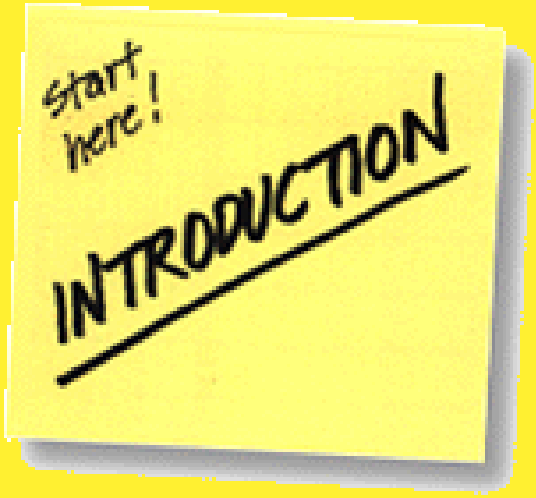
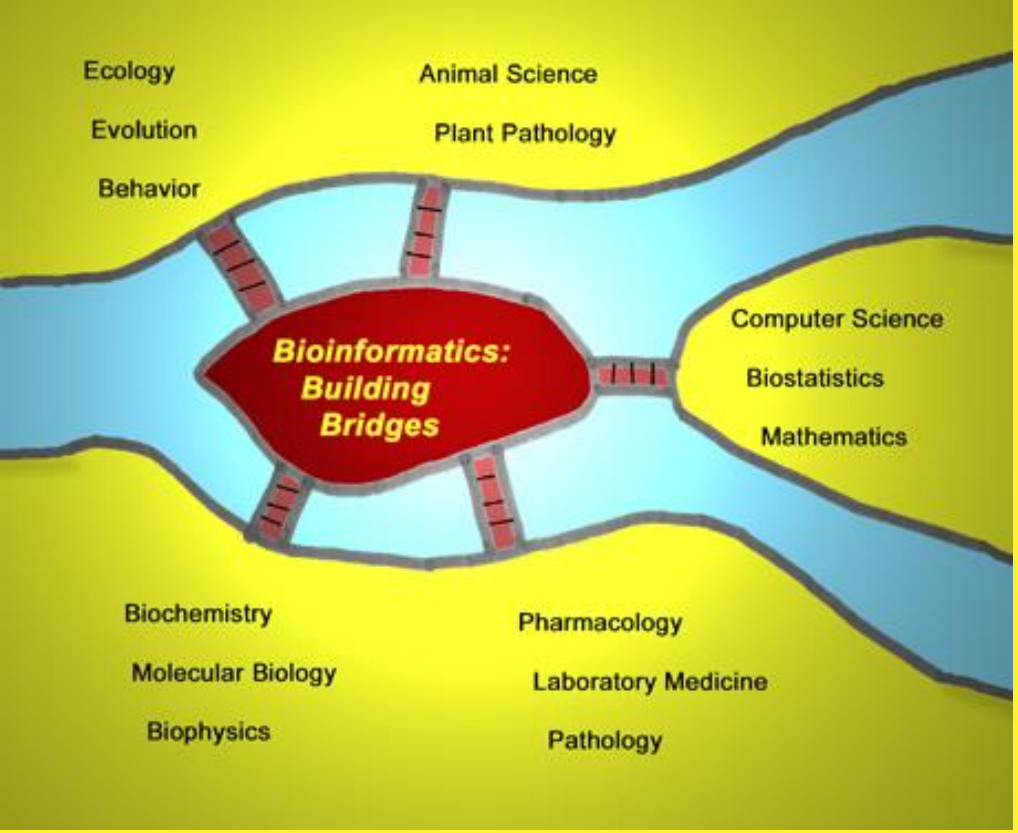
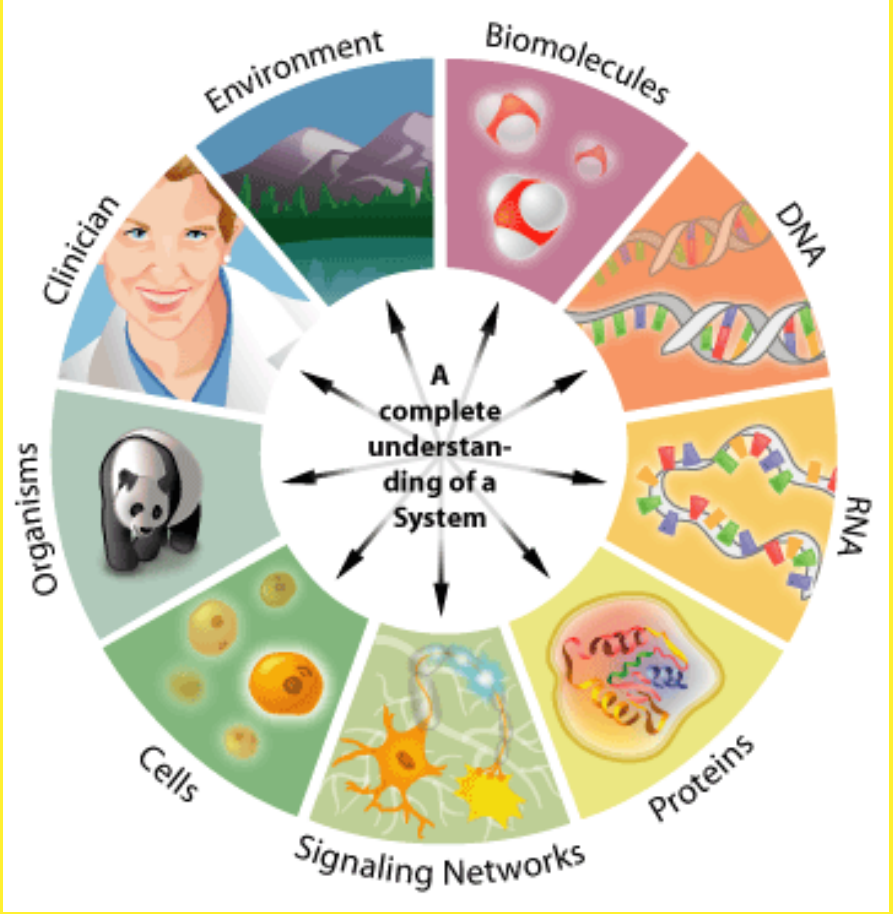



