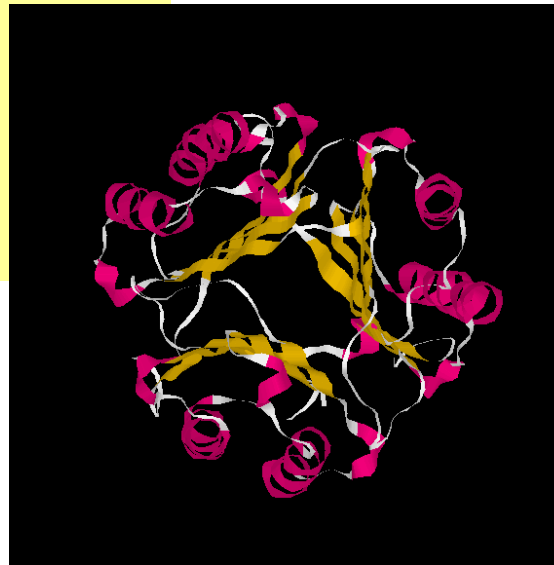
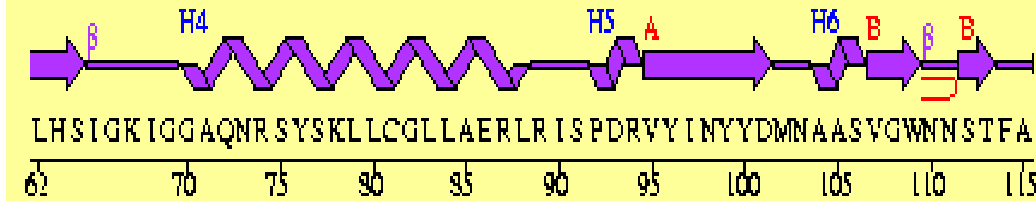
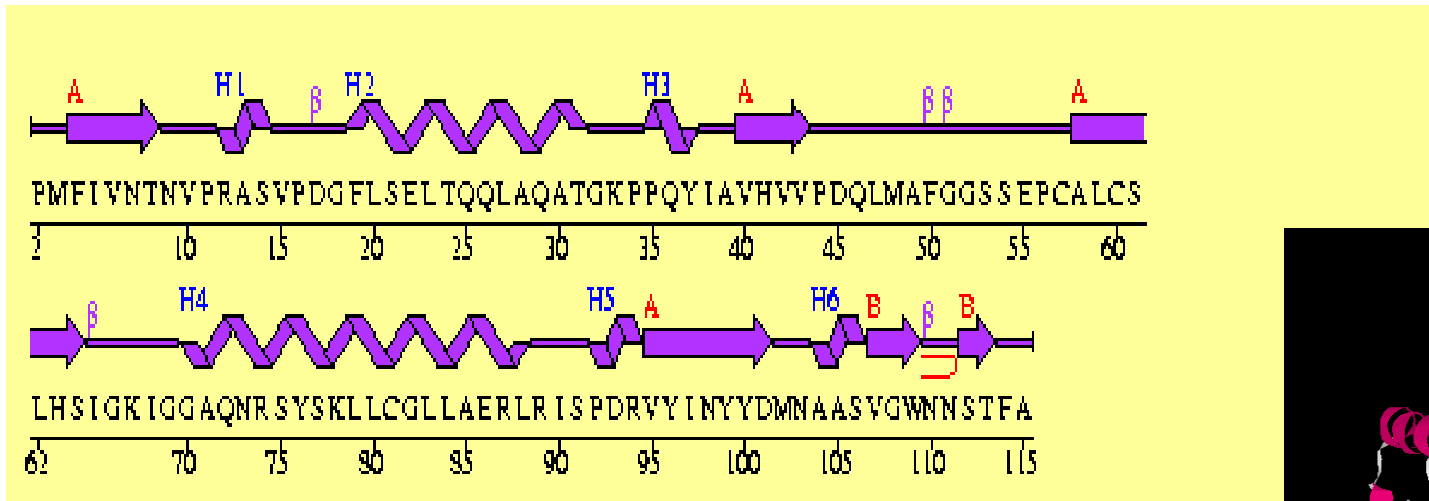


BIOL312

Bioinformatics and Computational Biology

Fezel Nizam



What is bioinformatics?

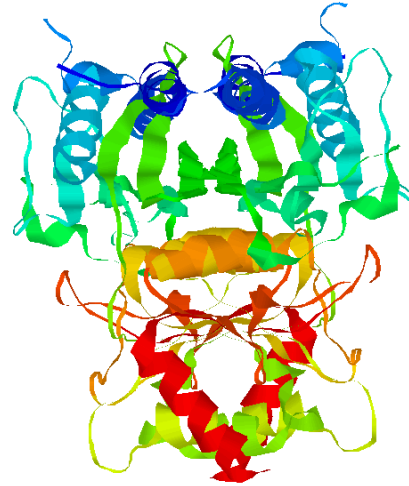
- an emerging interdisciplinary research area
- deals with the computational management and analysis of biological information:
 - genes,
 - genomes,
 - proteins,
 - cells,
 - ecological systems,
 - medical information,
 - robots,
 - artificial intelligence...

The Core of Bioinformatics to date

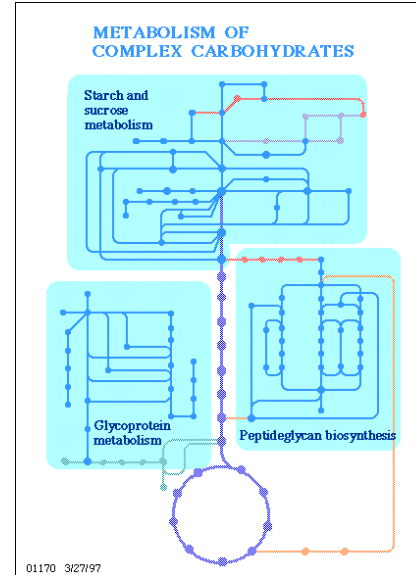
•Relationships between:

TDQAAFDTNIVTLTRFVM
EQGRKARGTGEMTQLLNS
LCTAVKAISTAVRKAGIA
HLYGIAGSTNVTGDQVKK
LDVLSNDLVINVLKSSFA
TCVLVTEEDKNAIIVEPE
KRGKYVVCFDPLDGSSNI
DCLVSI GTIFGIYRKNST
DEPSEKDALQPGRNLVAA
GYALYGSATMLV

sequence



3D structure



protein functions

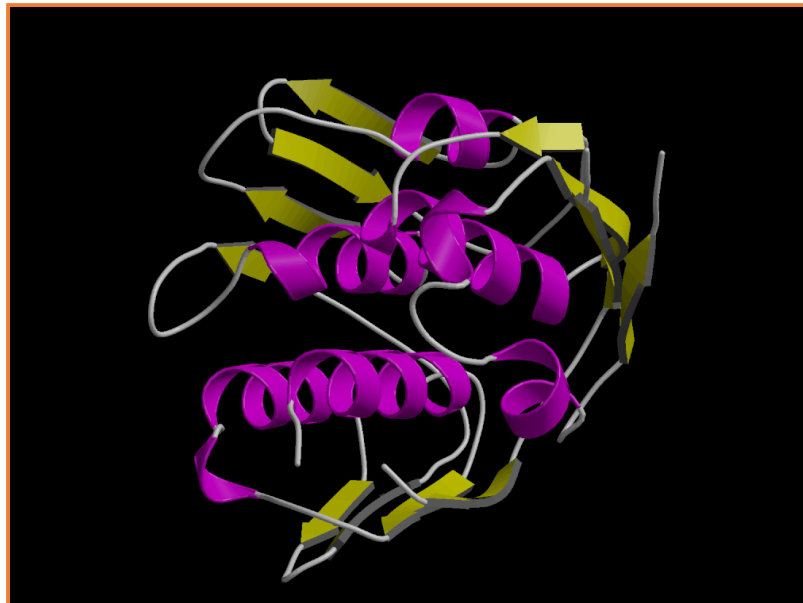
•Properties and evolution of genes, genomes, proteins, metabolic pathways in cell

•Use of this knowledge for prediction, modelling, and design

“The holy grail of bioinformatics”

```
GCTCCTCACTGTCTGTGTTTATTCTTTTAGCTTCTTCAGA  
TCTTTTAGTCTGAGGAAGCCTGGCATGTGCAAATGAAG  
TTAACCTAA...
```

> 500, 000 genes
sequenced to date



Expected number of unique
protein structures:

~ 700-1, 000

Basic concepts

- **conceptual foundations of bioinformatics:**
 - evolution**
 - protein folding**
 - protein function**
- **bioinformatics builds mathematical models of these processes -**
 - to infer relationships between components of complex biological systems**

What is bioinformatics?

- Two perspectives;
 - A set of tools & techniques to support biological science
 - Equivalent in scope to new assay methodology or new investigative techniques
 - A science that supports the systematic development and analysis of such tools
 - Investigation of the set of scientific principles forming the foundation for successful bioinformatics applications

Background of Students

- Computer & information science
 - Need to understand existing tools, scientific approach, and needs of biological research
- Biomedicine
 - Need to learn a set of tools and skills
 - May also need to understand the deeper scientific issues

The Gap I

- Biological scientists and investigators can't build their own tools
- Computer scientists don't know what tools to build

The Gap II

- Putting a biological investigator and a system implementer together in a room doesn't solve the problem
 - Barriers include:
 - Language
 - Methodology
 - Conceptualization

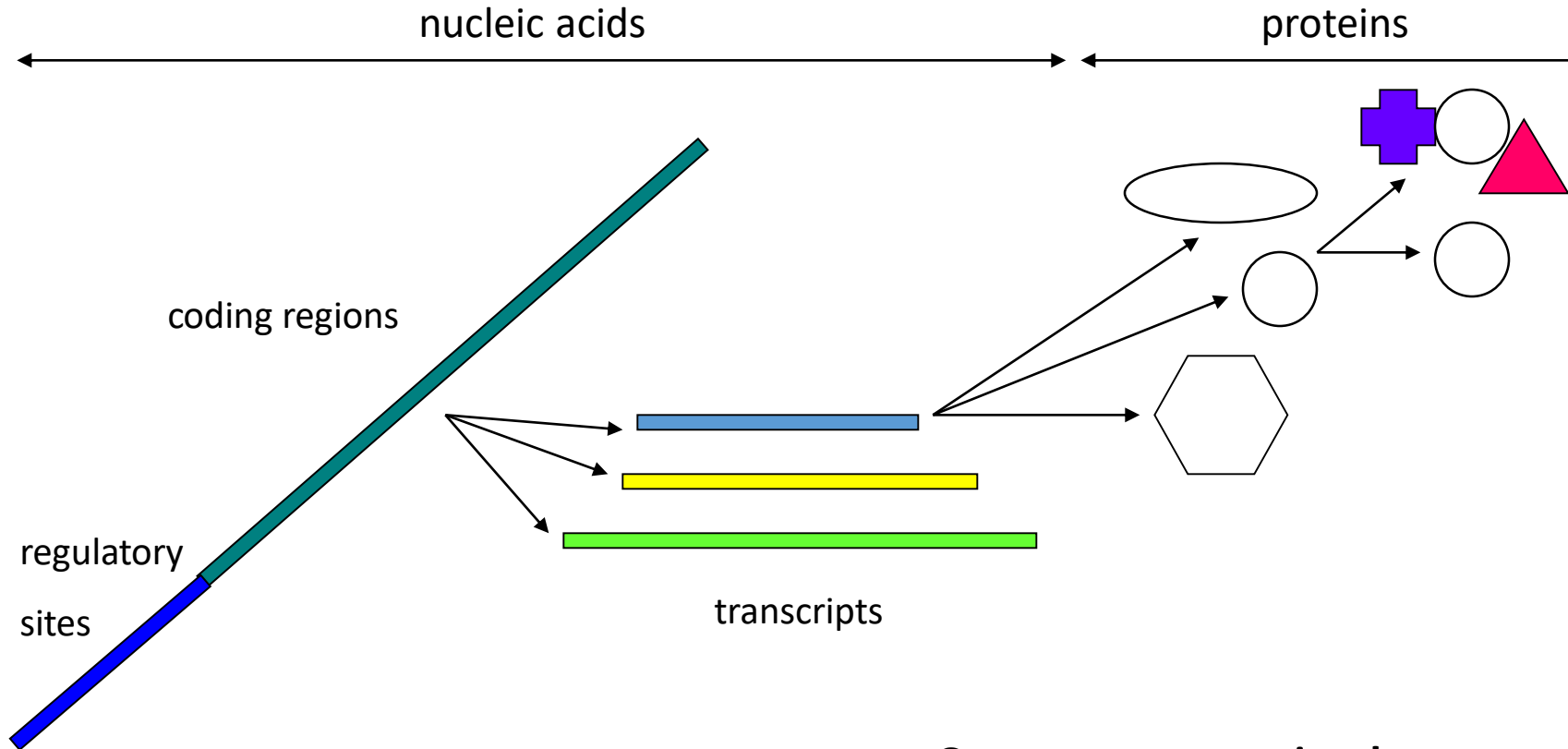
The Gap III

- Computer science is a “science of the artificial”
 - Mainly concerned with human artifacts i.e. creations limited mainly by conceptualization and imagination
- Biomedicine is a science of discovery
 - Mainly concerned with how organisms function, the limiting factors are often a result of limits of investigative methods and tools

- Large databases that can be accessed and analyzed with sophisticated tools have become central to biological research and education.
- The information content in the genomes of organisms, in the molecular dynamics of proteins, and in population dynamics, to name but a few areas, is **enormous**.
- Biologists are increasingly finding that the management of complex data sets is becoming a bottleneck for scientific advances.
- Therefore, **bioinformatics** is rapidly become a key technology in all fields of biology.

- **Molecular Bioinformatics** involves the use of computational tools to discover new information in complex data sets from the
- **one-dimensional** information of DNA through the
- **two-dimensional** information of RNA and the
- **three-dimensional** information of proteins,
- to the **four-dimensional** information of evolving living systems

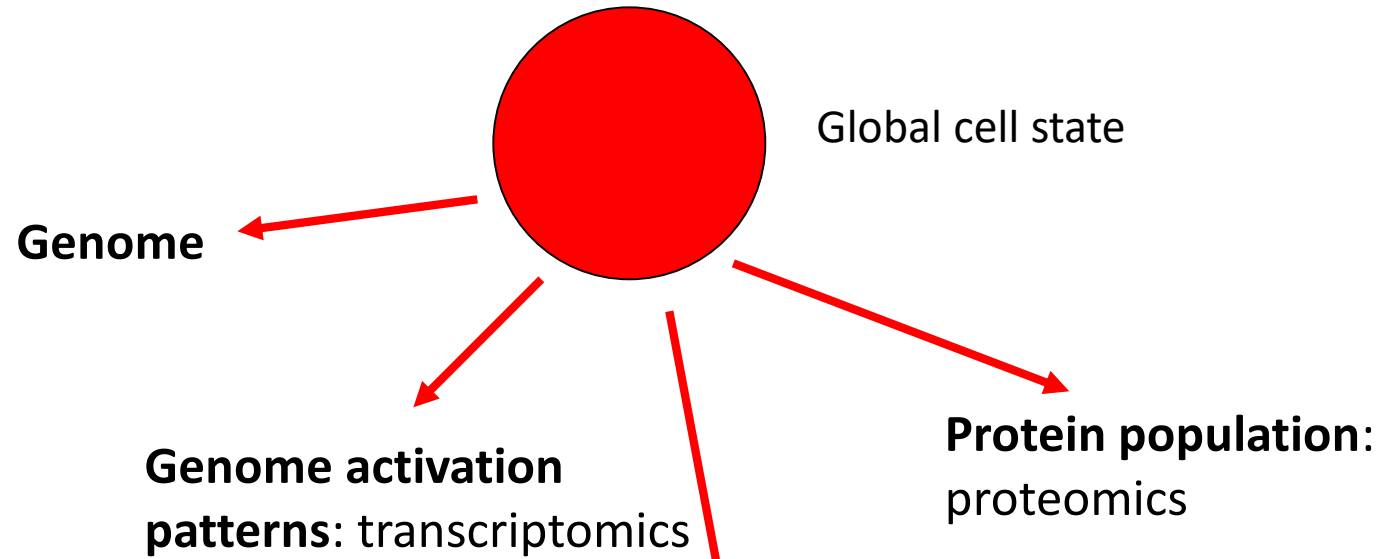
Information processing in cells



One-to-many mappings!

Context-dependence!

