations, molecular movements, docking predictions) 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

Bioinformatics 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

```
graph TD; subgraph Nucleotide_Sequence_Analysis [Nucleotide sequence analysis]; SPM[Sequencing project management] --> NSF[Nucleotide sequence file]; NSD[Sequence database browsing] --> NSF; MSE[Manual sequence entry] --> NSF; NSF --> SDB[Search databases for similar sequences]; NSF --> SPCR[Search for protein coding regions]; SDB --> SC[Sequence comparison]; SPCR --> DFE[Design further experiments<br/>• Restriction mapping<br/>• PCR planning]; SPCR --> C[coding]; SPCR --> NC[non-coding]; C --> TIP[Translate into protein]; TIP --> PSF[Protein sequence file]; NC --> SKM[Search for known motifs]; NC --> RSP[RNA structure prediction]; PSF --> SDB2[Search databases for similar sequences]; PSF --> SKM2[Search for known motifs]; PSF --> PSS[Predict secondary structure]; SDB2 --> SC2[Sequence comparison]; SKM2 --> SC2; SKM2 --> PTT[Predict tertiary structure]; PSS --> PTT; SC --> MSA[Multiple sequence analysis]; SC2 --> MSA; PTT --> MSA; MSA --> CMSA[Create a multiple sequence alignment]; CMSA --> EAL[Edit the alignment]; EAL --> FAP[Format the alignment for publication]; EAL --> MP[Molecular phylogeny]; EAL --> PFA[Protein family analysis]; end; subgraph Protein_Sequence_Analysis [Protein sequence analysis]; TIP --> PSF; PSF --> SDB2; PSF --> SKM2; PSF --> PSS; end;
```

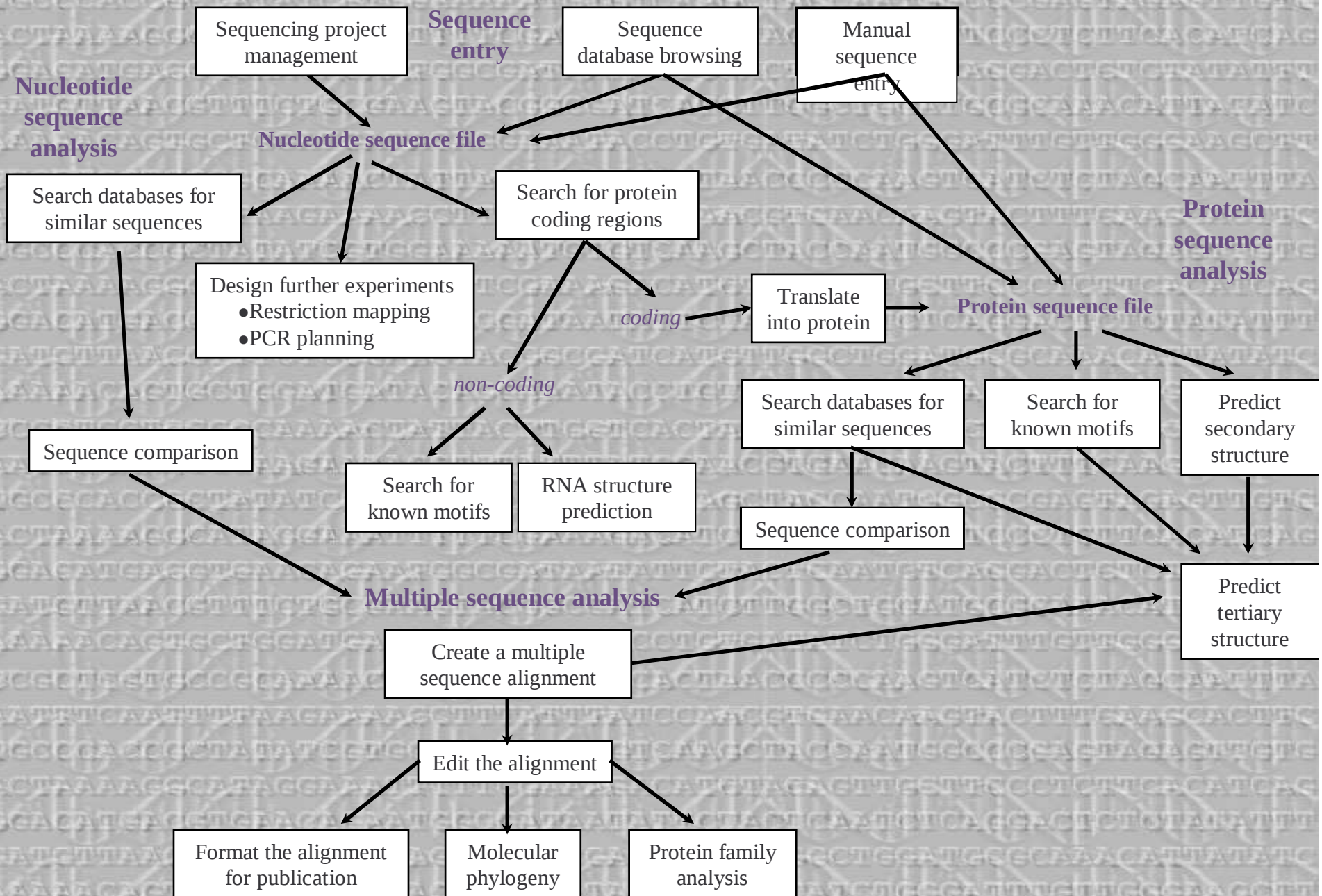
The flowchart illustrates the overview of sequence analysis, divided into Nucleotide sequence analysis and Protein sequence analysis.