# *Introduction to Bioinformatics*




# What is Bioinformatics?

# *NIH* – definitions

**What is Bioinformatics? -** Research, development, and application of computational tools and approaches for expanding the use of biological, medical, behavioral, and health data, including the means to acquire, store, organize, archive, analyze, or visualize such data.

**on molecular**

**What is Computational Biology?** The development and application of analytical and theoretical methods, mathematical modeling and computational simulation techniques to the study of biological, behavioral, and social data.

# *Molecular Bioinformatics*

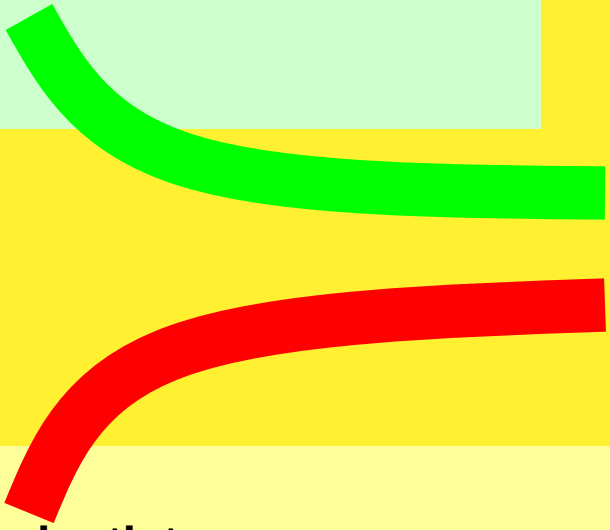
**Molecular Bioinformatics** involves the use of computational tools to discover new information in complex data sets (from the one-dimensional information of DNA through the two-dimensional information of RNA and the three-dimensional information of proteins, to the four-dimensional information of evolving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

Bioinformatics (Oxford English Dictionary):

The branch of science concerned with information and information flow in biological systems, esp. the use of computational methods in genetics and genomics.