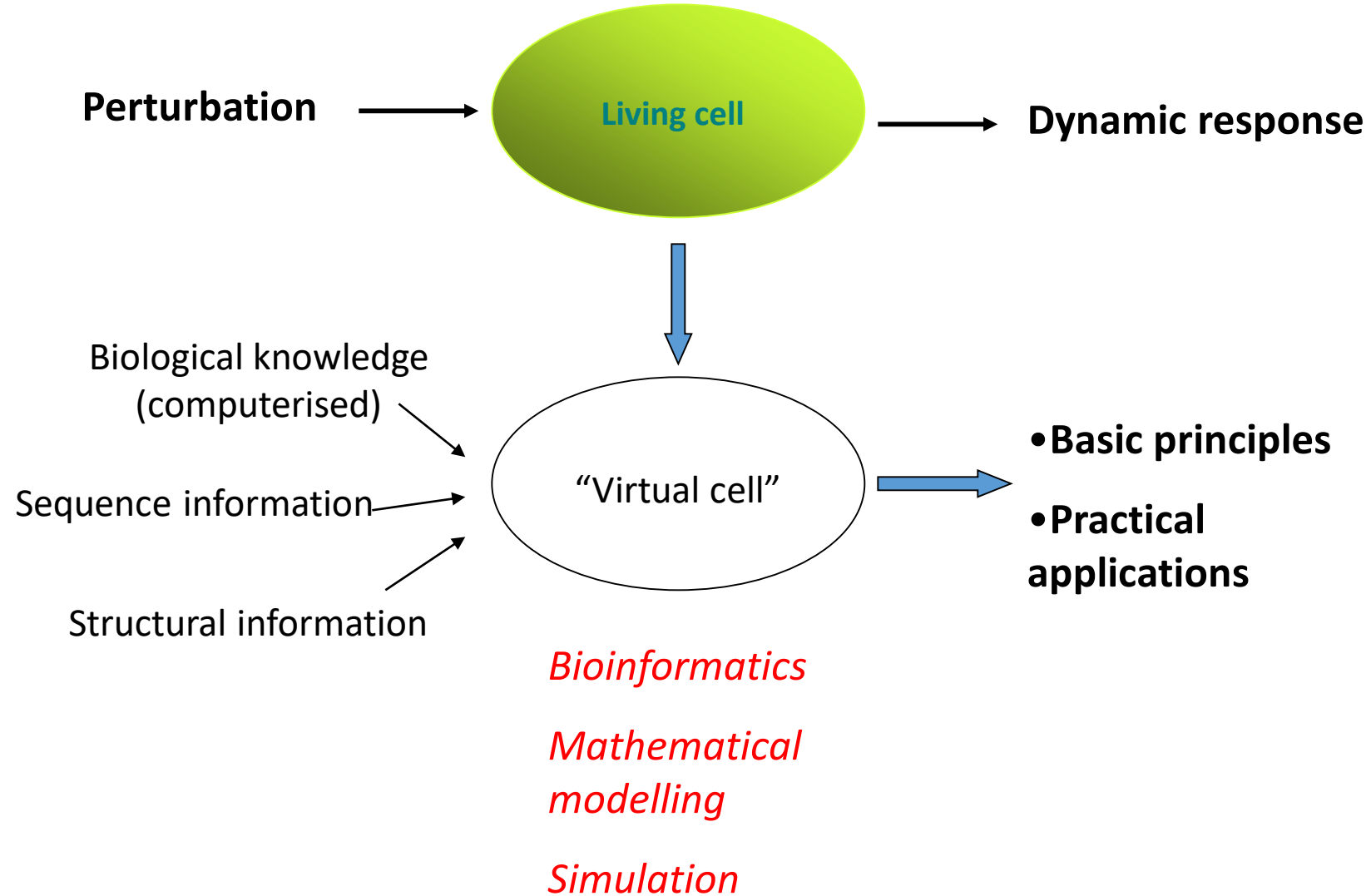
Global approaches: Toward a new Systems Biology



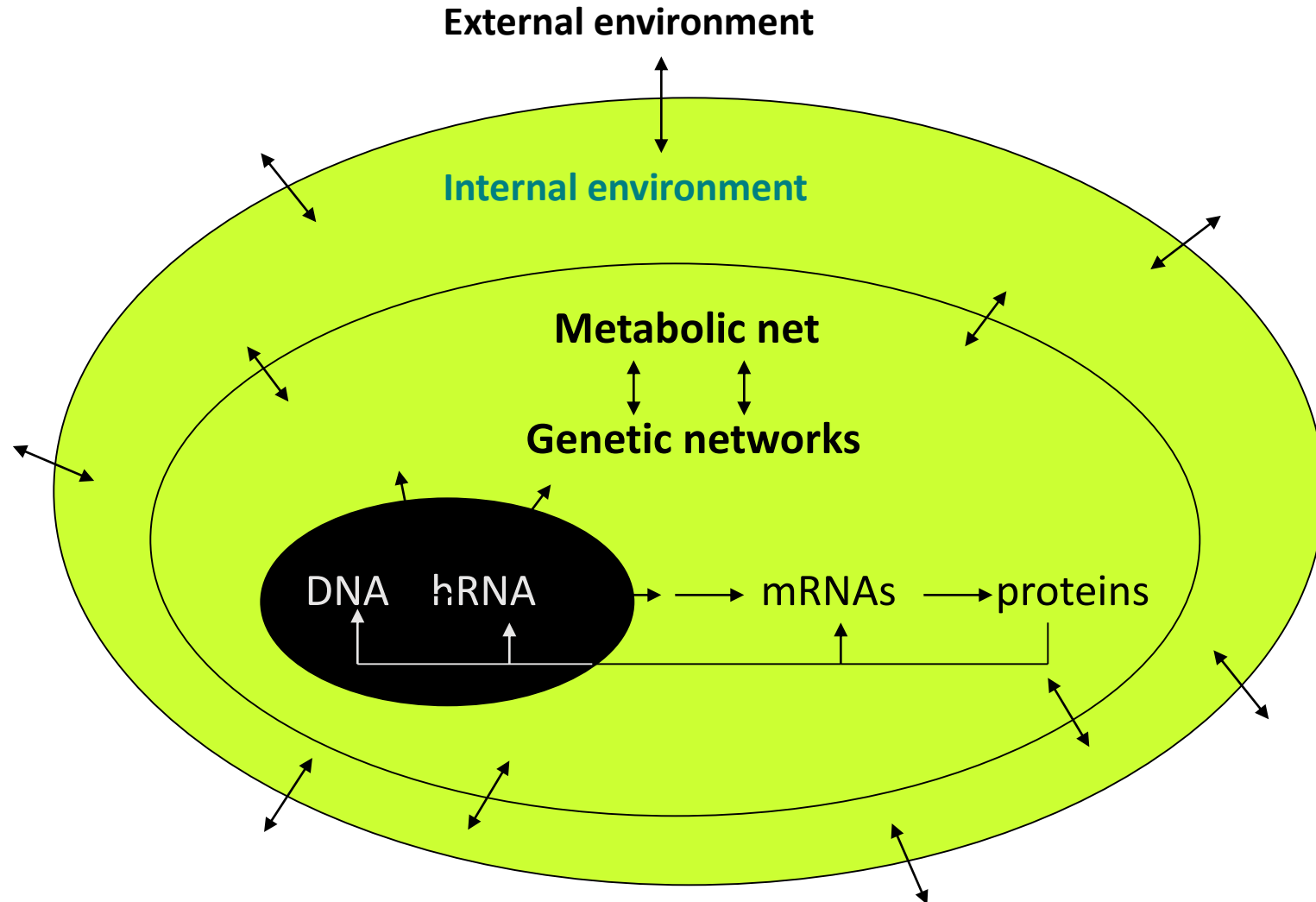
• How does the spatial and temporal organisation of living matter give rise to biological processes?

tissue imaging ↔ EM ↔ X-ray, NMR
cells
molecular complexes

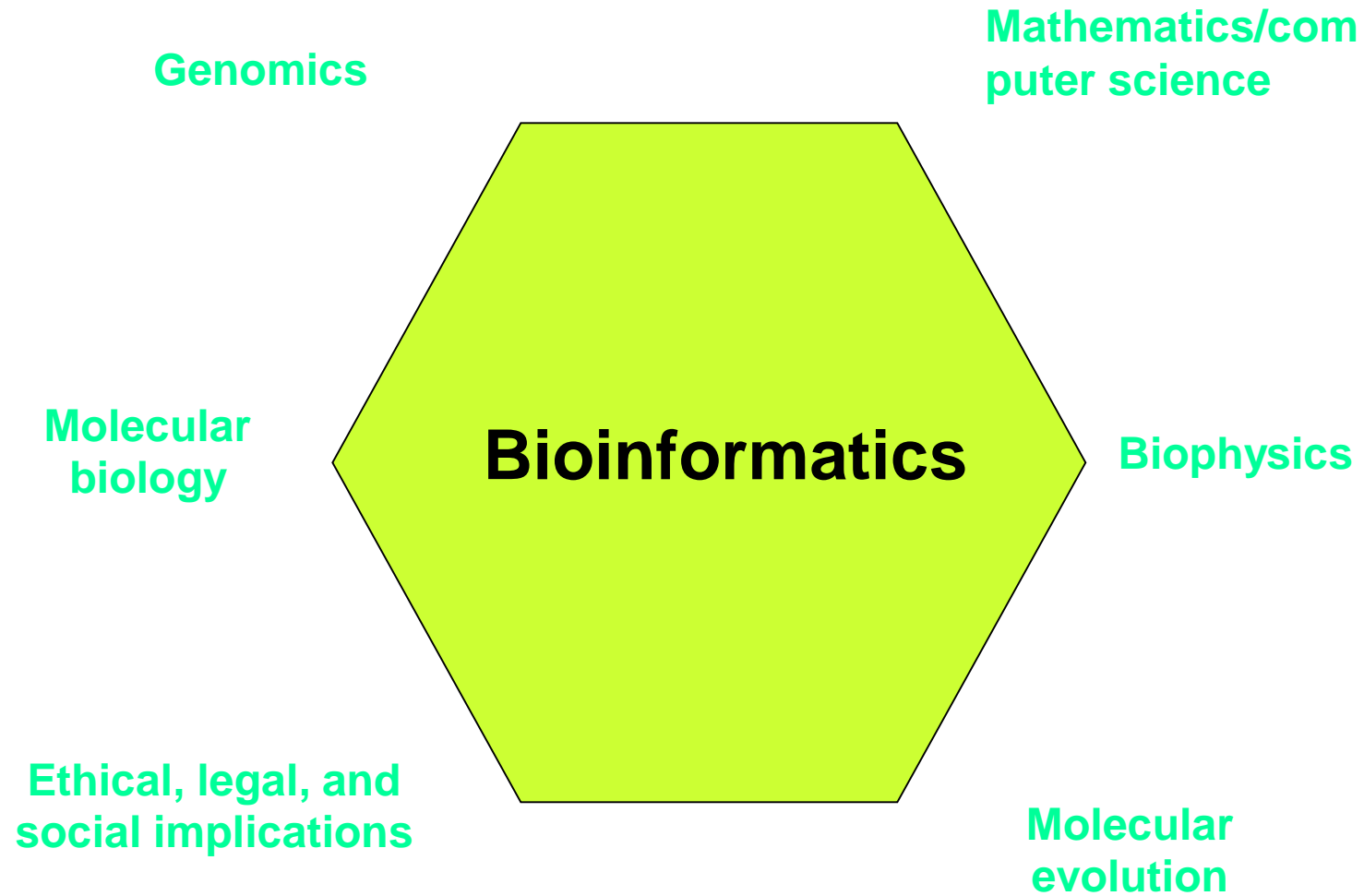
Global approaches: Toward a new Systems Biology



We do not know yet whether the information in the genome is sufficient to reconstruct an entire biological system. Information on building blocks not enough, information on their interactions is essential.



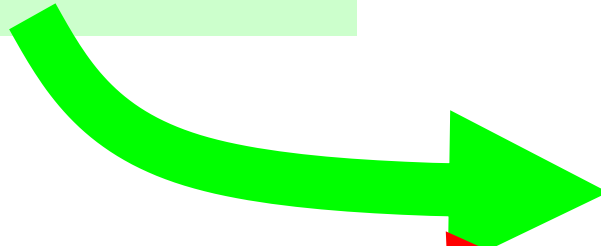
Bioinformatics in context



The field of science in which **biology**, **computer science** and **information technology** merge into a single discipline

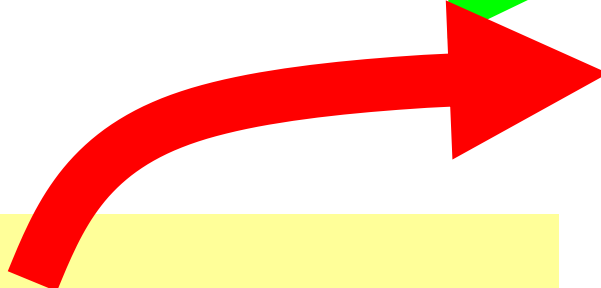
Biologists

collect molecular data:
DNA & Protein sequences,
gene expression, etc.



Bioinformaticians

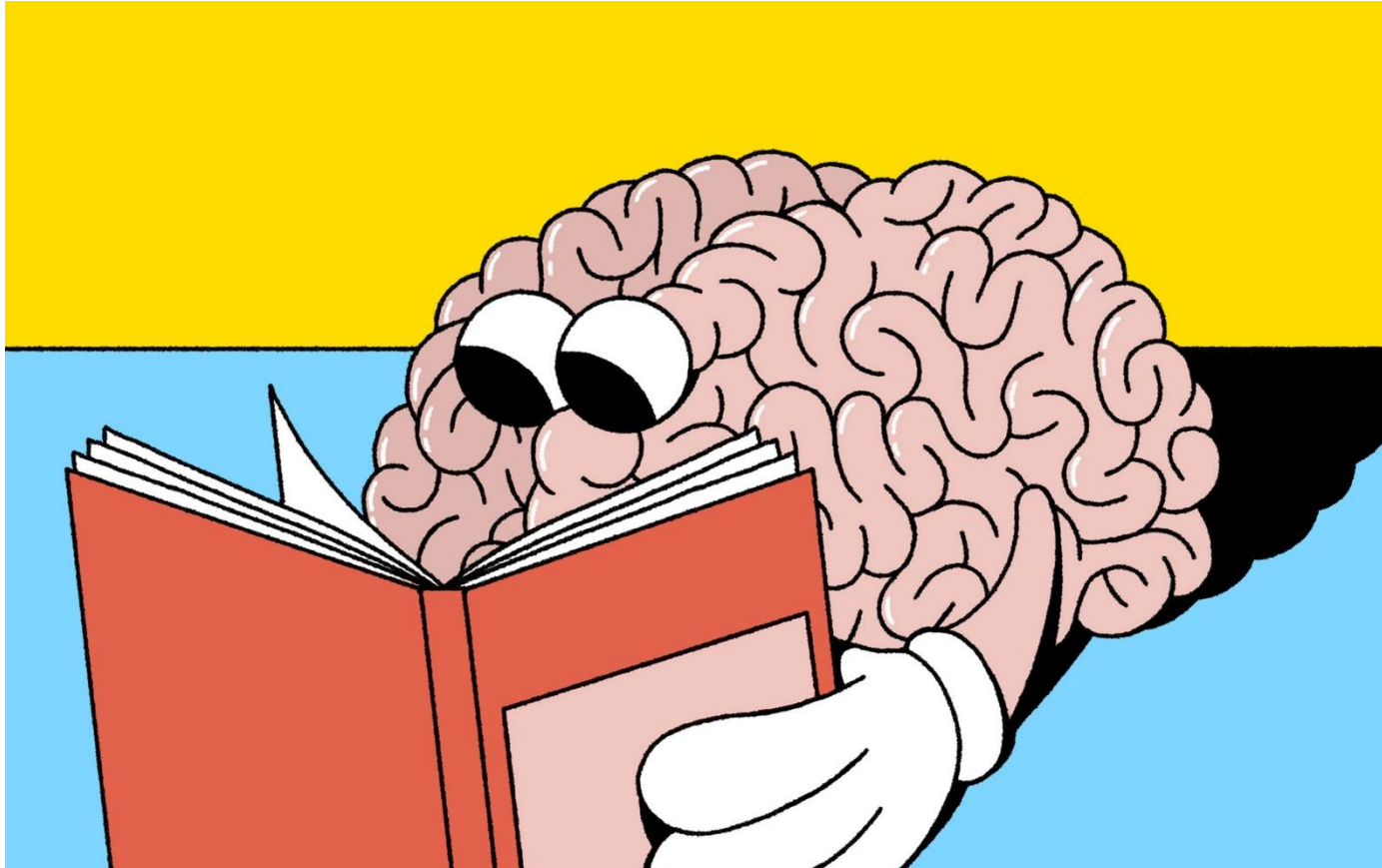
Study biological questions by
analyzing molecular data



Computer scientists

(+Mathematicians, Statisticians, etc.)
Develop tools, softwares, algorithms
to store and analyze the data.

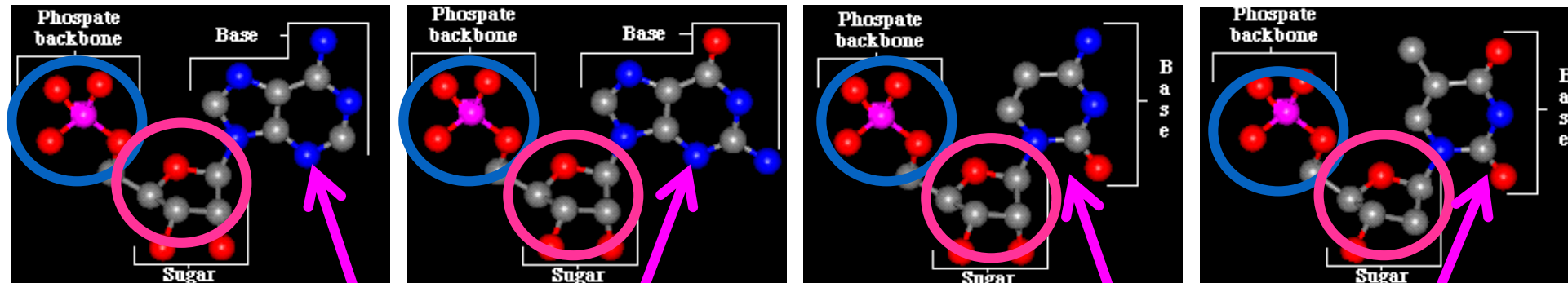
Some biological background....



The hereditary information of all living organisms, with the exception of some viruses, is carried by **deoxyribonucleic acid (DNA)** molecules.

2 purines:

2 pyrimidines:



adenine (A)

guanine (G)

cytosine (C)

thymine (T)

two rings

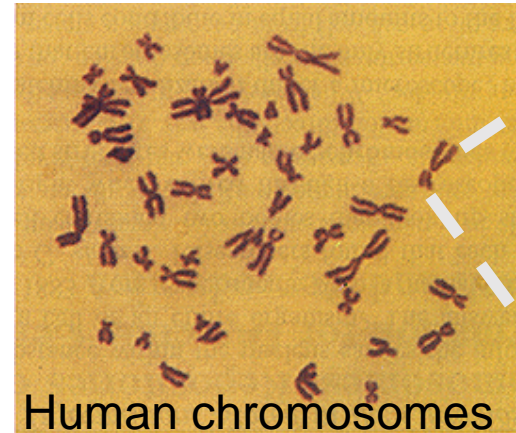
one ring

The entire complement of genetic material carried by an individual is called the **genome**.