Nucleotide sequence analysis:

- Sequencing project management leads to Nucleotide sequence file.
- Sequence database browsing leads to Nucleotide sequence file.
- Manual sequence entry leads to Nucleotide sequence file.
- Nucleotide sequence file leads to Search databases for similar sequences.
- Nucleotide sequence file leads to Search for protein coding regions.
- Search databases for similar sequences leads to Sequence comparison.
- Search for protein coding regions leads to Design further experiments (Restriction mapping, PCR planning).
- Search for protein coding regions leads to *coding* (Translate into protein).
- Search for protein coding regions leads to *non-coding* (Search for known motifs, RNA structure prediction).
- Translate into protein leads to Protein sequence file.
- Protein sequence file leads to Search databases for similar sequences.
- Protein sequence file leads to Search for known motifs.
- Protein sequence file leads to Predict secondary structure.
- Search databases for similar sequences leads to Sequence comparison.
- Search for known motifs leads to Sequence comparison.
- Search for known motifs leads to Predict tertiary structure.
- Predict secondary structure leads to Predict tertiary structure.
- Sequence comparison leads to Multiple sequence analysis.
- Sequence comparison leads to Predict tertiary structure.
- Predict tertiary structure leads to Multiple sequence analysis.
- Multiple sequence analysis leads to Create a multiple sequence alignment.
- Create a multiple sequence alignment leads to Edit the alignment.
- Edit the alignment leads to Format the alignment for publication.
- Edit the alignment leads to Molecular phylogeny.
- Edit the alignment leads to Protein family analysis.

Protein sequence analysis:

- Translate into protein leads to Protein sequence file.
- Protein sequence file leads to Search databases for similar sequences.
- Protein sequence file leads to Search for known motifs.
- Protein sequence file leads to Predict secondary structure.
- Search databases for similar sequences leads to Sequence comparison.
- Search for known motifs leads to Sequence comparison.
- Search for known motifs leads to Predict tertiary structure.
- Predict secondary structure leads to Predict tertiary structure.