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.

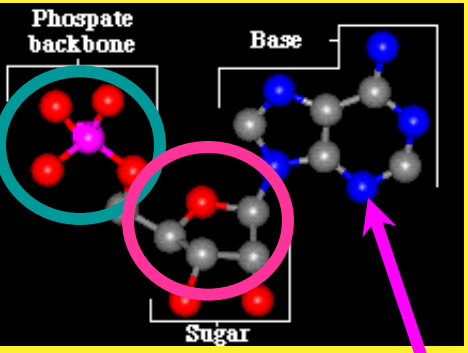
Some biological background….

A biologist

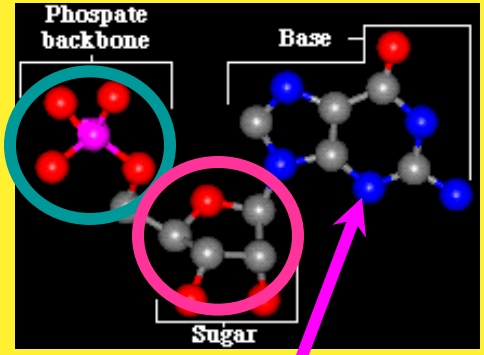
The hereditary information of all living organisms, with the exception of some viruses, is carried by **deoxyribonucleic acid (DNA)** molecules.

2 purines:

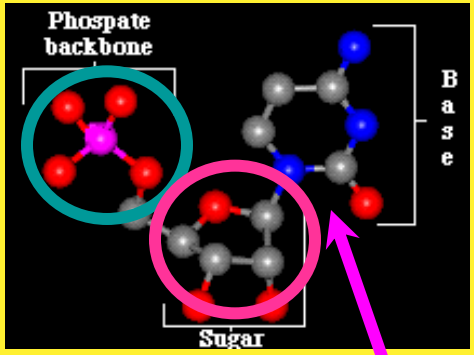
2 pyrimidines:



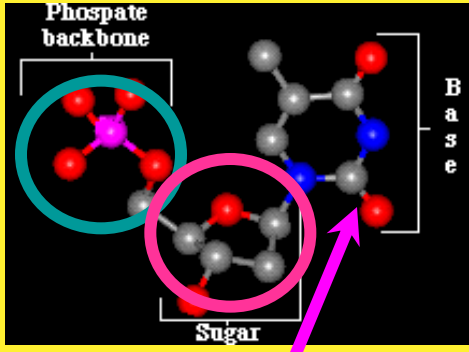

adenine (A)




guanine (G)




cytosine (C)




thymine (T)