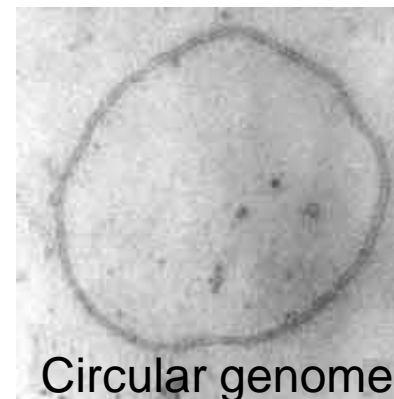
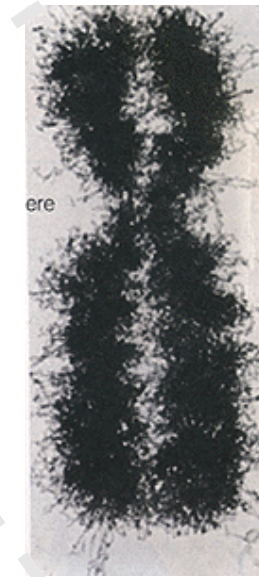
Eukaryotes may have up to 3 subcellular genomes:

1. Nuclear
2. Mitochondrial
3. Plastid

Bacteria have either circular or linear genomes and may also carry plasmids



Human chromosomes

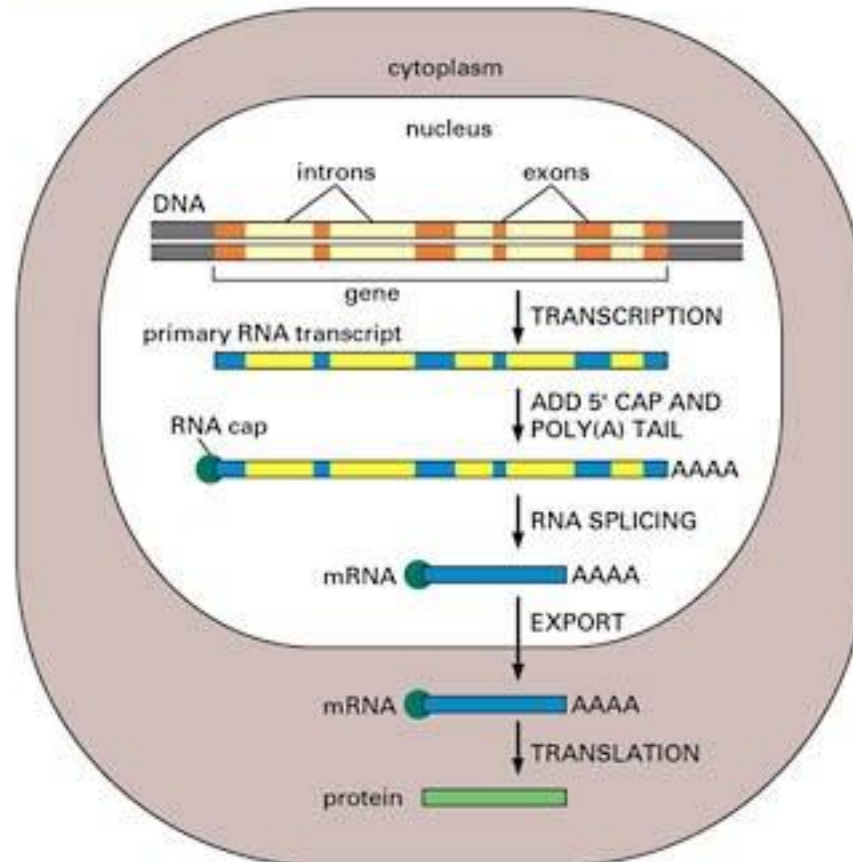


Circular genome

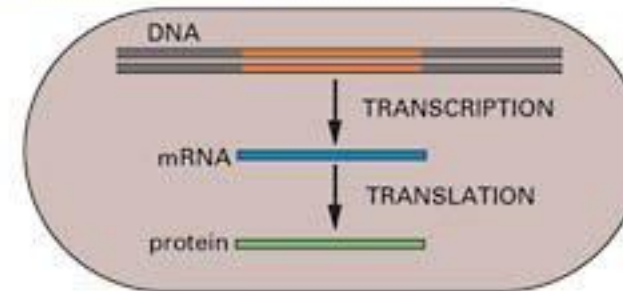
Central dogma: DNA makes RNA makes Protein

Modified dogma: DNA makes DNA and RNA, RNA makes DNA, RNA an Protein

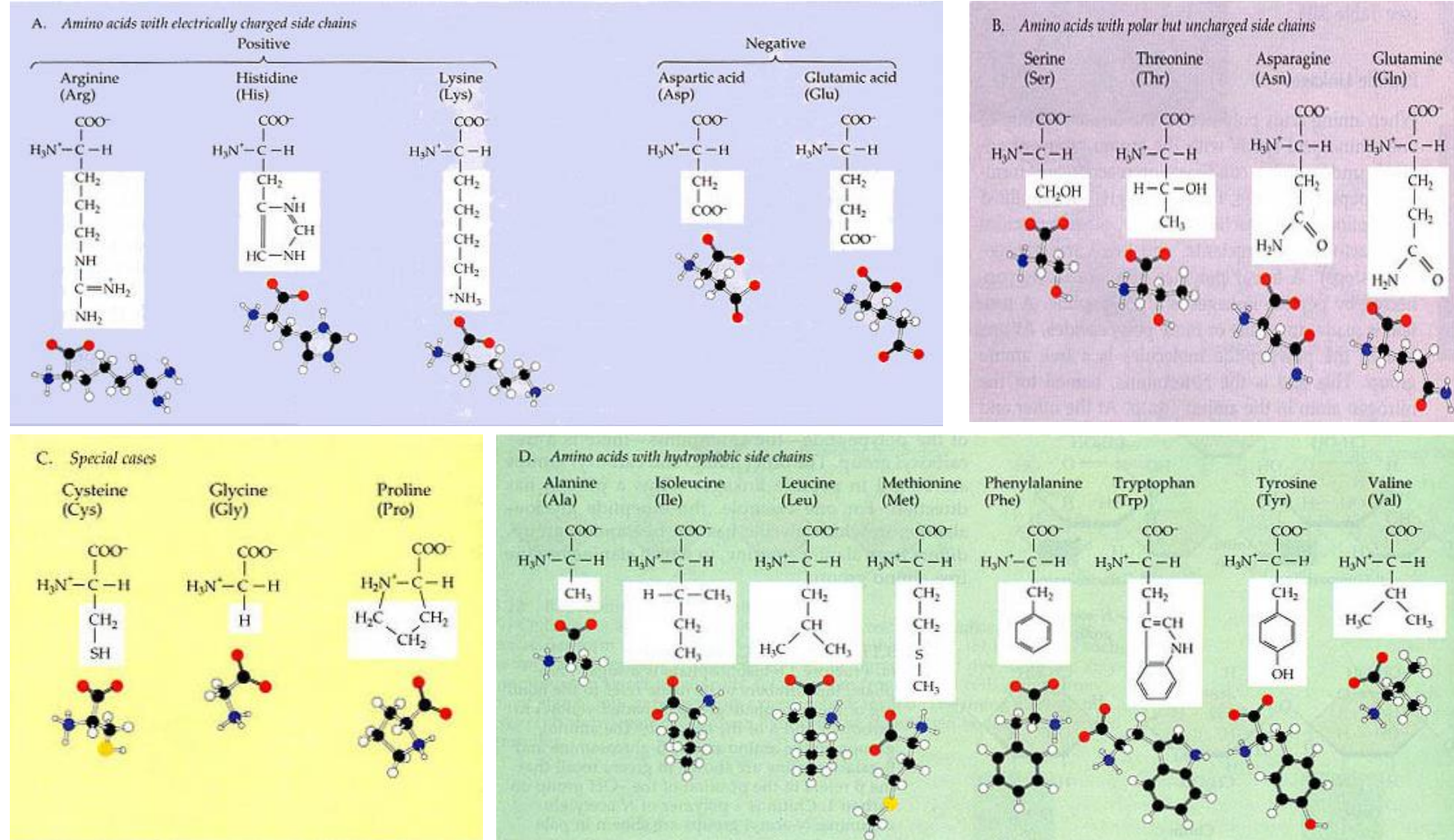
(A) EUCARYOTES



(B) PROCARYOTES



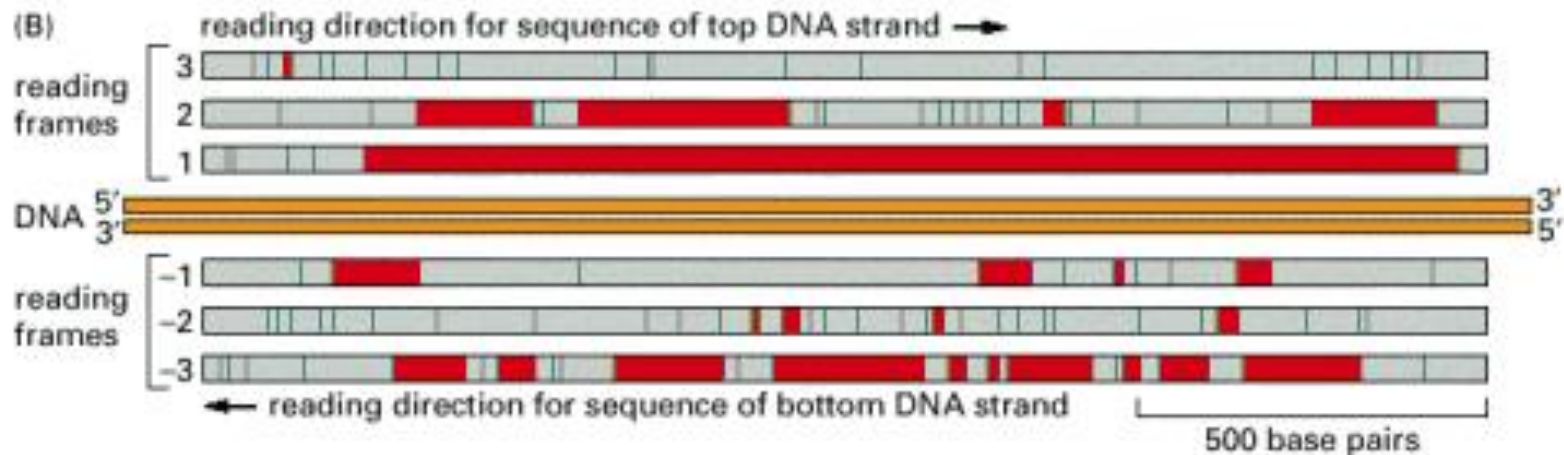
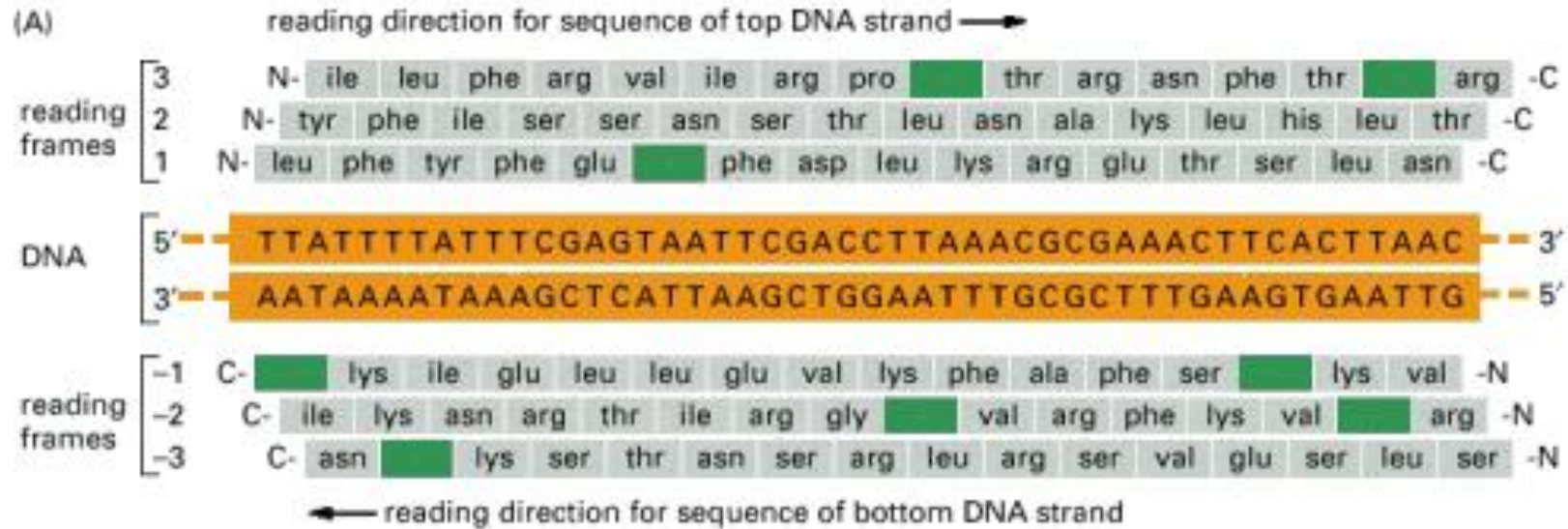
Amino acids - The protein building blocks



		Second letter				
		U	C	A	G	
First letter	U	UUU } Phe UUC } UUA } Leu UUG }	UCU } UCC } Ser UCA } UCG }	UAU } Tyr UAC } UAA Stop UAG Stop	UGU } Cys UGC } UGA Stop UGG Trp	U C A G
	C	CUU } CUC } Leu CUA } CUG }	CCU } CCC } Pro CCA } CCG }	CAU } His CAC } CAA } Gln CAG }	CGU } CGC } Arg CGA } CGG }	U C A G
	A	AUU } AUC } Ile AUA } AUG Met	ACU } ACC } Thr ACA } ACG }	AAU } Asn AAC } AAA } Lys AAG }	AGU } Ser AGC } AGA } Arg AGG }	U C A G
	G	GUU } GUC } Val GUA } GUG }	GCU } GCC } Ala GCA } GCG }	GAU } Asp GAC } GAA } Glu GAG }	GGU } GGC } Gly GGA } GGG }	U C A G

Third letter

Any region of the DNA sequence can, in principle, code for six different amino acid sequences, because any one of three different reading frames can be used to interpret each of the two strands.



AATGCGTATAGGC

DNA

DMPVERILEALAVE

amino acid

**“motifs”: regular
expressions, blocks,
profiles, fingerprints**

**e. g., alpha-helices, beta-
strands**

atomic co-ordinates

domains, folding units

Protein folding

A human Hemoglobin:



How does it all look on a computer monitor?

A cDNA sequence

```
>gi|14456711|ref|NM_000558.3| Homo sapiens hemoglobin, alpha 1 (HBA1), mRNA  
ACTCTTCTGGTCCCCACAGACTCAGAGAGAACCCACCATGGTGCTGTCTCCTGCCGACAAGACCAACGTCAAGGCCG  
CCTGGGGTAAGGTCGGCGCGCACGCTGGCGAGTATGGTGCGGAGGCCCTGGAGAGGATGTTTCCTGTCCTTCCCCACC  
ACCAAGACCTACTTCCCGCACTTCGACCTGAGCCACGGCTCTGCCCAGGTTAAGGGCCACGGCAAGAAGGTGGCCGA  
CGCGCTGACCAACGCCGTGGCGCACGTGGACGACATGCCCAACGCGCTGTCCGCCCTGAGCGACCTGCACGCGCACA  
AGCTTCGGGTGGACCCGGTCAACTTCAAGCTCCTAAGCCACTGCCTGCTGGTGACCCTGGCCGCCACCTCCCCGCC  
GAGTTCACCCCTGCGGTGCACGCCTCCCTGGACAAGTTCCTGGCTTCTGTGAGCACCGTGCTGACCTCCAAATACCG  
TTAAGCTGGAGCCTCGGTGGCCATGCTTCTTGCCCCTTGGGCCTCCCCCCAGCCCCTCCTCCCCTTCCTGCACCCGT  
ACCCCCGTGGTCTTTGAATAAAGTCTGAGTGGGCGGC
```

A cDNA sequence (reading frame)

```
>gi|14456711|ref|NM_000558.3| Homo sapiens hemoglobin, alpha 1 (HBA1), mRNA  
ACTCTTCTGGTCCCCACAGACTCAGAGAGAACCCACCATGGTGCTGTCTCCTGCCGACAAGACCAACGTCAAGGCC  
GCCTGGGGTAAGGTCGGCGCGCACGCTGGCGAGTATGGTGCGGAGGCCCTGGAGAGGATGTTCTGTCCTTCCCCAC  
CACCAAGACCTACTTCCCGCACTTCGACCTGAGCCACGGCTCTGCCCAGGTTAAGGGCCACGGCAAGAAGGTGGCCG  
ACGCGCTGACCAACGCCGTGGCGCACGTGGACGACATGCCCAACGCGCTGTCCGCCCTGAGCGACCTGCACGCGCAC  
AAGCTTCGGGTGGACCCGGTCAACTTCAAGCTCCTAAGCCACTGCCTGCTGGTGACCCTGGCCGCCACCTCCCCGC  
CGAGTTCACCCCTGCGGTGCACGCCTCCCTGGACAAGTTCCTGGCTTCTGTGAGCACCGTGCTGACCTCAAATACC  
GTTAAGCTGGAGCCTCGGTGGCCATGCTTCTTGCCCCTTGGGCCTCCCCCAGCCCCTCCTCCCCTTCTGCACCC  
GTACCCCGTGGTCTTTGAATAAAGTCTGAGTGGGCGGC
```

A protein sequence

```
>gi|4504347|ref|NP_000549.1| alpha 1 globin [Homo sapiens]  
MVLSPADKTNVKAAWGKVGAHAGEYGAELERMFLSFPTTKTYFPHFDLSHGSAQVKGHGKKVADALTNAVAH  
VDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTISKYR
```

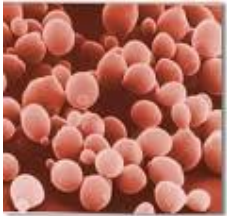
And, a whole genome...