Sequence comparison:

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

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

Comparing Genomes is Like Cryptography

CKQEBHEREYTWASUISCZMEISDFOGETHEBLPBGGOODFQSTLKSTUFFRTAC

DLUCEHEREZBRTTOISAWNDCDARJJPTHERROFGOODERGHCLSTUFFBRHA

Comparative Sequence Analysis

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

Species A

```
GATCGTCTAGAGTCTCGAGATC  
TCTCGAGATCTCTCGAGAGCTCT  
GTGATCTGAGGATTTAGCCACA  
CTTACCTCTCGAGAGATCTAGCA  
TCCGCTGAGCCGAGTTTCTACT  
CTCGAGGATCTGCTCGAGCTCA  
GAGGCTCGAGGCTCTCGAGCTCA  
TACCGAGATCTGAGATCTAGTAC  
AGAGTCTGAGGATCTAGCTTAC  
CTTCCAGGAGGATCTAGCTTAC  
AGCTCGAGGATCTAGCTTAC  
CTTCCAGGATCTAGCTTAC  
AGCTCGAGGATCTAGCTTAC  
AGCTCGAGGATCTAGCTTAC  
AGCTCGAGGATCTAGCTTAC  
AGCTCGAGGATCTAGCTTAC  
AGCTCGAGGATCTAGCTTAC
```

Species B

```
TATCGGCTAGAGTCTCGAGATC  
TCTCGAGATCTCTCGAGAGCTCT  
GTGATCTGAGGATTTAGCCACA  
CTTACCTCTCGAGAGATCTAGCA  
TCCGCTGAGCCGAGTTTCTACT  
CTCGAGGATCTGCTCGAGCTCA  
GAGGCTCGAGGCTCTCGAGCTCA  
TACCGAGATCTGAGATCTAGTAC  
AGAGTCTGAGGATCTAGCTTAC  
CTTCCAGGAGGATCTAGCTTAC  
AGCTCGAGGATCTAGCTTAC  
CTTCCAGGATCTAGCTTAC  
AGCTCGAGGATCTAGCTTAC  
AGCTCGAGGATCTAGCTTAC  
AGCTCGAGGATCTAGCTTAC  
AGCTCGAGGATCTAGCTTAC  
AGCTCGAGGATCTAGCTTAC
```

Compare



```
GATCGTCTAGAGTCTCGAGATC  
TCTCGAGATCTCTCGAGAGCTCT  
GTGATCTGAGGATTTAGCCACA  
CTTACCTCTCGAGAGATCTAGCA  
TCCGCTGAGCCGAGTTTCTACT  
CTCGAGGATCTGCTCGAGCTCA  
GAGGCTCGAGGCTCTCGAGCTCA  
TACCGAGATCTGAGATCTAGTAC  
AGAGTCTGAGGATCTAGCTTAC  
CTTCCAGGAGGATCTAGCTTAC  
AGCTCGAGGATCTAGCTTAC  
CTTCCAGGATCTAGCTTAC  
AGCTCGAGGATCTAGCTTAC  
AGCTCGAGGATCTAGCTTAC  
AGCTCGAGGATCTAGCTTAC  
AGCTCGAGGATCTAGCTTAC  
AGCTCGAGGATCTAGCTTAC
```

Sequences in Common (i.e., 'Conserved' or 'Constrained')

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  PGEPSTQDGNGHGHVAGTIAALNNSIGVLGVAPSAELYXXXXXXXXXXXXXXXXXXQGLE 110
          P      D NGHGH TH AG I + +      GVA + ++      +G+E
Sbjct: 234 PANDEASDLNGHGHHCAGIIGSKH----FGVAKNTKIVAVKVLRNNGEGTVSDVIKGIE 288
```

Multiple sequence alignment:

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