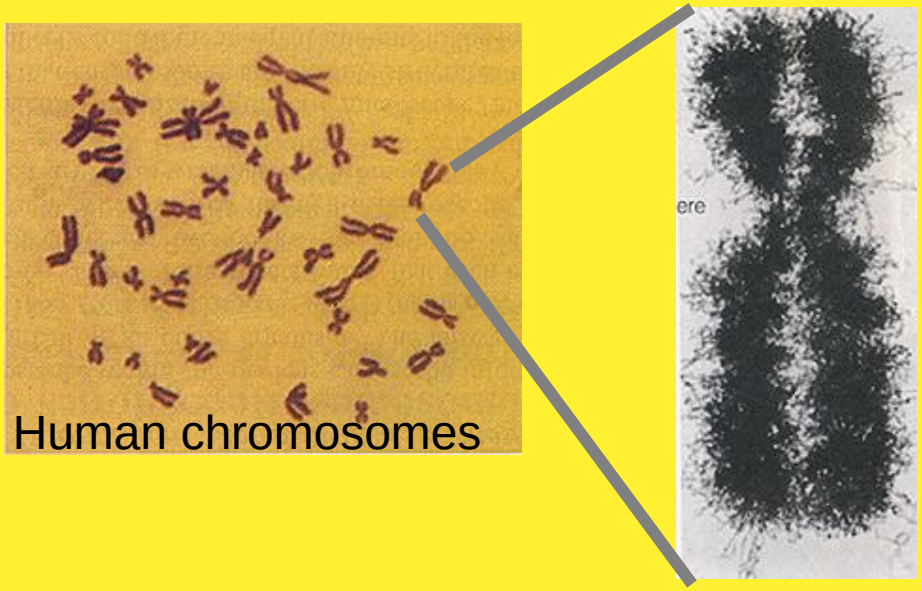
*two rings*

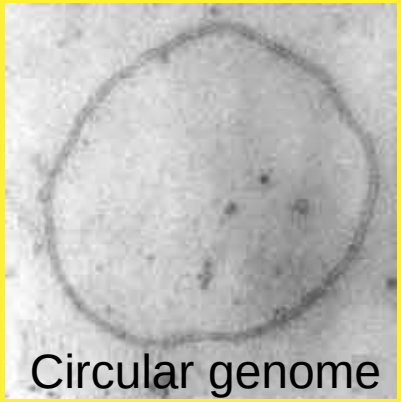
*one ring*

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

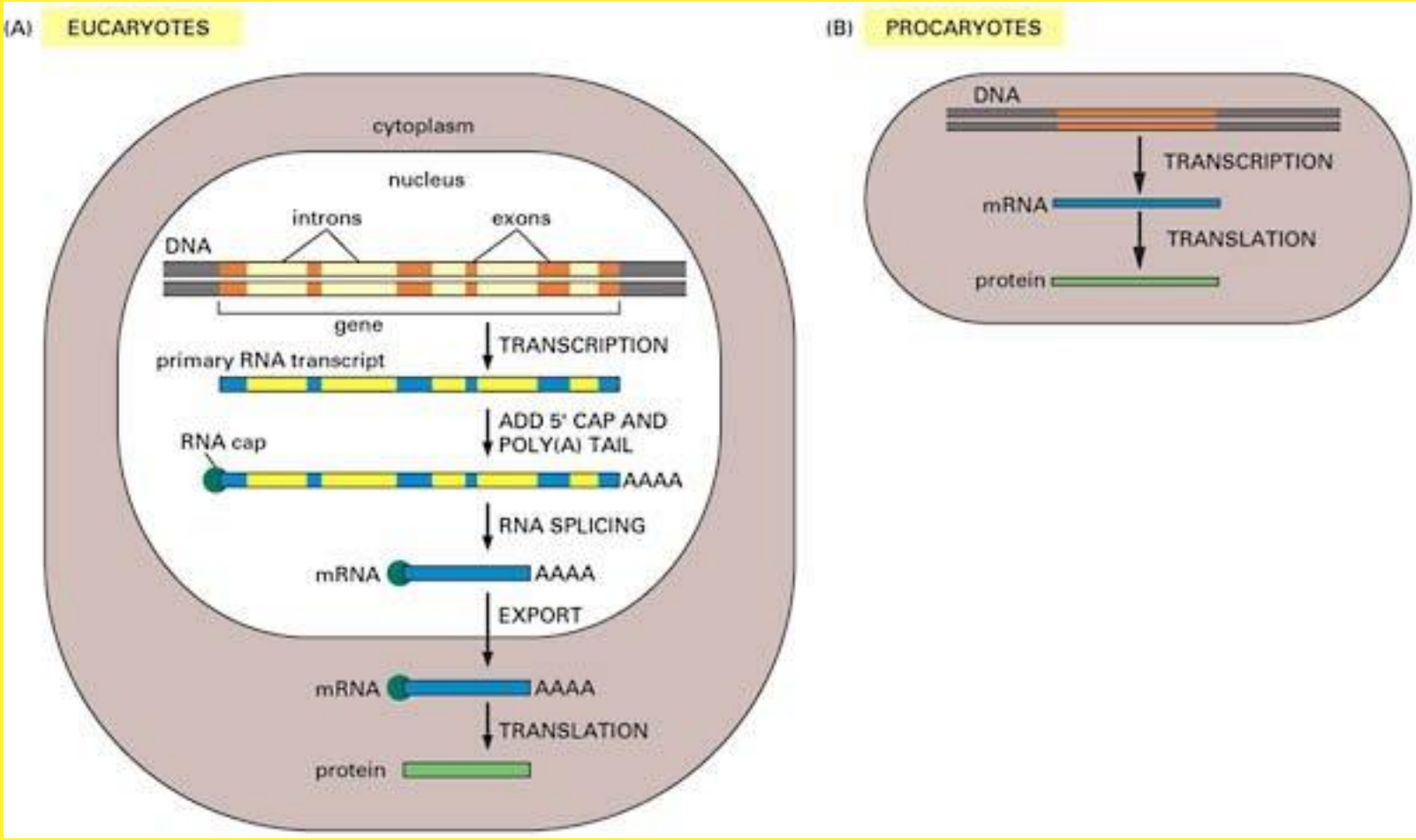
Bacteria have either circular or linear genomes and may also carry plasmids