ACTCTTCTGGTCCCCACAGACTCAGAGAGAACCCACCATGGTGCTGTCTCCTGCCGACAAGACCAACGTCAAGGCCGCCTGG
GGTAAGGTCGGCGCGCACGCTGGCGAGTATGGTGCGGAGGCCCTGGAGAGGATGTTCCCTGTCCTTCCCCACCACCAAGACCT
ACTTCCCGCACTTCGACCTGAGCCACGGCTCTGCCCAGGTTAAGGGCCACGGCAAGAAGGTGGCCGACGCGCTGACCAACGC
CGTGGCGCACGTGGACGACATGCCCAACGCGCTGTCCGCCCTGAGCGACCTGCACGCGCACAAAGCTTCGGGTGGACCCGGTC
AACTTCAAGCTCCTAAGCCACTGCCTGCTGGTGACCCTGGCCGCCACCTCCCCGCCGAGTTCACCCCTGCGGTGCACGCCT
CCCTGGACAAGTTCCTGGCTTCTGTGAGCACCGTGCTGACCTCCAAATACCGTTAAGCTGGAGCCTCGGTGGCCATGCTTCT
TGCCCCCTTGGGCCTCCCCCAGCCCCTCCTCCCCTTCTGCACCCGTACCCCGTGGTCTTTGAATAAAGTCTGAGTGGGCG
GCACTCTTCTGGTCCCCACAGACTCAGAGAGAACCCACCATGGTGCTGTCTCCTGCCGACAAGACCAACGTCAAGGCCGCCT
GGGGTAAGGTCGGCGCGCACGCTGGCGAGTATGGTGCGGAGGCCCTGGAGAGGATGTTCCCTGTCCTTCCCCACCACCAAGAC
CTACTTCCCGCACTTCGACCTGAGCCACGGCTCTGCCCAGGTTAAGGGCCACGGCAAGAAGGTGGCCGACGCGCTGACCAAC
GCCGTGGCGCACGTGGACGACATGCCCAACGCGCTGTCCGCCCTGAGCGACCTGCACGCGCACAAAGCTTCGGGTGGACCCGG
TCAACTTCAAGCTCCTAAGCCACTGCCTGCTGGTGACCCTGGCCGCCACCTCCCCGCCGAGTTCACCCCTGCGGTGCACGC
CTCCCTGGACAAGTTCCTGGCTTCTGTGAGCACCGTGCTGACCTCCAAATACCGTTAAGCTGGAGCCTCGGTGGCCATGCTT
CTTGCCCCCTTGGGCCTCCCCCAGCCCCTCCTCCCCTTCTGCACCCGTACCCCGTGGTCTTTGAATAAAGTCTGAGTGGG
CGGCACTCTTCTGGTCCCCACAGACTCAGAGAGAACCCACCATGGTGCTGTCTCCTGCCGACAAGACCAACGTCAAGGCCGC
CTGGGGTAAGGTCGGCGCGCACGCTGGCGAGTATGGTGCGGAGGCCCTGGAGAGGATGTTCCCTGTCCTTCCCCACCACCAAG
ACCTACTTCCCGCACTTCGACCTGAGCCACGGCTCTGCCCAGGTTAAGGGCCACGGCAAGAAGGTGGCCGACGCGCTGACCA
ACGCCGTGGCGCACGTGGACGACATGCCCAACGCGCTGTCCGCCCTGAGCGACCTGCACGCGCACAAAGCTTCGGGTGGACCC
GGTCAACTTCAAGCTCCTAAGCCACTGCCTGCTGGTGACCCTGGCCGCCACCTCCCCGCCGAGTTCACCCCTGCGGTGCAC
GCCTCCCTGGACAAGTTCCTGGCTTCTGTGAGCACCGTGCTGACCTCCAAATACCGTTAAGCTGGAGCCTCGGTGGCCATGC
TTCTTGCCCCCTTGGGCCTCCCCCAGCCCCTCCTCCCCTTCTGCACCCGTACCCCGTGGTCTTTGAATAAAGTCTGAGTG
GGCGGCGCCGTGGCGCACGTGGACGACATGCCCAACGCGCTGTCCGCCCTGAGCGACCTGCACGCGCACAAAGCTTCGGGTGG
ACCCGGTCAACTTCAAGCTCCTAAGCCACTGCCTGCTGGTGACCCTGGCCGCCACCTCCCCGCCGAGTTCACCCCTGCGGT
GCACGCCTCCCTGGACAAGTTCCTGGCTTCTGTGAGCACCGTGCTGACCTCCAAATACCGTTAAGCTGGAGCCTCGGTGGCC
ATGCTTCTTGCCCCCTTGGGCCTCCCCCAGCCCCTCCTCCCCTTCTGCACCCGTACCCCGTGGTCTTTGAATAAAGTCTG
AGTGGGCGGCACTCTTCTGGTCCCCACAGACTCAGAGAGAACCCACCATGGTGCTGTCTCCTGCCGACAAGACCAACGTCAA
GGCCGCCTGGGGTAAGGTCGGCGCGCACGCTGGCGAGTATGGTGCGGAGGCCCTGGAGAGGATGTTCCCTGTCCTTCCCCACC
ACCAAGACCTACTTCCCGCACTTCGACCTGAGCCACGGCTCTGCCCAGGTTAAGGGCCACGGCAAGAAGGTGGCCG...

How big are whole genomes?



E. coli 4.6×10^6 nucleotides
– Approx. 4,000 genes



Yeast 15×10^6 nucleotides
– Approx. 6,000 genes

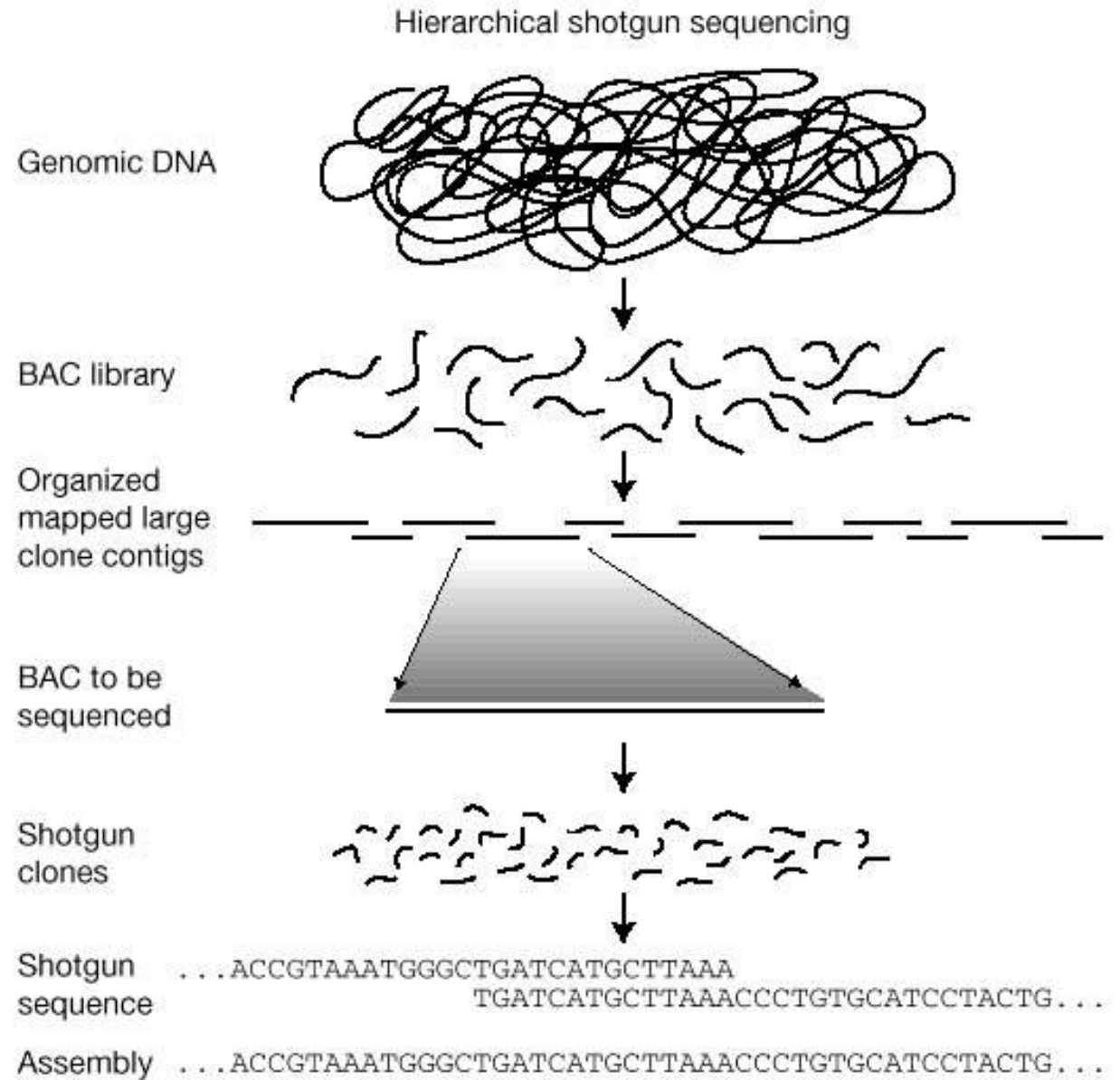


Human 3×10^9 nucleotides
– Approx. 30,000 genes

Smallest human chromosome 50×10^6 nucleotides

What do we actually do with bioinformatics?

Sequence assembly



(next generation sequencing)

Genome annotation

UCSC Genome Browser on Human Mar. 2006 Assembly (hg18)

move <<<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x

position/search chr7:150,287,621-156,821,424 jump clear size 6,533,804 bp. configure

chr7 (q36.1-q36.3) 31.1 33 q34 35

Scale chr7: 151000000 152000000 153000000 154000000 155000000 156000000

RefSeq Genes
Vertebrate Multiz Alignment & Conservation (44 Species)
Placental Mammal Basewise Conservation by PhyloP

Multiz Alignments of 44 Vertebrates

Rhesus
Tarsier
Mouse
Dog
Elephant
Opossum
Platypus
Chicken
Lizard
X_tropicalis
Stickleback

SNPs (130) Simple Nucleotide Polymorphisms (dbSNP build 130)

SNPs (129) Simple Nucleotide Polymorphisms (dbSNP build 129)

Affy SNP 6.0
Affy SNP 6.0 SV
Illumina 1M-Duo

SNP Genotyping Arrays

Repeating Elements by RepeatMasker

SINE
LINE
LTR
DNA
Simple
Low Complexity
Satellite
RNA
Other
Unknown

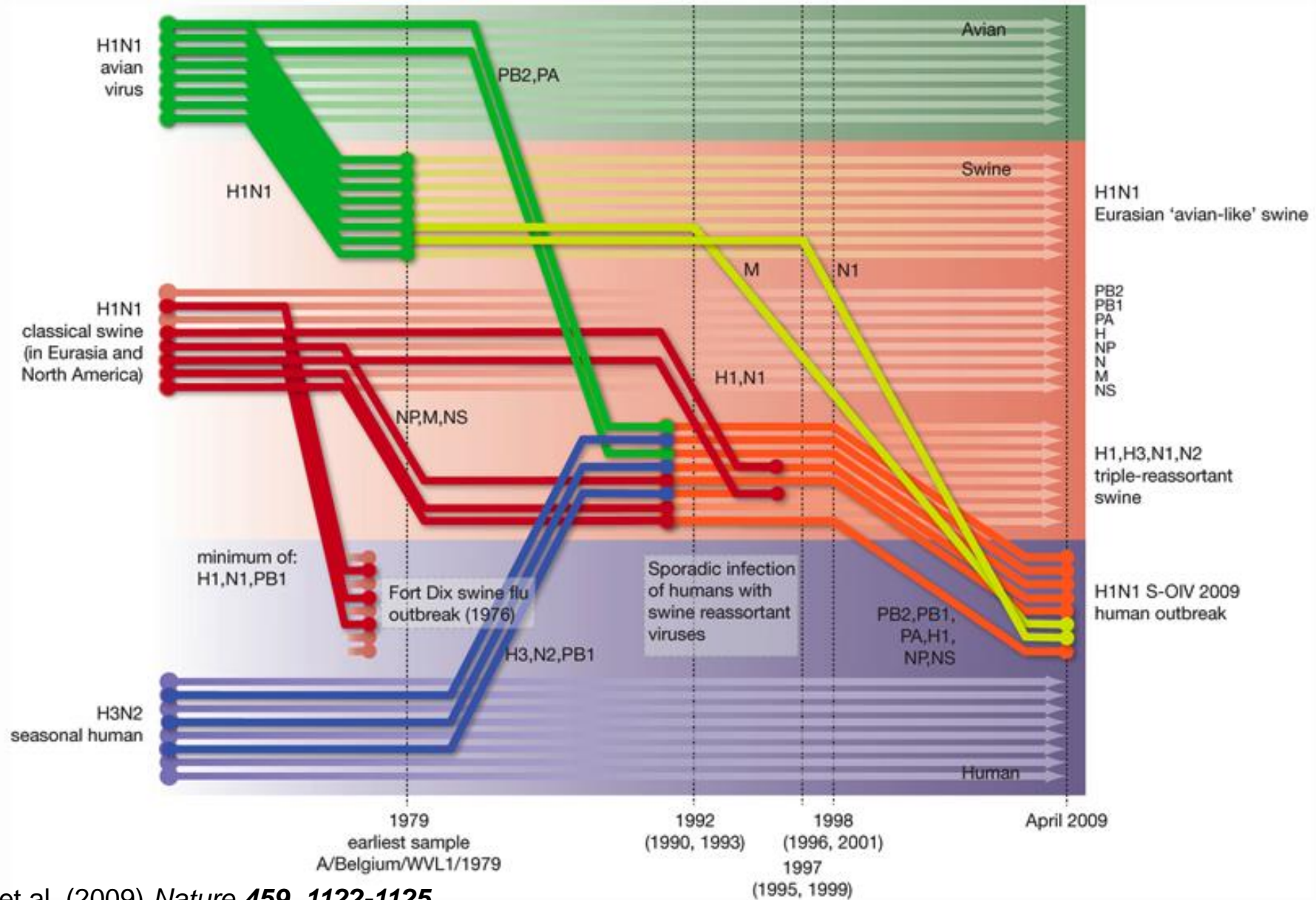
move start Click on a feature for details. Click on base position to zoom in around cursor. Click move end

< 2.0 > gray/blue bars on left for track options and descriptions.

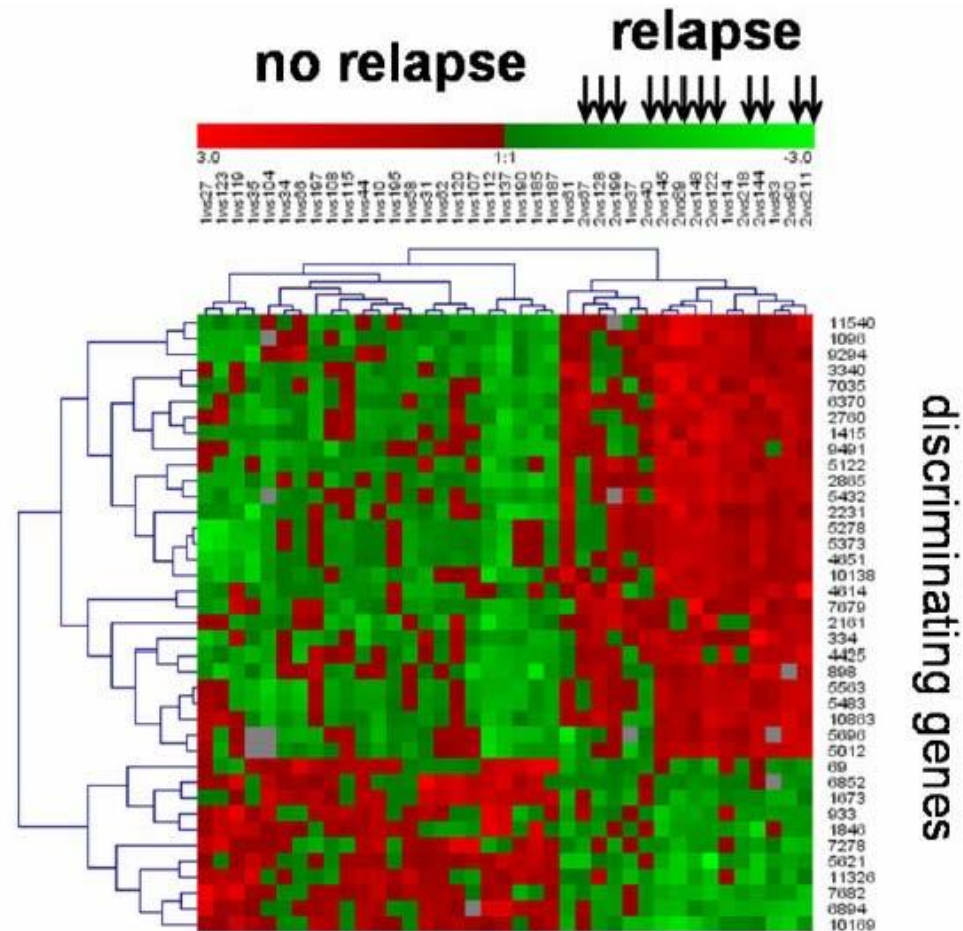
default tracks hide all add custom tracks configure reverse refresh