```
109Y:C_ss      10  -----EEE-----HHHHHH-----MMMMMM--M--MMMMMMMM--MMMMMM--M-- 83
109Y:C         10  PALDSLALDLTLRRCGELRLTLAELRRLDAGTILEVT-GISPG--HATLCHGEQVVAEGELVDVEG---RLGLQITRLVTR 83
EHEC_SepQ     232  LSNETVMNRIYCAIGTVHIDIHMLRNVKKDDIINSDDGYHLFG--GCQLIRNNTTIAYGSIVKINE---DFYFTVSIVCD- 305
Yp_YscQ       226  VQLEDLPQTLVMEIGRLTLPLGEIKQLAVGQTLACQ-THCYG--EVNICLNQGSVGRGSLRLCDE---KLVVRIAQWGLQ 299
Ct_YscQ       302  ATPLPSATKIVAEEVARYSLSVGEFLKLGPGSVLQFDGVHPTL--GVDIIILNGAKVGRGNIIALQD---VLGIRVLEV--- 373
At_AGRc988p   116  EMIMDIPIDVQIVLGTSRMLVSGIMSLEEGATIALD-RKIGE--PVEIMVNGRRIARGEITVLEDDDTFRFGVKLIEVLST 192
Ba_FliN       53  -TILNVPVNITVELGTSRIKIKDFLKFSGKSMLILN-ESIKE--PLNIFMNGHLIASGEIVVLEE---KYGIRITNIKNS 125
Ci_FliN       33  APVFDVPVNISAVLGRANMSVAQLLQLGQGSILELD-RKVGE--AIDIYVNNRLVARGEVNVVDE---RLGVTMTETIKD 106
Cc_FliN       22  SQVNAVNVESISVLLGRSILPMQQLLRMGRGAVIPLD-AKTHD--EVWILANNHPIARGEIQISDD---RIAIQVTRAADV 95
Cj_FliY       199  NLIMDVRLPVRVRIGNKKMLLKDVLTMDIGSVVELN-QLAND--PLEILIGDKRIAYGEVVIVDG---NFGVQITEIGSK 272
Cj_FliN       18  EDILDITVDFVSELGTTNMSVAELLKLEVGSVIDLE-KPAGE--SVELYINKRIFGKGEVMVYEK---NLAIRINEILDS 91