Human chromosomes

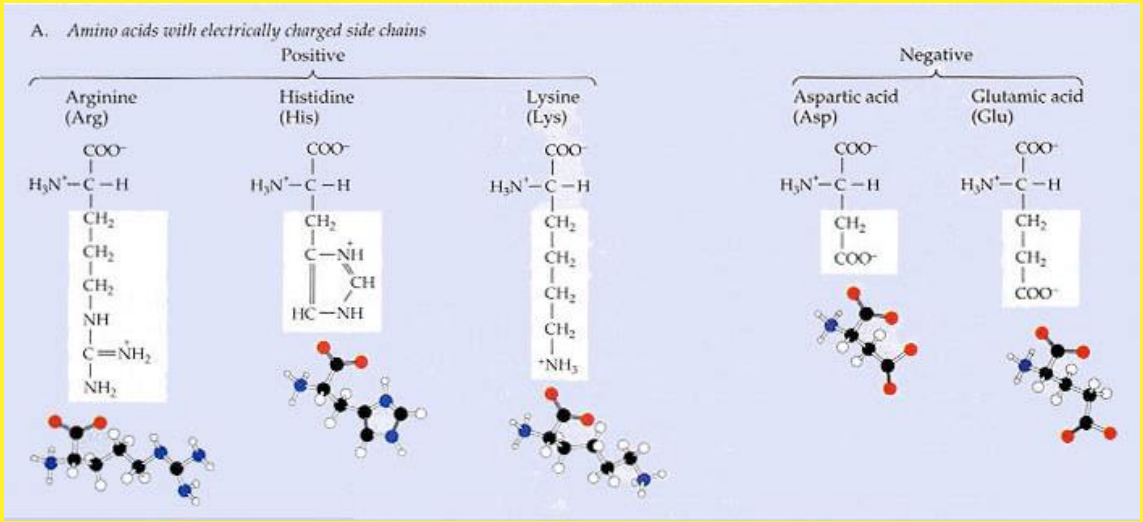

Circular genome

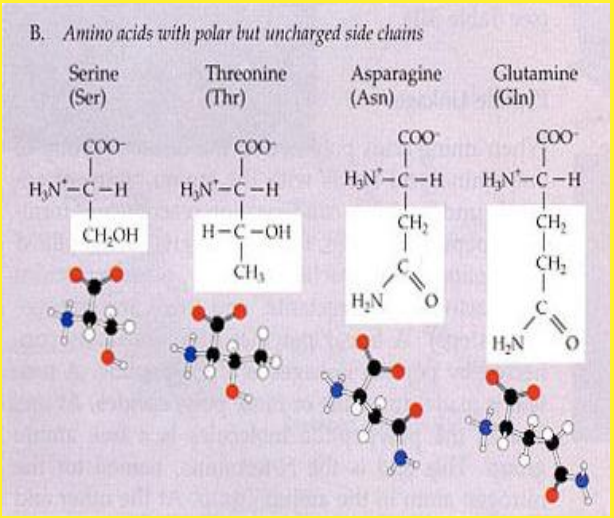
**Central dogma: DNA makes RNA makes Protein**