Molecular evolution

Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic

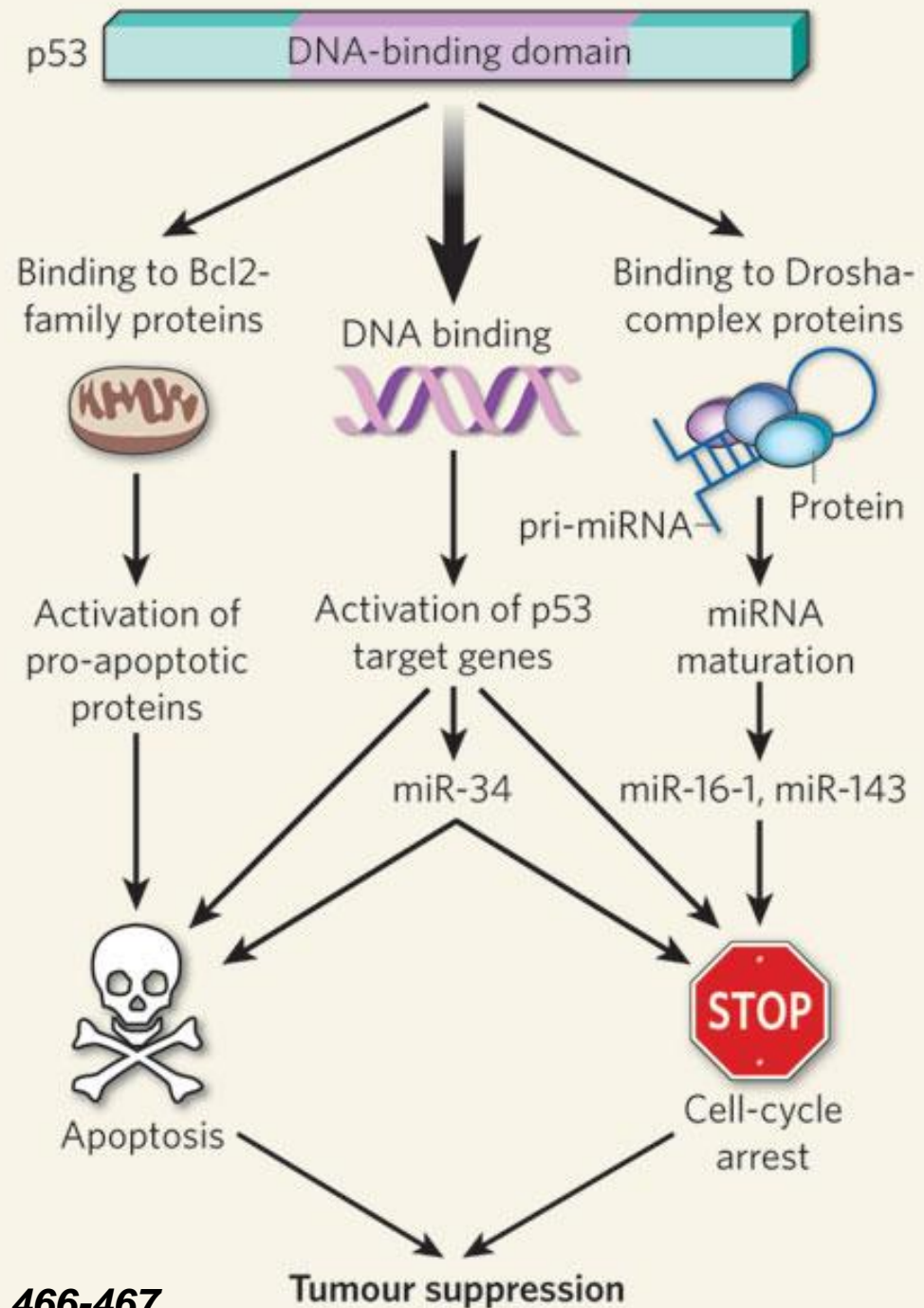


Analysis of gene expression

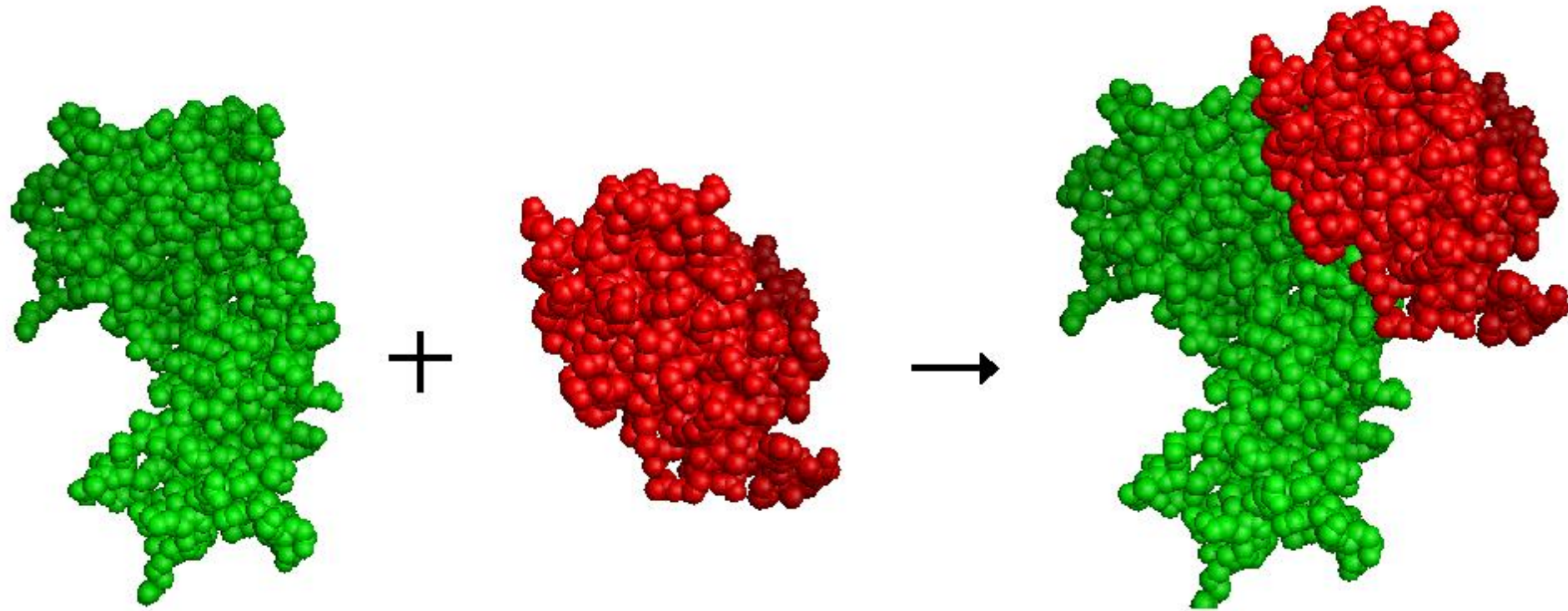


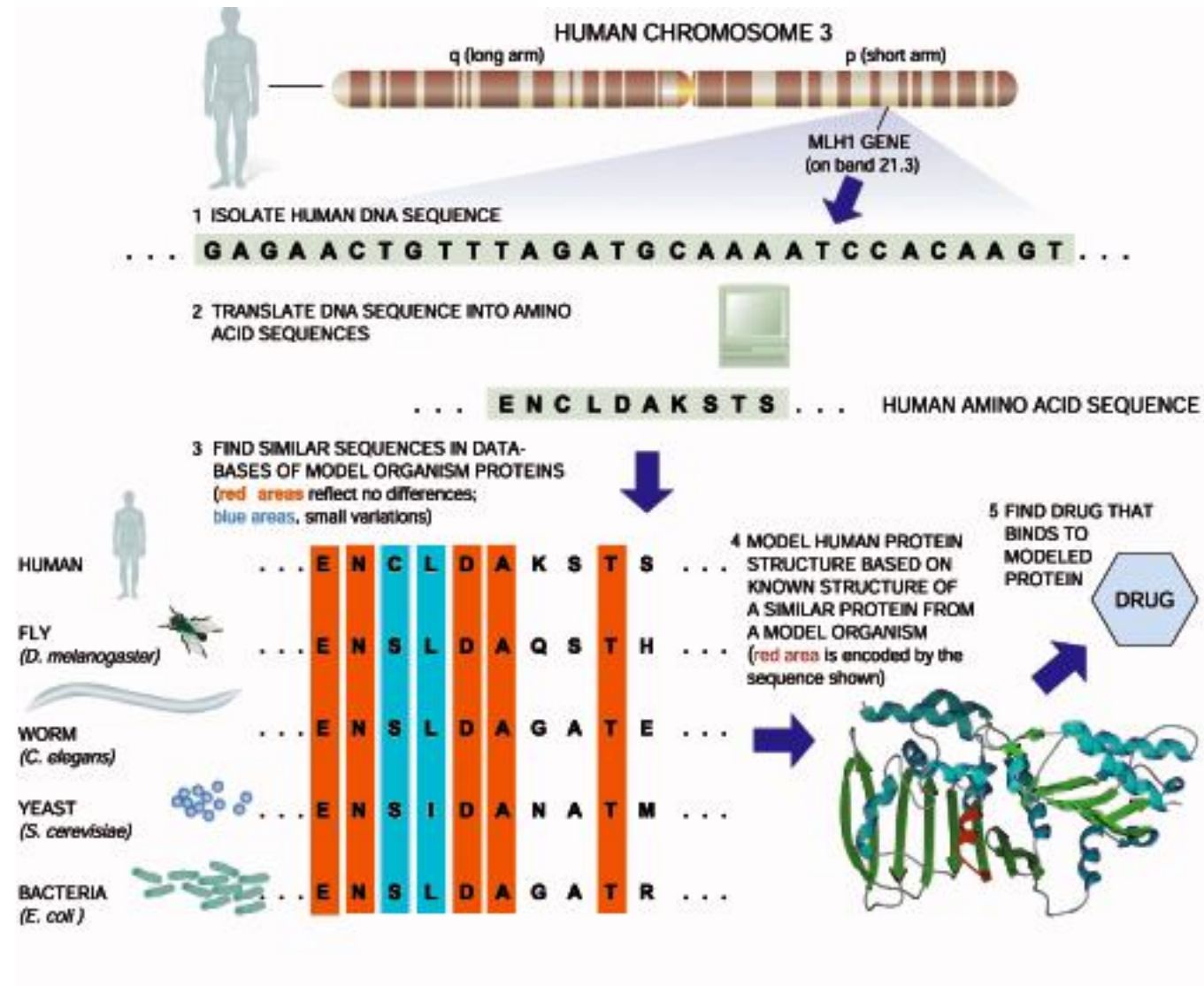
Gene expression profile of relapsing versus non-relapsing Wilms tumors. A set of 39 genes discriminates between the two classes of tumors.

Analysis of regulation

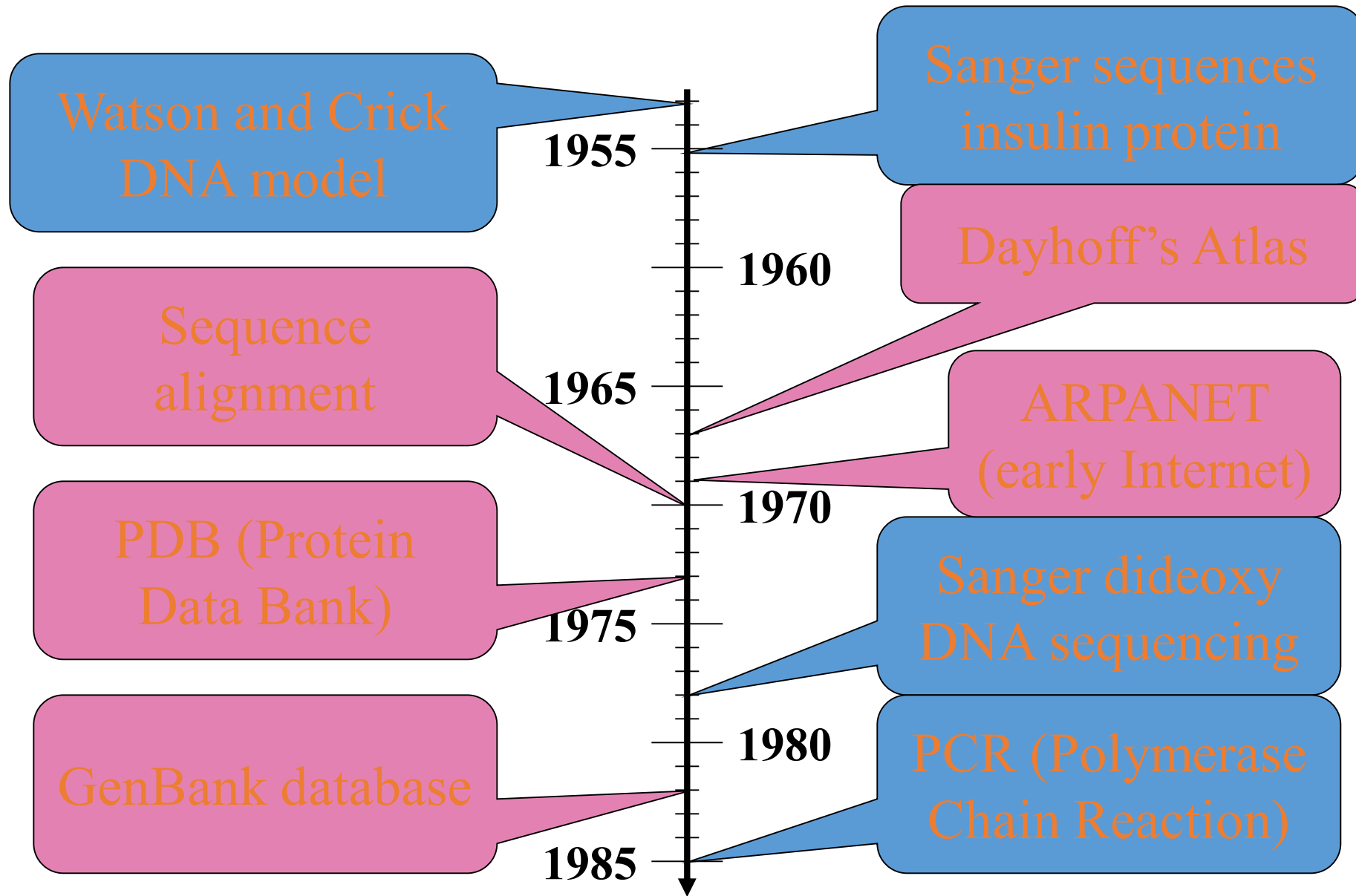


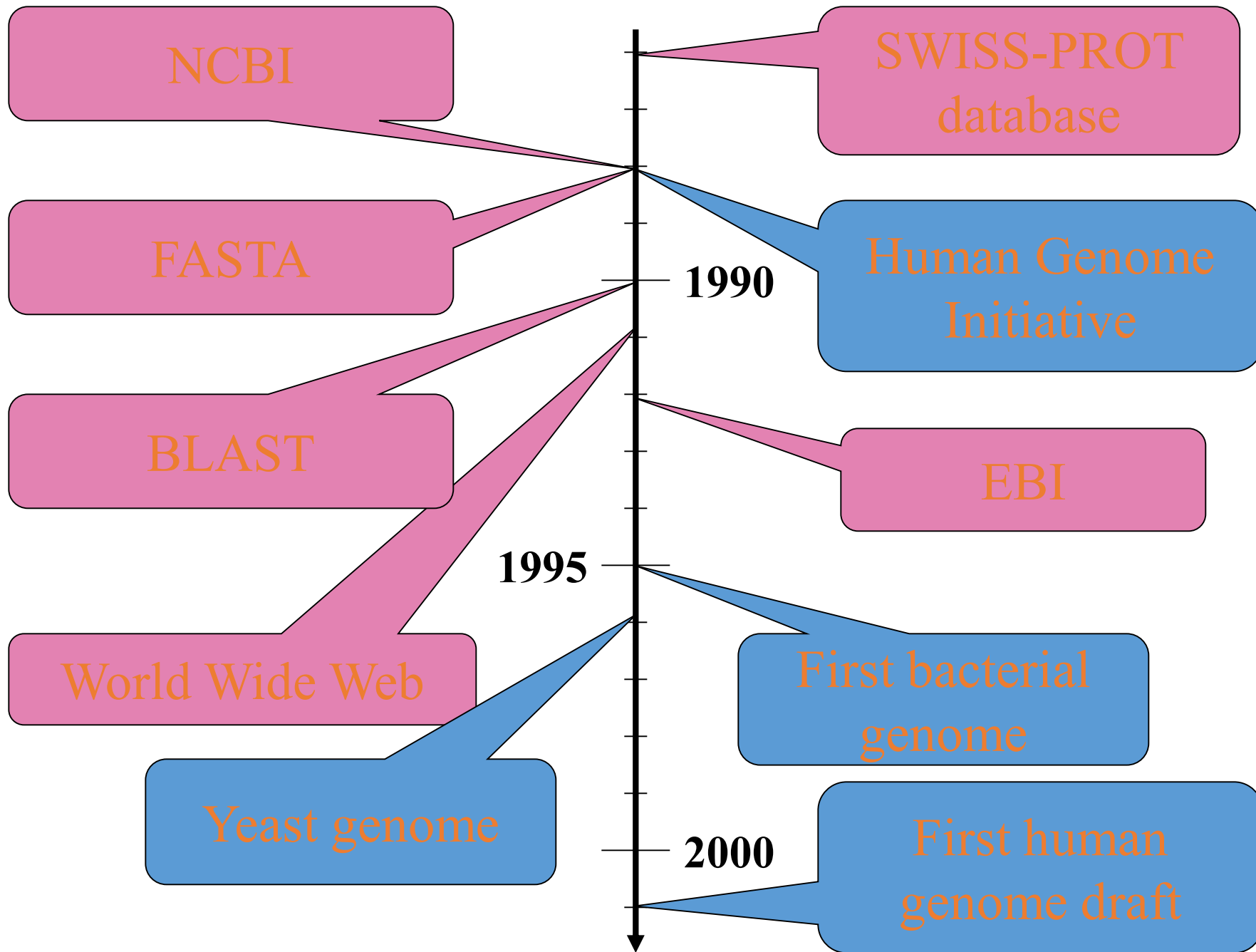
Protein structure prediction
Protein docking





From DNA to Genome





Origin of bioinformatics and biological databases:

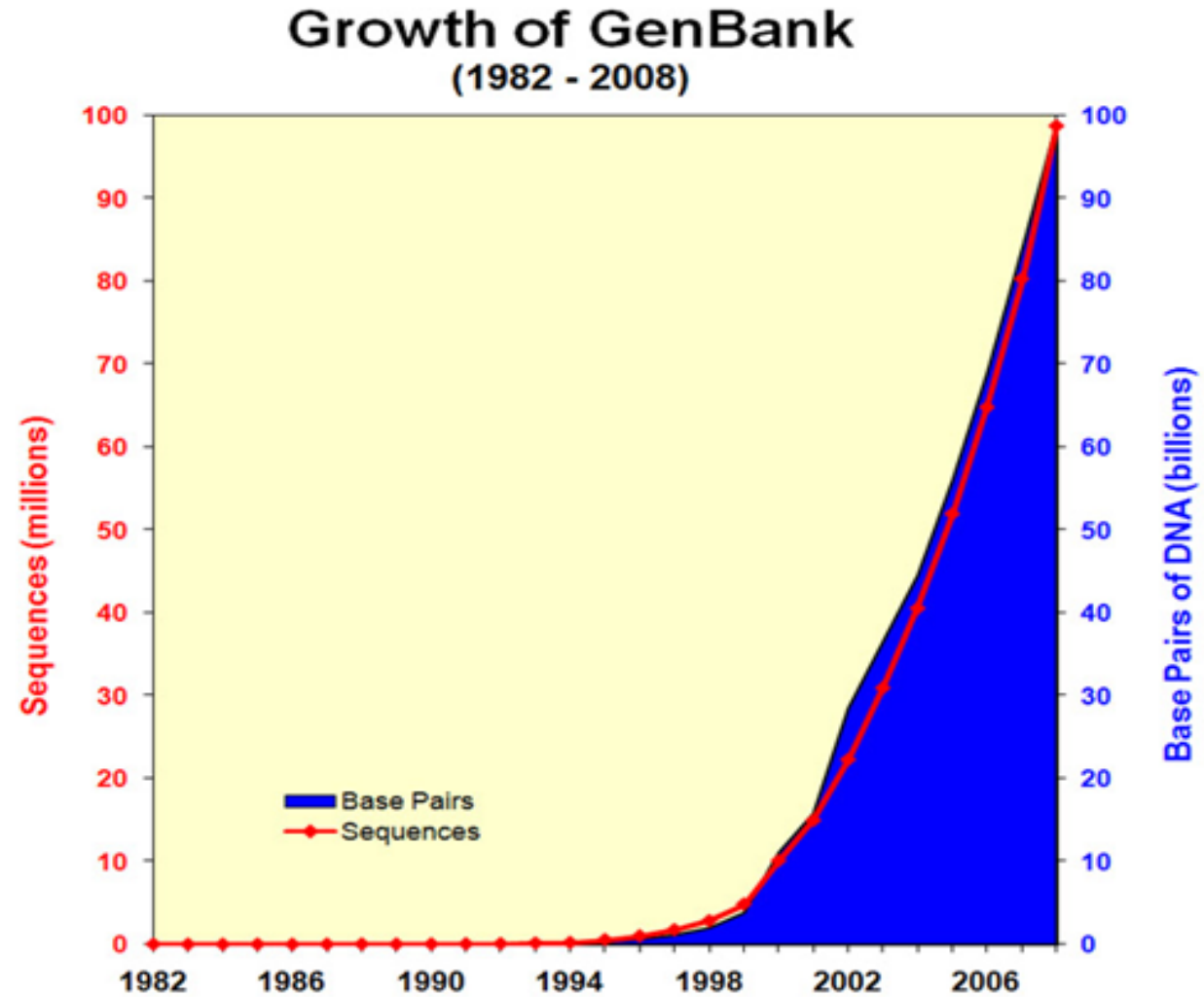
The first protein sequence reported was that of bovine insulin in **1956**, consisting of 51 residues.