EHEC_EpaO     254  ----ELPVKIEFVLGKKIMNLYEIDELCAKRIISLL-PESEK--NIEIRVNGALTGYGELVEVDD---KLGVEIHSWLSG 323
Pa_PscQ       233  HELDQLPIPVSEFEVGRRTLDLHTLSTLQPGSLLDLD-SALDG--EVRILANQRCLGIGELVRLQD---RLGVRVTRLFGH 306
Rs_HrcQ       276  VPLEQLEVPVHLELAVMGMPALAEALAQPHVLTLPVKIRDV--SVRLVCHGQTLGHGQLVAVGE---QLGLQIASIGKH 350
St_SsaQ       245  VELEQIPQQVLFVVGGRASLEIGQLRQLKTGDVLPVG-GCFAP--EVTIRVNDRIIGQGELIACGN---EFMVRITRWYLC 318
St_SpaO       225  PGLNQPLPVKLEFVLYRKNVTLAELEAMGQQQLSLP-TNAEL--NVEIMANGVLLGNSELVQMND---TLGVEIHEWLSE 298
Yp_FliN       57  SLFSRIPVTLTLEVASVEIPLSELLTVNNDSDVIELD-KLAGE--PLDIRVNGIMFGQAEVVVINE---KYGLRIININSQ 130
Aa_FliN       37  QHFS DIPVEVEVVVGRANKTLGELLAMGIGSVIEID-REPKD--LVDIKVNGKLIAGKELVIIDG---KIGVKIKEVVKE 110
Li_Lin0701    28  RQVDNIGVNLIVRLGKKEMPVGDIAELSIGDVLEVE-KKPGH--KVEIFLDEKKVIGEGAILMDE---NFGIVISEID-- 99
Li_Lin0706    2  KINHTIPLRIDFELGRTKQPVGSLLDVKKGTVFRLE-DSTGN--VVKITISGKCIGYGEILTKDG---KMFVKITKLGE 75
Li_Lin0708    441  QILEDIPVTLEVVFGTAKVKLEKFIWCEKDVIILK-ESMNE--PLVLALNGVTIGKGILVRVDD---HFGIQMTELVR- 513
Ml_HrcQ       279  EDLDDVEIMLVFECGRWPIPLGELRSAGEGHIFELG-RPIQD--PVDILANGQCIGRGDIVRIGD---TLGIRLRGRRLGC 352