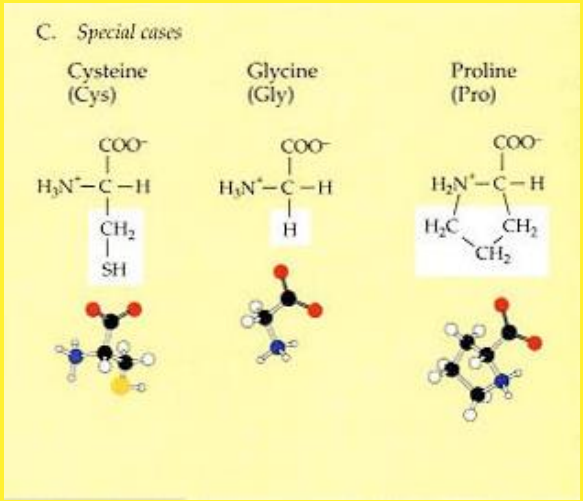
**Modified dogma: DNA makes DNA and RNA, RNA makes DNA, RNA an Protein**

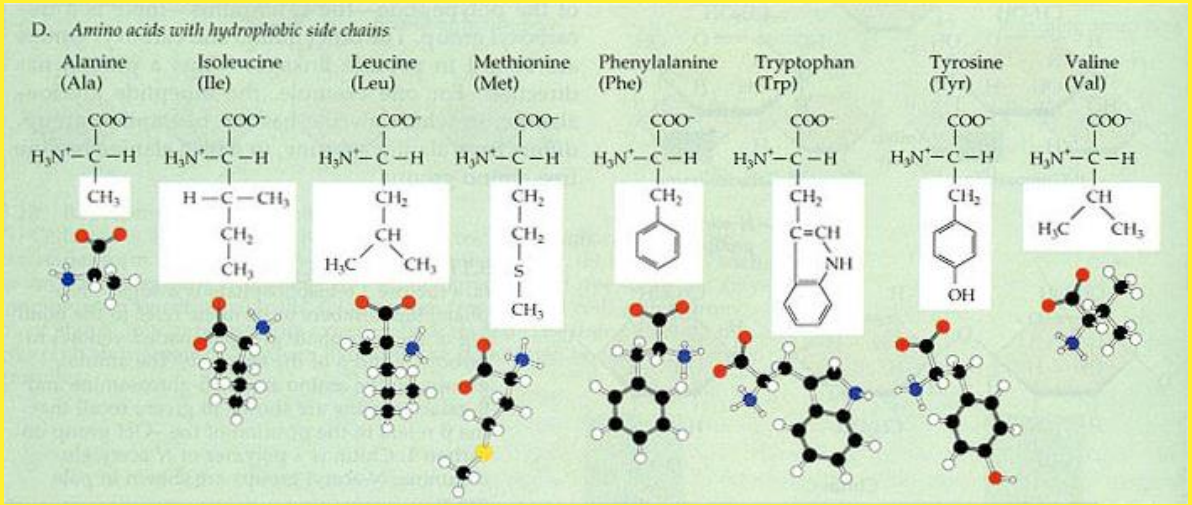


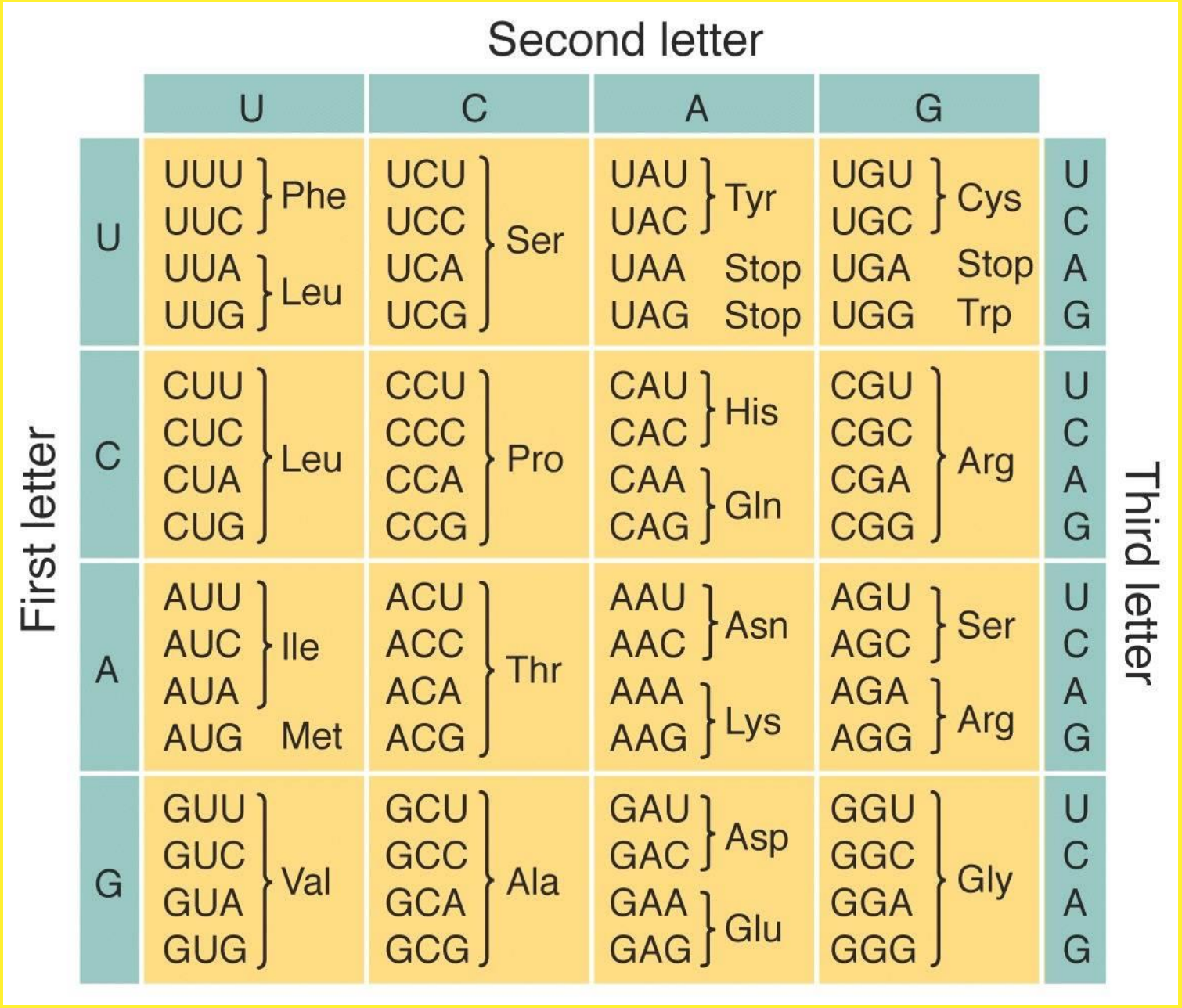
**Amino acids** - The protein building blocks