Nearly a decade later, the first nucleic acid sequence was reported, that of yeast tRNA^{alanine} with 77 bases.

In 1965, **Dayhoff** gathered all the available sequence data to create the first bioinformatic database (*Atlas of Protein Sequence and Structure*).

The Protein DataBank followed in 1972 with a collection of ten X-ray crystallographic protein structures. The SWISSPROT protein sequence database began in 1987.

Nucleotides

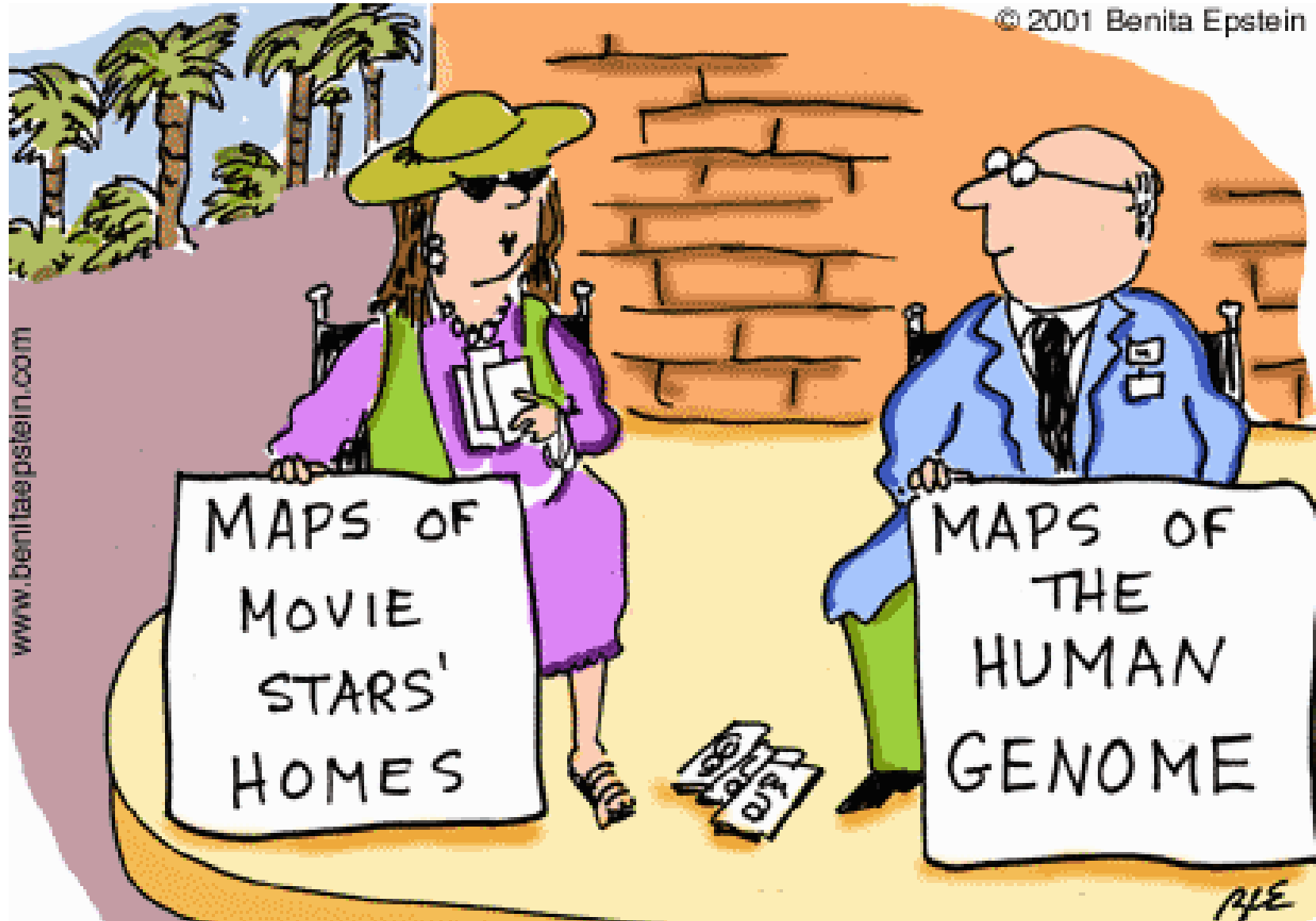


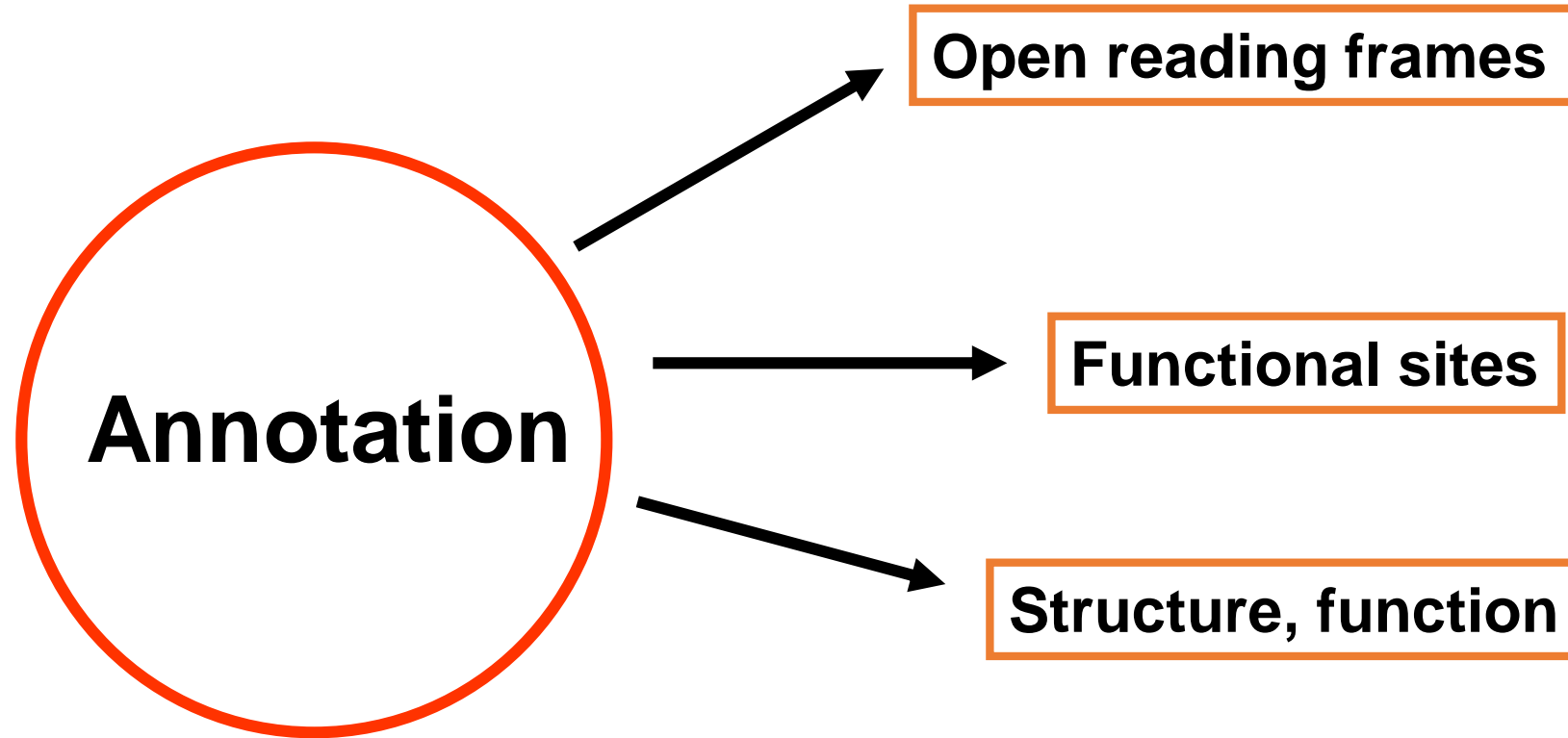
Complete Genomes

as of August 2011:

Eukaryotes	37
Prokaryotes	1708
Total	1745

What can we do with sequences and other type of molecular information?





**CCTGACAAATTCGACGTGCGGCATTGCATGCAGACGTGCATG
CGTGCAAATAATCAATGTGGACTTTTTCTGCGATTATGGAAGAA
CTTTGTTACGCGTTTTTTGTCATGGCTTTGGTCCCGCTTTGTTC
AGAATGCTTTTAATAAGCGGGGTTACCGGTTTGGTTAGCGAGA
AGAGCCAGTAAAAGACGCAGTGACGGAGATGTCTGATG CAA
TAT GGA CAA TTG GTT TCT TCT CTG AAT**
..... TGAAAACGTA

promoter

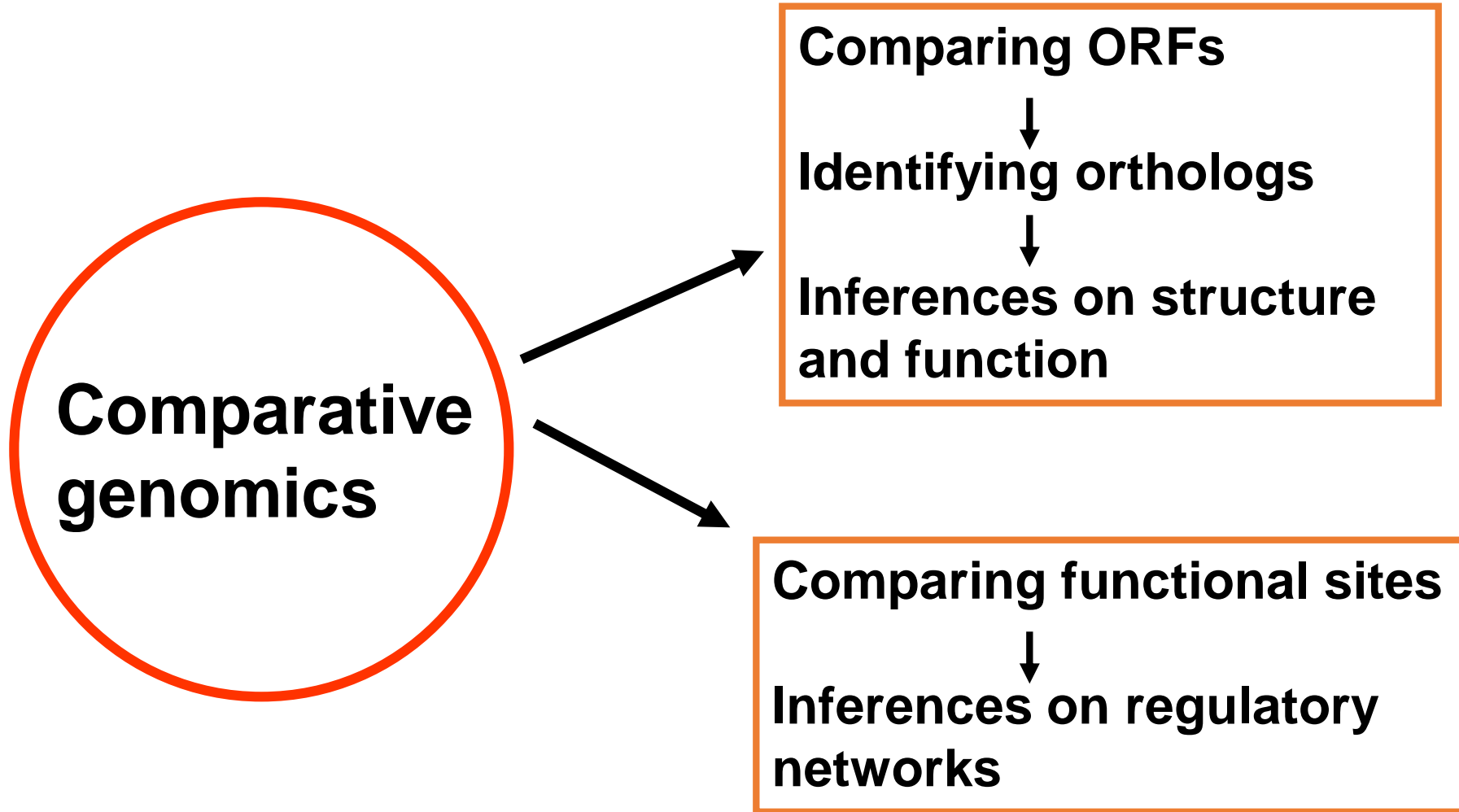
TF binding site

CCTGACAAATTTCGACGTGCCGCATTGCATGCAGACGTGCATG
 CGTGCAAATAATCAATGTGGACTTTTCTGCGATTATGGAAGA(A)
 CTTTGTTACGCGTTTTTGTTCATGGCTTTGGTCCCGCTTTGTTC
 AGAATGCTTTTAATAAGCGGGGTTACCGGTTTGGTTAGCGAGA
 AGAGCCAGTAAAAGACGCAGTGACGGAGATGTCTGATG CAA
TAT GGA CAA TTG GTT TCT TCT CTG AAT
 TGAAAAACGTA

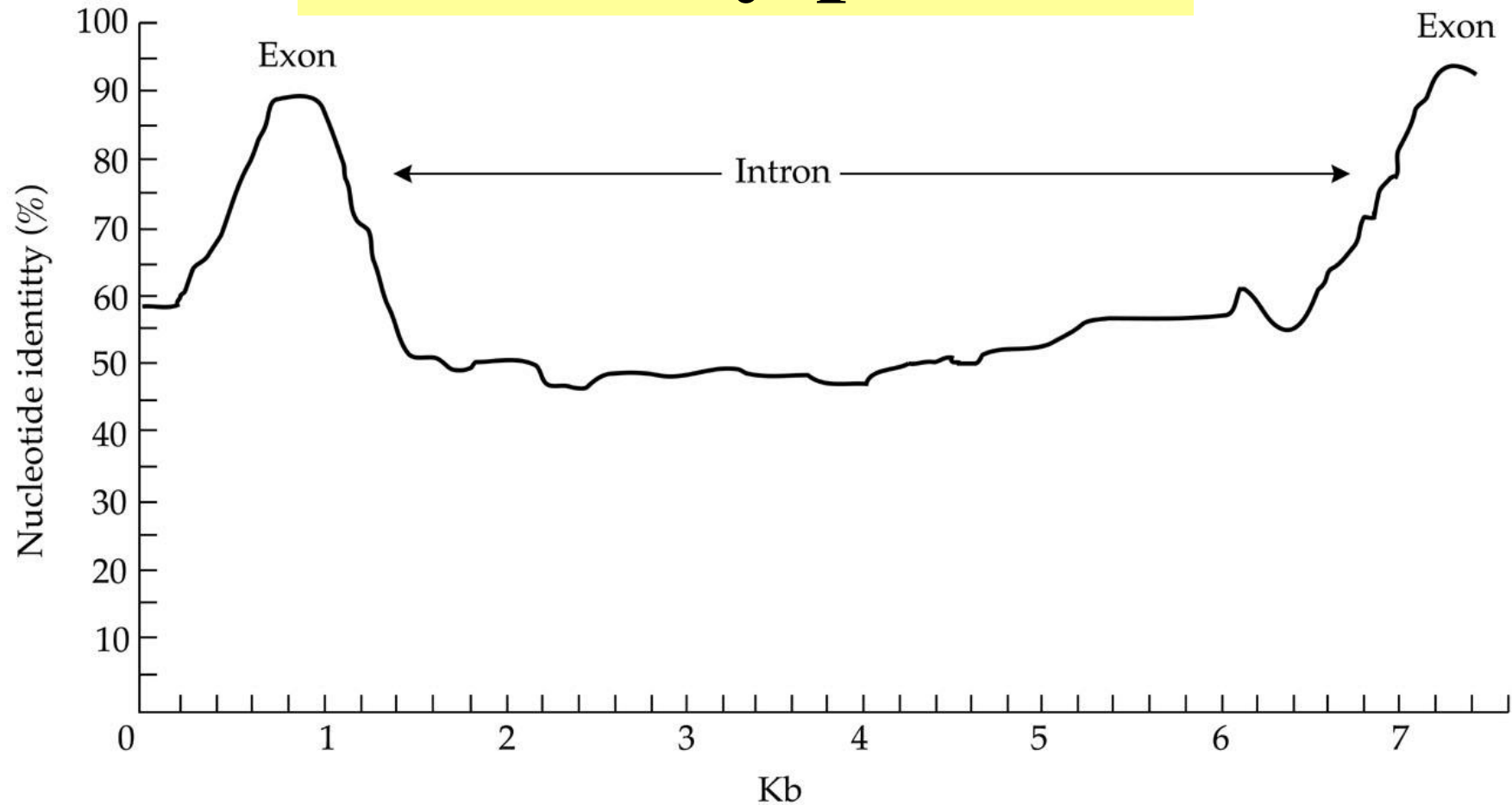
Transcription Start Site

Ribosome binding Site

ORF = Open Reading Frame
 CDS = Coding Sequence



Similarity profiles

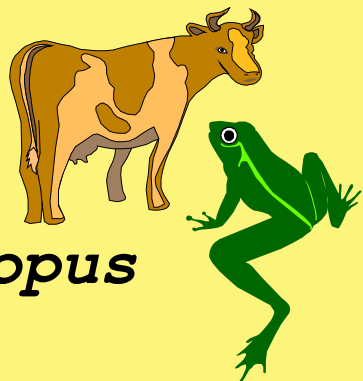


Researchers can learned a great deal about the structure and function of human genes by examining their counterparts in model organisms.