At_FliM       237  QVKR-SQVTLEARIKLETTLTTLRTISRLVAGDVIPFQ-DLKQDDIGVEVSANGSKLYNCEFGKSGD---RYMVRVKNNVST 311
Hp_FliM       252  AILS-IPLTLSARLCEPEVPLRQIMQMQPGDVLPVH-LTEAL----SLLVEGQPIFEAAPGERGG---QAALNLTRRHVR 322
Vp_FliM       250  EEIMDCPVNFRVNLLKDISLRDIMELQPGDIPIE-MPEHA----TMFIEDLPTYRVKMRSED---KLAVQVSQEI 321
Tp_FliM       253  KRIMTAQIPVVAELGTSELTIEEFLSLEVGDCTILD-KSVTD--PLTVLVGDKPKFLGQAGRVNR---KQAVQILDHDIR 326
Bb_FliM       263  EKLENTAMPLVAEIGEVKLKVREILSLDKGDVNLN-SSLINK-DLTLKVGTKEKFKCRMGLMGN---KVSQVITEKIGD 337
Tp_FliM       252  DKLSTVDMDVVAEVSGLRLSVRDILGLRVGDIIRLH-DTHVGD-PFVLSIGNRKKFLCQPGVVGK---KIAAQILERIES 326
Rp_FliM       252  ALLSGVSDMIVFLGAVELSLKEMLDLDVGDITRLN-KIAND--EVSVVYHKKKRYLASVGFQGY---RKTIQIKEVIYS 325
Sf_SpaO       219  FNYDDINVKVDFILLEKNMTINELKMYVENELFKFP-DDIVK--HVNIVNGSLVGHGELVSIED---GYGIEISSWMVK 292
Bp_SctQ       231  TRIGELELPVQFEIDTVSLPIDQLSALEPGYVIELP-VAVTD-ARLRLVVHGQTVGYGELVAVGE---HLGVRIIRMAHR 305
Ea_HrcQ       259  PQLASPLSLEVRCRDTALTGLGRLQAGSVVTLN-NVTPG--EAGLYHGDTLIARGELVVDVEG---HLGLQLTQLLLT 332
Consensus/80% ..h.pl.h.l.h.lsp.pb.l.plbpb..spll.l.....hpl.hssp.hhbtphh.....pbslpl.p.h..
```