Alignment preproinsulin

Xenopus

Bos



MALWMQCLP-LVLVLLFSTPNTTEALANQHL

MALWTRLRPLLALLALWPPPPARAFVNQHL

**** : * * . * : * : . . * : . * : ****

Xenopus

Bos

CGSHLVEALYLVCGRGFFYYPKIKRDIEQ

CGSHLVEALYLVCGERGFFYTPKARREVEG

***** : ***** ** : * : : *

Xenopus

Bos

AQVNGPQDNELDG-MQFQPQEYQMKRGIV

PQVG---ALELAGGPGAGGLEGPPQKRGIV

. ** . ** * * *****

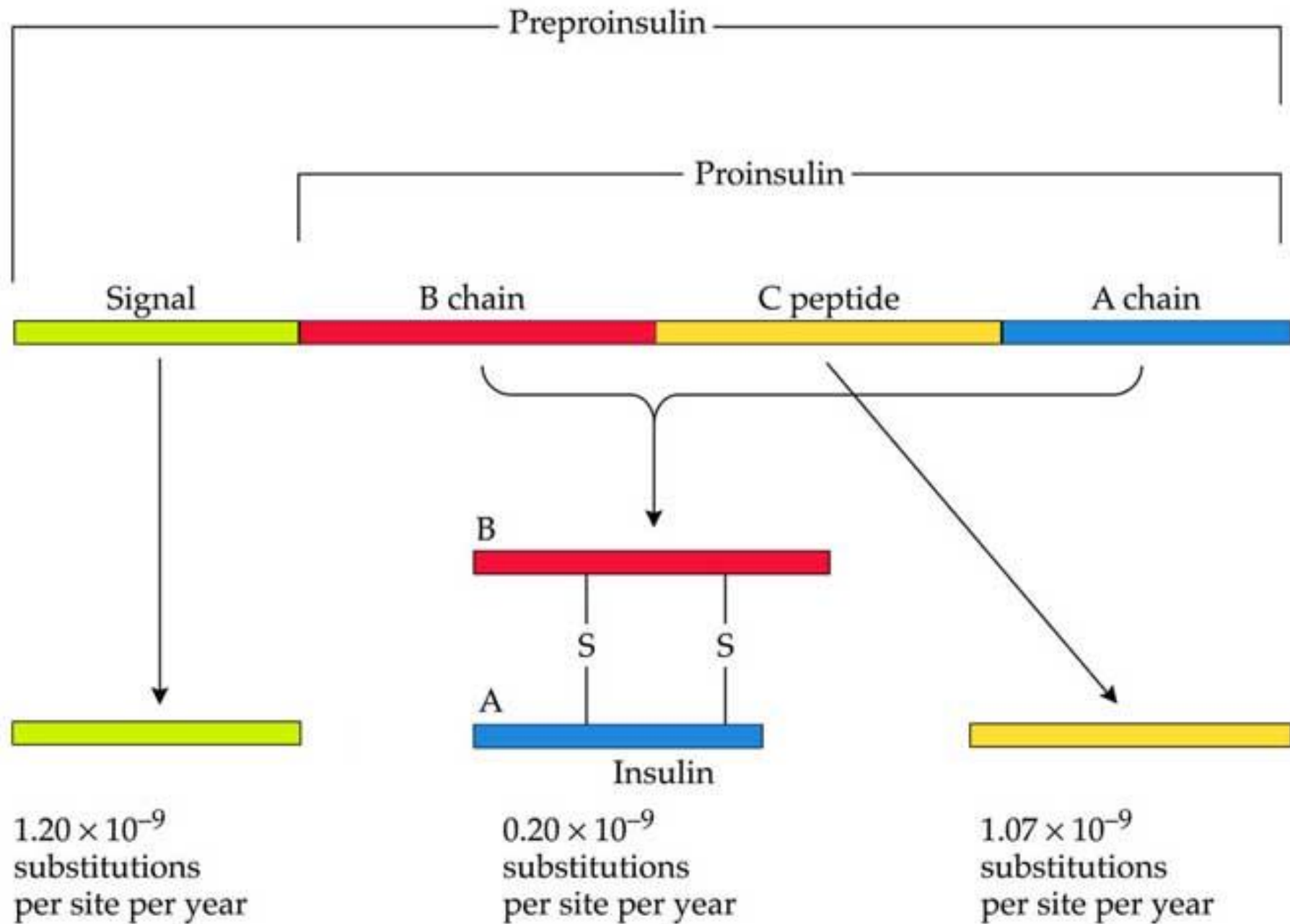
Xenopus

Bos

EQCCHSTCSLFQLENYCN

EQCCASVCSLYQLENYCN

**** * . **** : *****



Ultraconserved Elements in the Human Genome

Gill Bejerano, Michael Pheasant, Igor Makunin, Stuart Stephen, W. James Kent, John S. Mattick, & David Haussler (*Science* 2004. 304:1321-1325)

There are 481 segments longer than 200 base pairs (bp) that are absolutely conserved (100% identity with no insertions or deletions) between orthologous regions of the human, rat, and mouse genomes. Nearly all of these segments are also conserved in the chicken and dog genomes, with an average of 95 and 99% identity, respectively. Many are also significantly conserved in fish. These ultraconserved elements of the human genome are most often located either overlapping exons in genes involved in RNA processing or in introns or nearby genes involved in the regulation of transcription and development.

There are 156 intergenic, untranscribed, ultraconserved segments

**Junk:
Supporting evidence**

Megabase deletions of gene deserts result in viable mice

Marcelo A. Nóbrega*, Yiwen Zhu*, Ingrid Plajzer-Frick, Veena Afzal & Edward M. Rubin

DOE Joint Genome Institute Walnut Creek, California 94598, USA, and Genomics Division Lawrence Berkeley National Laboratory Berkeley, California 94720, USA

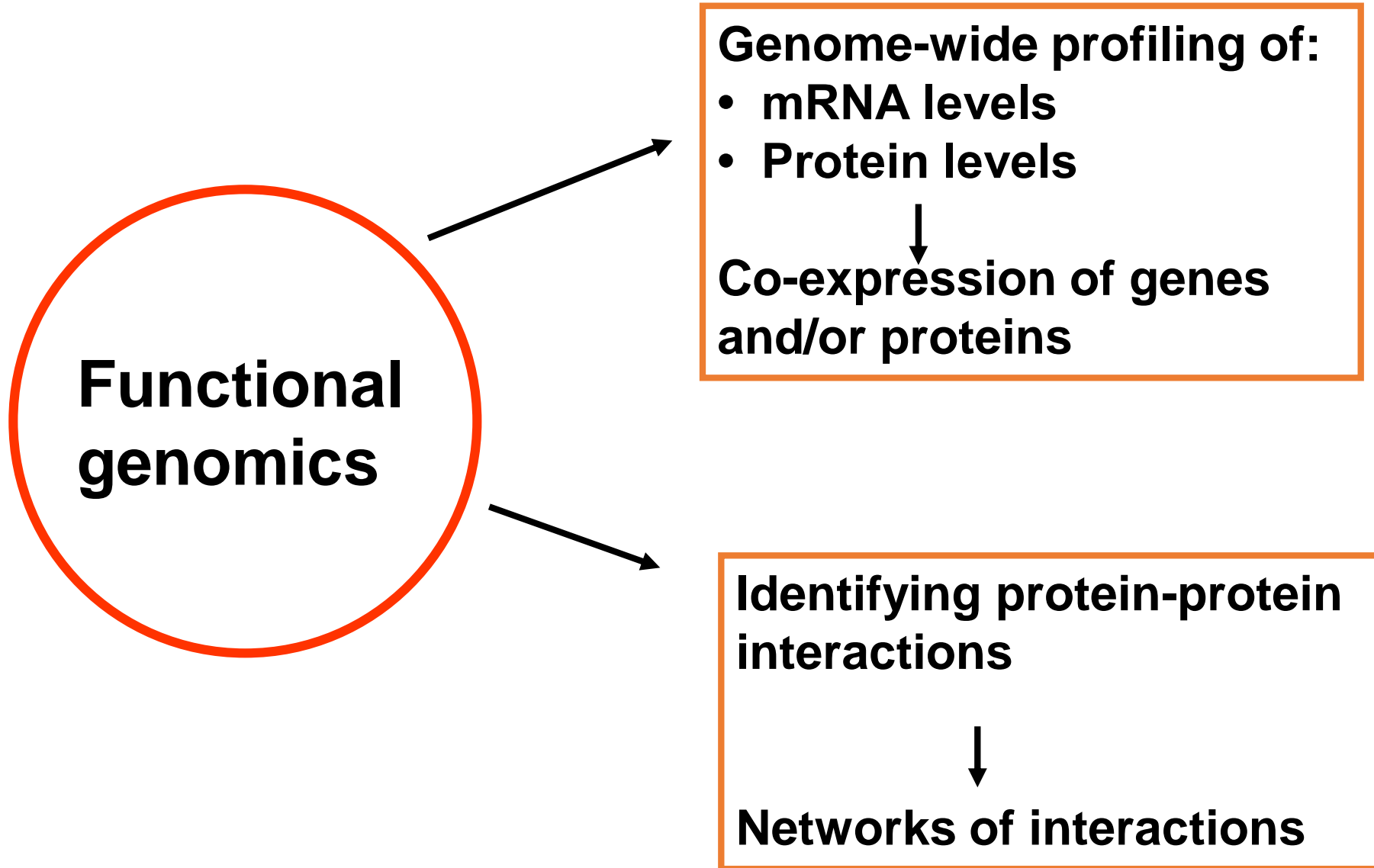
* These authors contributed equally to this work

The functional importance of the roughly 98% of mammalian genomes not corresponding to protein coding sequences remains largely undetermined¹. Here we show that some large-scale deletions of the non-coding DNA referred to as gene deserts²⁻⁴ can be well tolerated by an organism. We deleted two large non-coding intervals, 1,511 kilobases and 845 kilobases in length, from the mouse genome. Viable mice homozygous for the deletions were generated and were indistinguishable from wild-type littermates with regard to morphology, reproductive fitness, growth, longevity and a variety of physiological parameters related to homeostasis. Further detailed analysis of the deleted regions revealed that multiple genes bracketing the deletions showed significant expression differences in the homozygous mice. Together, the two deleted segments harbour 1,243 non-coding sequences conserved between humans and rodents (more than 100 base pairs, 70% identity). Some of the deleted sequences might encode for functions unidentified in our screen; nonetheless, our results support the existence of potentially functional non-coding sequences in the genomes of mammals.

Junk is real!

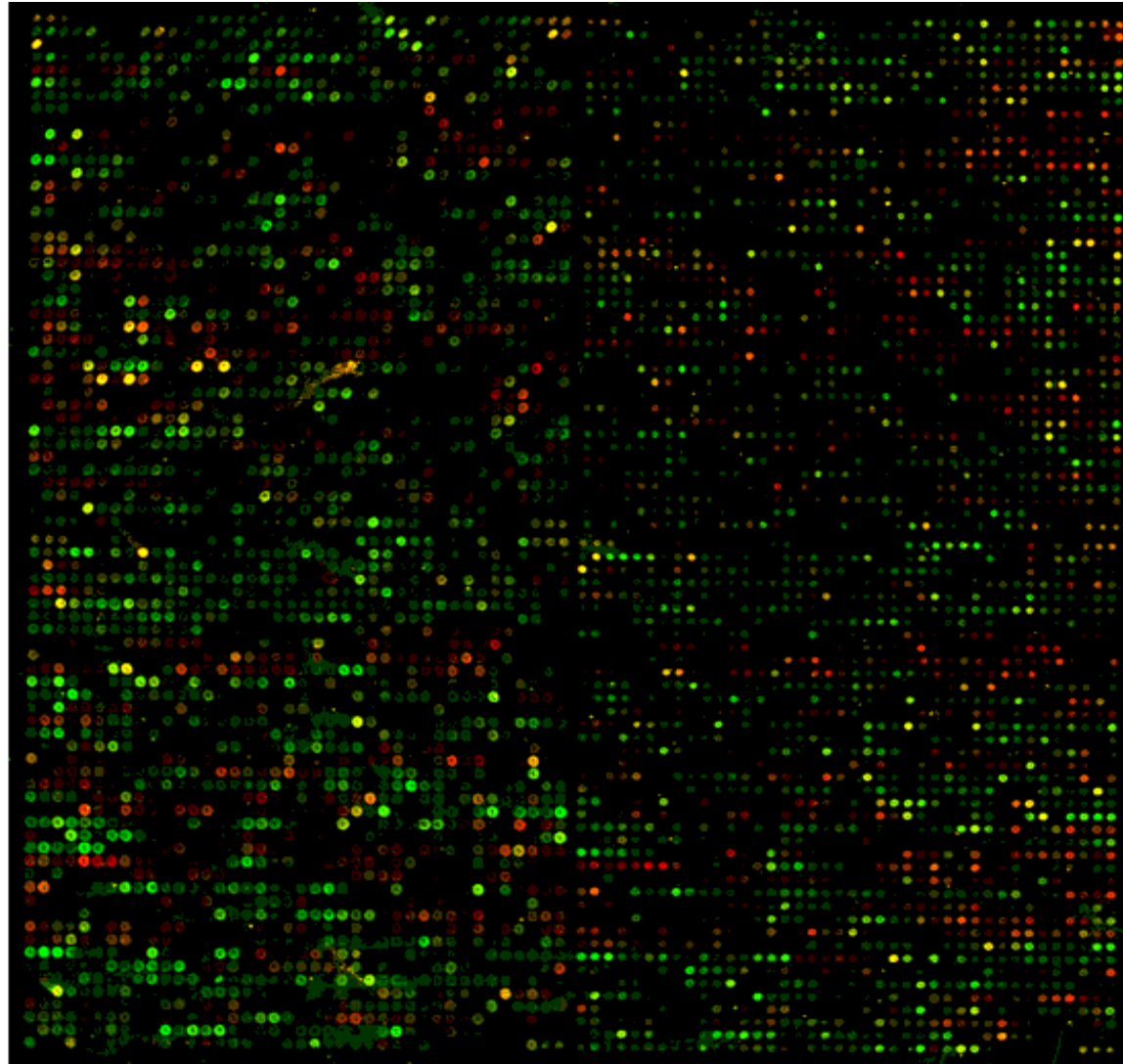
Nature 431, 988-993 (21 October 2004)

The genome of an organism is frequently referred to as its 'book



Understanding the function of genes and other parts of the genome

The
complete
S. cerevisiae
genome
on a
microarray
chip
hybridised
to RNA
from cultures in
anaerobic and
aerobic
stationary phase





**Structural
genomics**



**Assign structure to all
proteins encoded in
a genome**



Yes, if you train quickly, you can become a **SWISS-PROT** annotator before the human proteome is done, but first eat your dinner!

Biological databases

Database or databank?

Initially

- Databank (in UK)
- Database (in the USA)

Solution

- The abbreviation *db*

What is a Database?

A **structured collection** of data held in computer storage; *esp.* one that incorporates software to make it accessible in a variety of ways; *transf.*, any **large collection** of information.

database management: the organization and manipulation of data in a database.

database management system (DBMS): a software package that provides all the functions required for database management.

database system: a database together with a database management system.

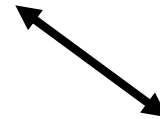
What is a database?

- A collection of data
 - structured
 - searchable (index) -> table of contents
 - updated periodically (release) -> new edition
 - cross-referenced ([hyperlinks](#)) -> links with other db
- Includes also associated tools (software) necessary for access, updating, information insertion, information deletion....
- Data storage management: flat files, relational databases...

Database: a « relational » example

Relational database (« table file »):

Teacher	Accession number	Education
Amos	1	Biochemistry
Dan	2	Genetics
John	3	Scientology



Course	Year	Involved teachers
Advanced Pottery	2000; 2001	1; 2
Ballet for Fat People	2001; 2002	2; 3

Why biological databases?

- Exponential growth in biological data.
- Data (genomic sequences, 3D structures, 2D gel analysis, MS analysis, Microarrays....) are no longer published in a conventional manner, but directly submitted to databases.
- Essential tools for biological research. The only way to publish massive amounts of data without using all the paper in the world.

Distribution of sequences

- Books, articles 1968 -> 1985
- Computer tapes 1982 -> 1992
- Floppy disks 1984 -> 1990
- CD-ROM 1989 ->
- FTP 1989 ->
- On-line services 1982 -> 1994
- WWW 1993 ->
- DVD 2001 ->

Some statistics

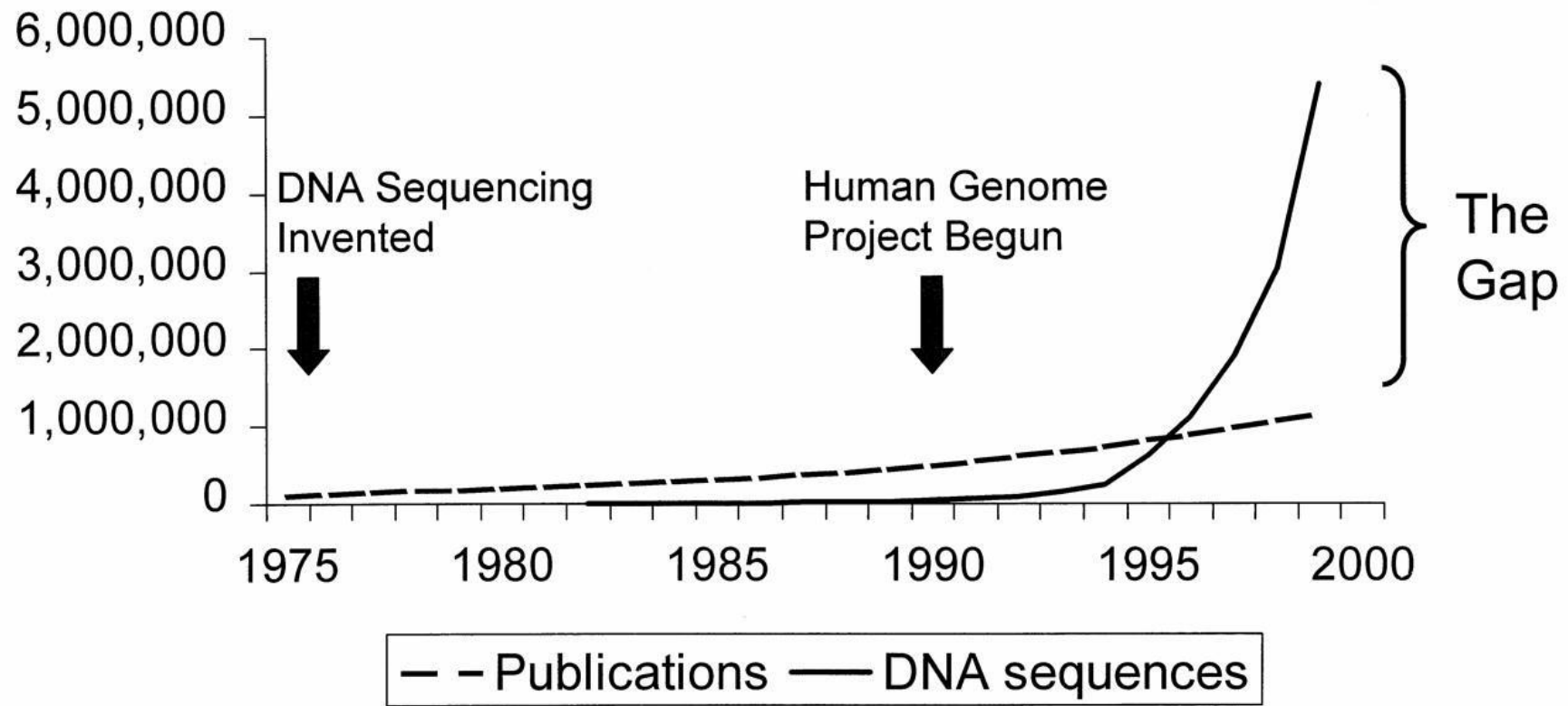
- More than 1000 different 'biological' databases
- Variable size: <100Kb to >20Gb
 - DNA: > 20 Gb
 - Protein: 1 Gb
 - 3D structure: 5 Gb
 - Other: smaller
- Update frequency: **daily** to **annually** to **seldom** to **forget about it**.
- Usually accessible through the web (some free, some not)

■ Some databases in the field of molecular biology...