Phylogeny inference:

Analysis of sequences allows evolutionary relationships to be determined

E.coli

C.botulinum

C.cadavers

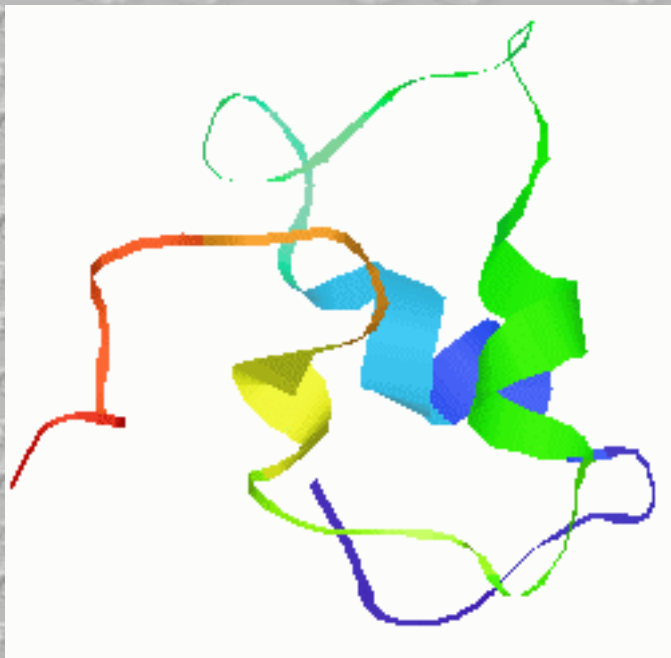
C.butyricum

B.subtilis

B.cereus

Protein Structure :

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



Mapping by Homology

Sequence 1

-A-K-L-M-

Sequence 2

-I-Q-D-M-

Dynamic Programming

Dayhoff Matrix

Sequence Alignment

-A-K-L-M-

| | |

-I-Q-D-M-

Profiling

Sequence

-A-K-L-M-

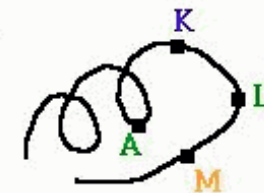
Structure with defined environment for each residue



Dynamic Programming

3D - 1D scoring table

Sequence to structure alignment



Threading

Sequence

-A-K-L-M-

Fold



Dynamic Programming

Pair-wise potential function

Optimum location of sequence on fold



“...**UNDERSTAND** and organise the information...”

		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	atc gatc gatattgggattgggga	atc gatc gatattgggattgggga atc gatc gatattgggattgggga	atc gatc gatattgggattgggga atc gatc gatattgggattgggga atc gatc gatattgggattgggga atc gatc gatattgggattgggga atc gatc gatattgggattgggga	
	gene finding	↓				