AATDB, AceDb, ACUTS, ADB, AFDB, AGIS, AMSdb,
ARR, AsDb, BBDB, BCGD, Beanref, Biolmage,
BioMagResBank, BIOMDB, BLOCKS, BovGBASE,
BOVMAP, BSORF, BTKbase, CANSITE, CarbBank,
CARBHYD, CATH, CAZY, CCDC, CD4OLbase, CGAP,
ChickGBASE, Colibri, COPE, CottonDB, CSNDB, CUTG,
CyanoBase, dbCFC, dbEST, dbSTS, DDBJ, DGP, DictyDb,
Picty_cDB, DIP, DOGS, DOMO, DPD, DPInteract, ECDC,
ECGC, ECO2DBASE, EcoCyc, EcoGene, EMBL, EMD db,
ENZYME, EPD, EpoDB, ESTHER, FlyBase, FlyView,
GCRDB, GDB, GENATLAS, Genbank, GeneCards,
Genline, GenLink, GENOTK, GenProtEC, GIFTS,
GPCRDB, GRAP, GRBase, gRNAsdb, GRR, GSDB,
HAEMB, HAMSTERS, HEART-2DPAGE, HEXadb, HGMD,
HIDB, HIDC, HIVdb, HotMolecBase, HOVERGEN, HPDB,
HSC-2DPAGE, ICN, ICTVDB, IL2RGbase, IMGT, Kabat,
KDNA, KEGG, Klotho, LGIC, MAD, MaizeDb, MDB,
Medline, Mendel, MEROPS, MGDB, MGI, MHCPEP5
Micado, MitoDat, MITOMAP, MJDB, MmtDB, Mol-R-Us,
MPDB, MRR, MutBase, MycDB, NDB, NRSub, O-lycBase,
OMIA, OMIM, OPD, ORDB, OWL, PAHdb, PatBase, PDB,
PDD, Pfam, PhosphoBase, PigBASE, PIR, PKR, PMD,
PPDB, PRESAGE, PRINTS, ProDom, Prolysis, PROSITE,
PROTOMAP, RatMAP, RDP, REBASE, RGP, SBASE,
SCOP, SeqAnaiRef, SGD, SGP, SheepMap, Soybase,
SPAD, SRNA db, SRPDB, STACK, StyGene, Sub2D,
SubtiList, SWISS-2DPAGE, SWISS-3DIMAGE, SWISS-
MODEL Repository, SWISS-PROT, TelDB, TGN, tmRDB,
TOPS, TRANSFAC, TRR, UniGene, URNADB, V BASE,
VDRR, VectorDB, WDCM, WIT, WormPep, YEPD, YPD,
YPM, etc !!!!

Categories of databases for Life Sciences

- Sequences (DNA, protein)
- Genomics
- Mutation/polymorphism
- Protein domain/family
- Proteomics (2D gel, Mass Spectrometry)
- 3D structure
- Metabolic networks
- Regulatory networks
- Bibliography
- Expression (Microarrays,...)
- Specialized



()



Primary biological databases

- *Protein*

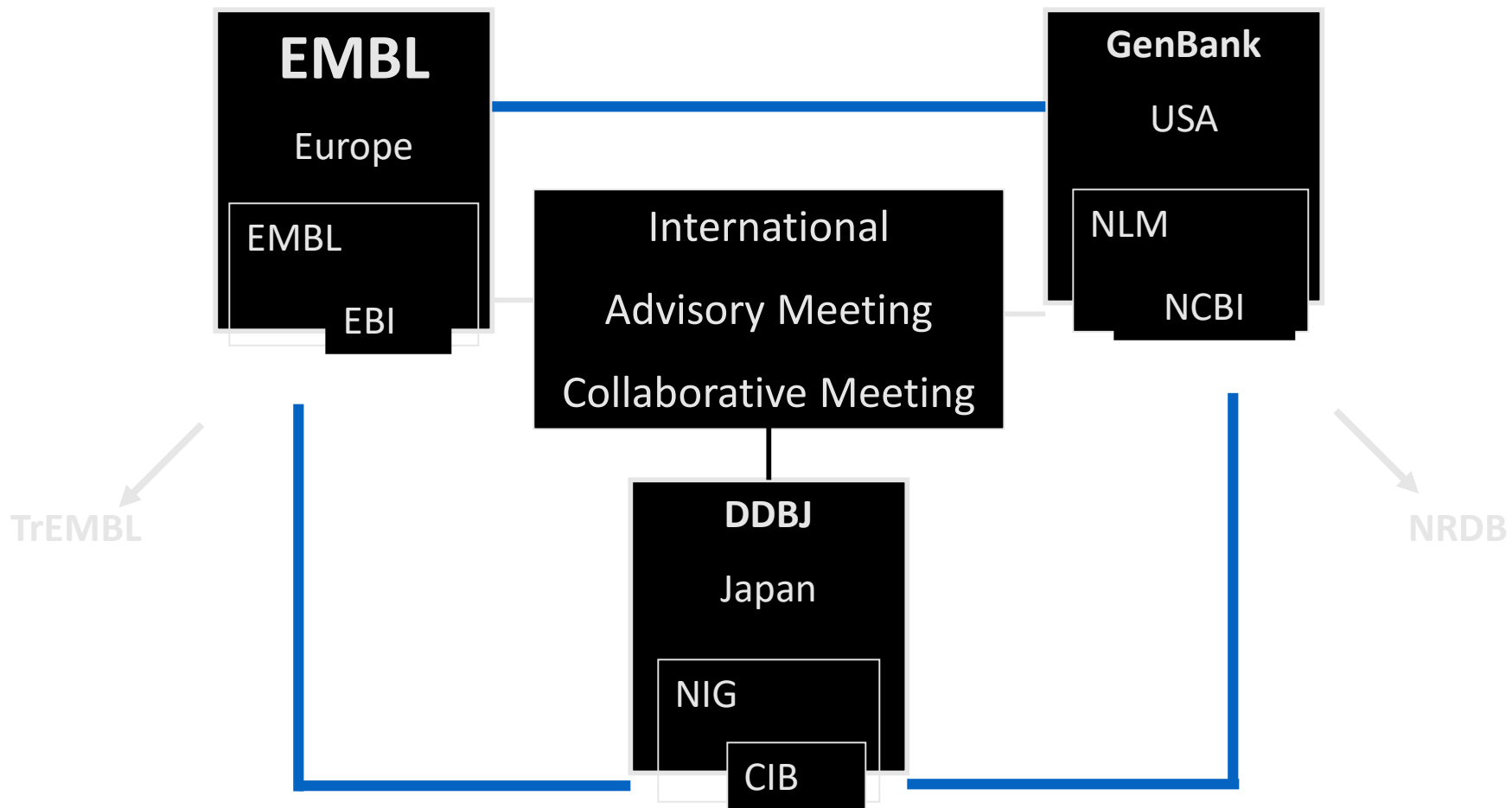
PIR

MIPS

SWISS-PROT

TrEMBL

NRL-3D



Biological data is highly **complex** and **interrelated**. Vast amount of biological information needs to be stored organized and indexed so that the information can be retrieved and used.

Biological databases are libraries of life sciences information, collected from scientific experiments, published literature, high-throughput experiment technology, and computational analyses.

- ✓ They contain information from research areas including genomics, proteomics, metabolomics, microarray gene expression, and phylogenetics.
- ✓ Information contained in biological databases includes gene function, structure, localization (both cellular and chromosomal), clinical effects of mutations as well as similarities of biological sequences and structures.

- Biological databases are an important tool in assisting scientists to understand and explain a host of biological phenomena like;
 - ✓ the structure of biomolecules and their interaction,
 - ✓ the whole metabolism of organisms
 - ✓ understanding the evolution of species.
- This knowledge helps facilitate the fight against **diseases**, assists in the **development of medications** and in discovering basic **relationships** amongst species in the history of life.
- There are five major types of databases namely
 - ✓ nucleotide databases,
 - ✓ protein databases,
 - ✓ protein structure databases,
 - ✓ metabolic pathway databases and
 - ✓ the bibliographic databases.

NCBI (National centre for Biotechnological information) :

NCBI is one of the leading online resources known for providing Biological sequence information.

As a national resource for molecular biology information, NCBI's mission is to develop new information technologies to aid in the understanding of fundamental molecular and genetic processes that control health and disease.

More specifically, the NCBI has been charged with creating automated systems for storing and analyzing knowledge about molecular biology, biochemistry, and genetics.

- NCBI is connected to various other sequence databases in order to be more efficient in answering sequence queries.
- The user queries and sequence information are delivered through NCBI's search tool called the "entrez".

NCBI:

<http://www.ncbi.nlm.nih.gov>

EBI:

<http://www.ebi.ac.uk/>

DDBJ:

<http://www.ddbj.nig.ac.jp/>

Entrez:

The NCBI database accepts queries and delivers data via a custom made search engine called Entrez. The Home page of NCBI has a search box which directs the user to entrez. Entrez is internally connected to various biological databases which increases the probability of getting the correct information.

GenBank:

The GenBank sequence database is an open access, annotated collection of all publicly available nucleotide sequences and their protein translations.

- This database is produced and maintained by the National Center for Biotechnology Information (NCBI)
- GenBank and its collaborators receive sequences produced in laboratories throughout the world from more than 100,000 distinct organisms.
- In more than 20 years since its establishment, GenBank has become the most important and most influential database for research in almost all biological fields, whose data were accessed and cited by millions of researchers around the world.

Literature Databases:

[Bookshelf](#): A collection of searchable biomedical books linked to PubMed.

[PubMed](#): Allows searching by author names, journal titles, and a new Preview/Index option. PubMed database provides access to over 12 million MEDLINE citations back to the mid-1960's. It includes History and Clipboard options which may enhance your search session.

[PubMed Central](#): The U.S. National Library of Medicine digital archive of life science journal literature.

[OMIM](#): Online Mendelian Inheritance in Man is a database of human genes and genetic disorders (also OMIA).



National Center for
Biotechnology Information

Search

Search

Clear

NCBI Home

Site Map (A-Z)

All Resources

Chemicals & Bioassays

Data & Software

DNA & RNA

Domains & Structures

Genes & Expression

Genetics & Medicine

Genomes & Maps

Homology

Literature

Proteins

Sequence Analysis

Taxonomy

Training & Tutorials

Variation

Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

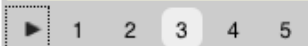
[About the NCBI](#) | [Mission](#) | [Organization](#) | [Research](#) | [RSS Feeds](#)

Get Started

- [Tools](#): Analyze data using NCBI software
- [Downloads](#): Get NCBI data or software
- [How-To's](#): Learn how to accomplish specific tasks at NCBI
- [Submissions](#): Submit data to GenBank or other NCBI databases

Genomic Structural Variation

dbVar archives large scale genomic variation data and associates defined variants with phenotypic information.



Popular Resources

- [BLAST](#)
- [Bookshelf](#)
- [Gene](#)
- [Genome](#)
- [Nucleotide](#)
- [OMIM](#)
- [Protein](#)
- [PubChem](#)
- [PubMed](#)
- [PubMed Central](#)
- [SNP](#)

NCBI News

[Preliminary genomic assemblies from two isolates from the European E. coli outbreak now available](#)

07 Jun 2011

[Preliminary genomic assemblies of two isolates are in the Nucleotide](#)

[New version of Cn3D \(v.4.3\) now available](#)

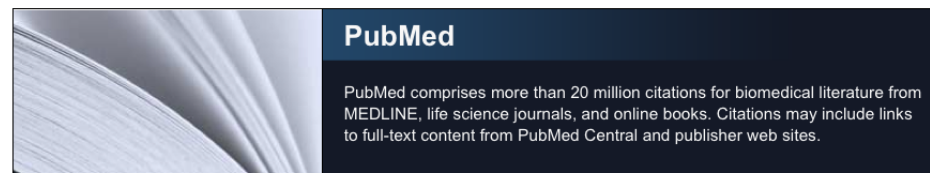
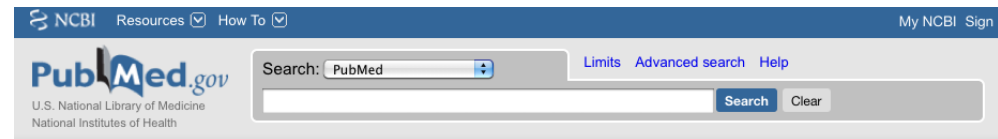
07 Jun 2011

[A new version of this popular 3D molecular visualization program](#)

[More...](#)

PubMed (Medline)

- MEDLINE covers the fields of medicine, nursing, dentistry, veterinary medicine, public health, and **preclinical sciences**
- Contains citations from approximately 5,200 worldwide journals in 37 languages; 60 languages for older journals.
- Contains over 20 million citations since 1948
- Contains links to biological db and to some journals
- New records are added to PreMEDLINE daily!



Using PubMed

[PubMed Quick Start Guide](#)

[Full Text Articles](#)

[PubMed FAQs](#)

[PubMed Tutorials](#)

[New and Noteworthy](#) 

PubMed Tools

[PubMed Mobile](#)

[Single Citation Matcher](#)

[Batch Citation Matcher](#)

[Clinical Queries](#)

[Topic-Specific Queries](#)

More Resources

[MeSH Database](#)

[Journals in NCBI Databases](#)

[Clinical Trials](#)

[E-Utilities](#)

[LinkOut](#)

PubMed :

This is an online Bibliographic database which has a collection of the research papers, journals and other bibliographic data. The Database is internally connected with other Bibliographic databases like Medline, Biomedcentral etc.

OMIM:

OMIM stand for Online Mendelian Inheritance in Man. This database contains information about the genetical disorders.

- ✓ OMIM gives complete data on the diseases the genetical background behind it and also the corresponding journal resources.

The OMIM (Online Mendelian Inheritance in Man)

- Genes and genetic disorders
- Edited by team at Johns Hopkins
- Updated daily

MIM Number Prefixes

- * gene with known sequence
- + gene with known sequence and phenotype
- # phenotype description, molecular basis known
- % mendelian phenotype or locus, molecular basis unknown
- no prefix other, mainly phenotypes with suspected mendelian basis

Searching OMIM

- Search Fields
 - Name of trait, e.g., hypertension
 - Cytogenetic location, e.g., 1p31.6
 - Inheritance, e.g., autosomal dominant
 - Gene, e.g., coagulation factor VIII

BLAST: BLAST stands for Basic Local Alignment Search Tool.

- BLAST is a tool that is used to find the sequences homologous to a particular sequence. BLAST compares all the sequences in the database with the one that is searched for and provides many hits which are usually arranged in the increasing order of the scores obtained.
- BLAST is available at the URL <http://blast.ncbi.nlm.nih.gov/>
- BLAST uses PAM and BLOSUM matrices for scoring the alignment.

(PAM: **P**oint **A**ccepted **M**utation or **P**ercent **A**ccepted **M**utation – represents accepted point mutation per 100 aa residues.)

Database of SNP's:

This database contains data about SNP's (Single Nucleotide polymorphism)

Pubchem :

This contains data about the chemical compounds that are used for insillico analysis

OMIA:

This database is similar to OMIM, but contains data about the diseases of all the other animals at the genetic level except human.

- **ENSEMBL-Genome Browser:** is a gene annotation system which creates and stores predicted gene locations. It is freely available.
- **UCSC Genome Bioinformatics** (<http://genome.ucsc.edu/>): It is a genome browser where the latest human genome assembly is freely available to view.
- **UniProt** (<http://www.uniprot.org/>) :Includes protein sequences and their functional information.
- **Protein Data Bank (PDB)** (<http://www.rcsb.org/pdb/home/home.do>) : Includes three-dimensional structural data of proteins as well as nucleic acids.

Retrieve the gene sequence in FASTA format

Intro:

- A gene is a molecular unit of heredity of a living organism. It is a name given to some stretches of DNA and RNA that code for a type of protein or for an RNA chain that has a function in the organism.
- Knowledge of gene sequences has become indispensable for basic biological research, other research branches utilizing sequencing, and in numerous applied fields such as diagnostic, biotechnology, forensic biology and biological systematics.
- In bioinformatics, FASTA format is a text-based format for representing either nucleotide sequences or peptide sequences, in which nucleotides or amino acids are represented using single-letter codes.

- The format also allows for sequence names and comments to precede the sequences.
- The format originates from the FASTA software package, but has now become a standard in the field of bioinformatics.
- The simplicity of FASTA format makes it easy to manipulate and parse sequences using text-processing tools and scripting languages.

- **Retrieve any one FASTA sequence of GABA transaminase in Human, mouse, pig and chick.**

Aim: To retrieve any one FASTA sequence of GABA transaminase in Human, mouse, pig and chick

Introduction:

- 4-aminobutyrate aminotransferase (or GABA transaminase) is an enzyme which catalyzes the conversion of 4-aminobutanoic acid (GABA) and 2-oxoglutarate into succinic semialdehyde and glutamate.

Method:

1. Open NCBI <http://www.ncbi.nlm.nih.gov/>
2. Choose the Protein Database and enter GABA transaminase in the search box
3. Click on Advanced Search Tab and Choose the Organism option from the drop down menu.
4. Enter Homo sapiens and the results are displayed.
5. The above steps can be repeated by entering the Organism name as Sus scrofa(Pig), Mus musculus(Mouse) and Gallus gallus(Chick).