		Protein Sequence	ALMNAKKKPPQRT	ALMNAKKKPPQRT ALMNAKKKPPQRT	ALMNAKKKPPQRT ALMNAKKKPPQRT ALMNAKKKPPQRT ALMNAKKKPPQRT ALMNAKKKPPQRT	
	structure prediction	↓				
		Protein Structure		 	   	
	geometry calculation	↓				
		Protein Surface				
	molecular simulation	↓				
		Force Field				
	structure docking	↓				
		Ligand Complex				

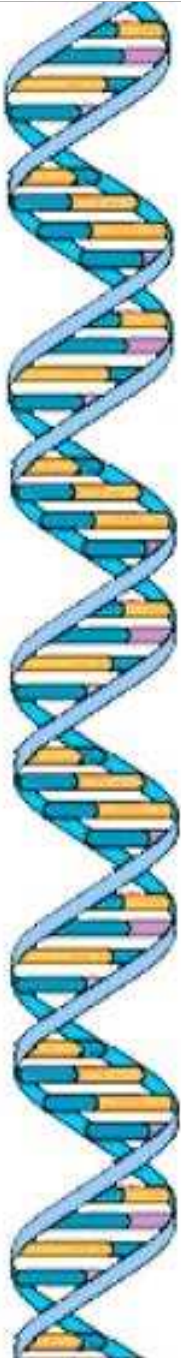
Challenges in bioinformatics

- **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

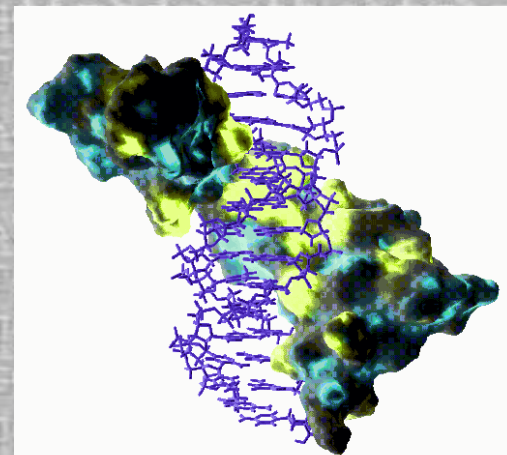
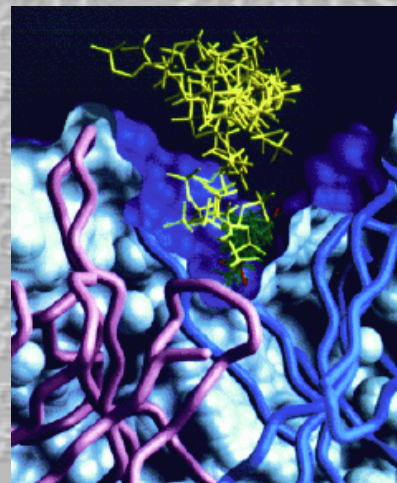
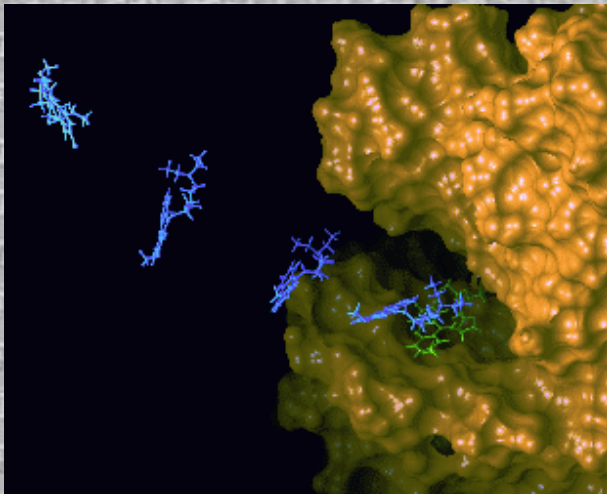
Bioinformatics is not just using a computer
to store data, or speed up biology

With bioinformatics, you do biological
hypothesis testing on a computer

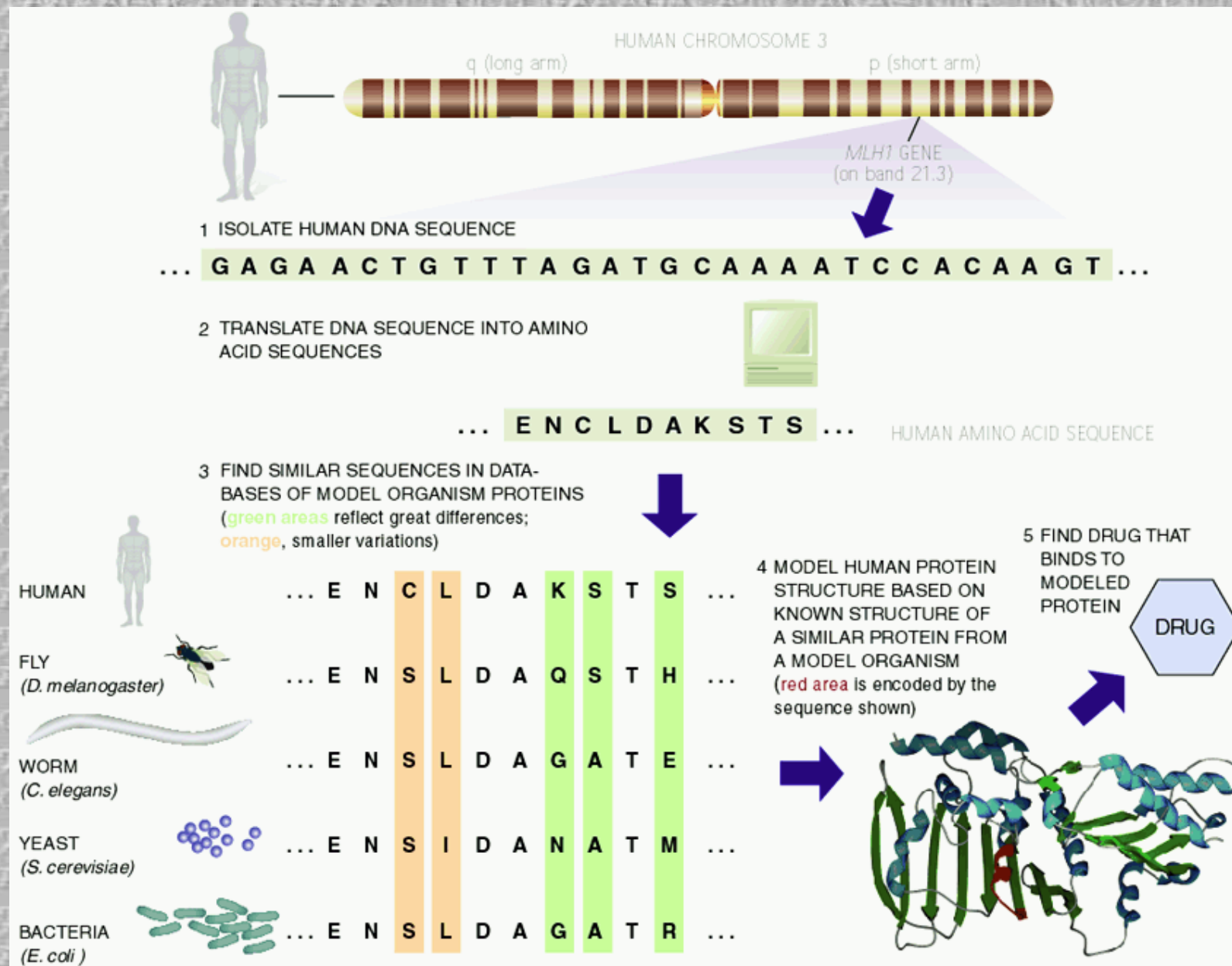
Major Application I: Designing Drugs

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

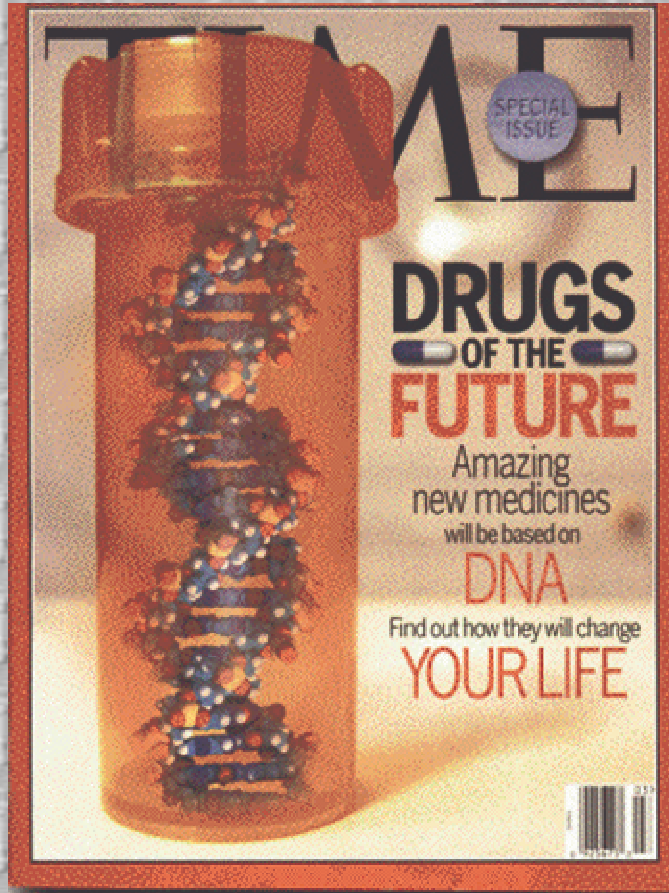
(From left to right, figures adapted from Olsen Group Docking Page at Scripps, Dyson NMR Group Web page at Scripps, and from Computational Chemistry Page at Cornell Theory Center).



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