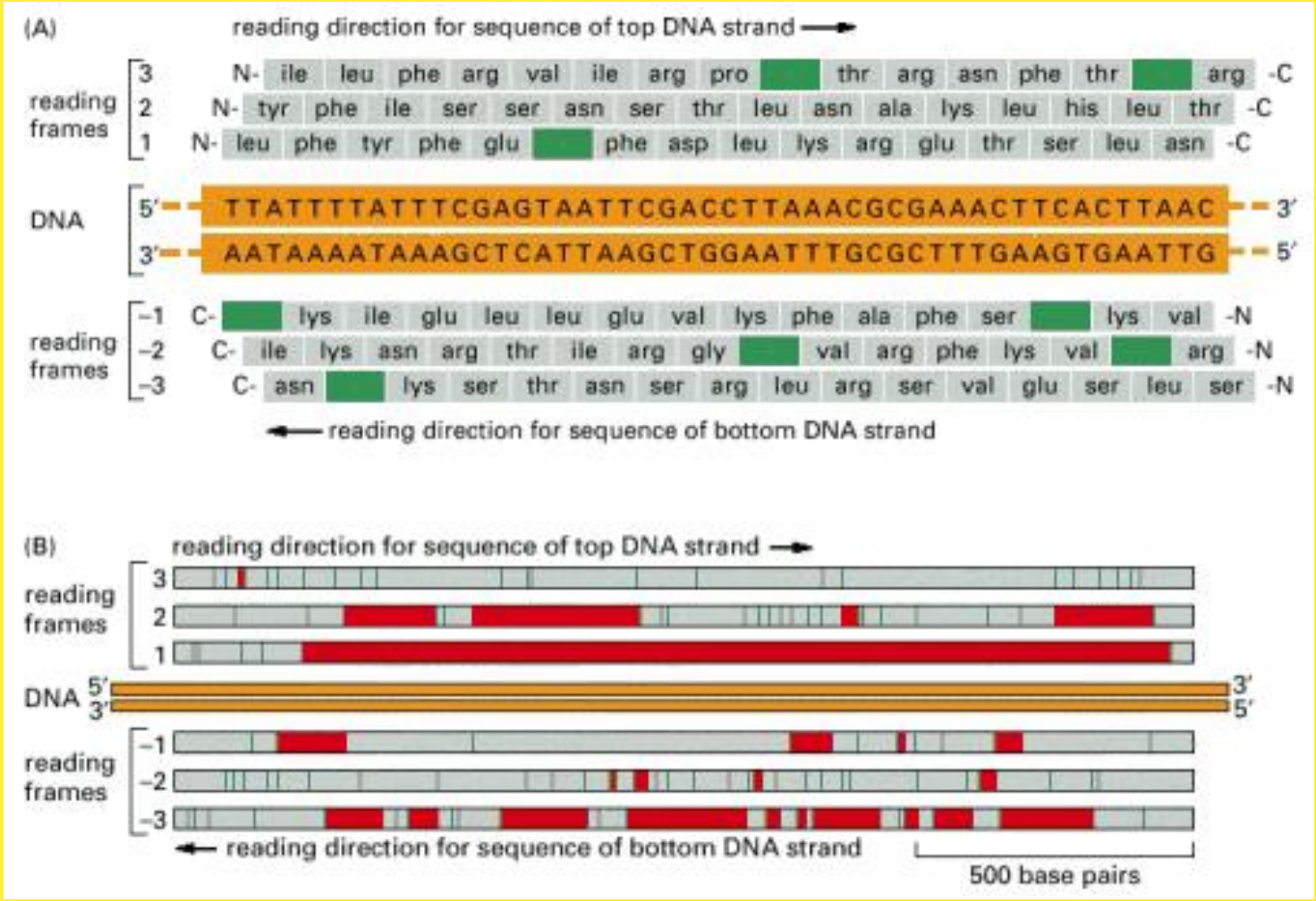




Second letter
U
C
A
G
First letter
Third letter
U
UUU Phe
UUC
UUA Leu
UUG
UCU Ser
UCC
UCA
UCG
UAU Tyr
UAC
UAA Stop
UAG Stop
UGU Cys
UGC
UGA Stop
UGG Trp
C
CUU Leu
CUC
CUA
CUG
CCU Pro
CCC
CCA
CCG
CAU His
CAC
CAA Gln
CAG
CGU Arg
CGC
CGA
CGG
A
AUU Ile
AUC
AUA
AUG Met
ACU Thr
ACC
ACA
ACG
AAU Asn
AAC
AAA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

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.



Protein folding

A human Hemoglobin:

How does it all looks like on a computer monitor?

## A cDNA sequence

>gi|14456711|ref|NM_000558.3| **Homo sapiens hemoglobin, alpha** 1 (HBA1), mRNA
ACTCTTCTGGTCCCCACAGACTCAGAGAGAACCCACCATGGTGCTGTCTCCTGCCGACAAGACCAA
CGTCAAGGCCGCCTGGGGTAAGGTCGGCGCGCACGCTGGCGAGTATGGTGCGGAGGCCCTGGAG
AGGATGTTCCTGTCCTTCCCCACCACCAAGACCTACTTCCCGCACTTCGACCTGAGCCACGGCTCT
GCCCAGGTTAAGGGCCACGGCAAGAAGGTGGCCGACGCGCTGACCAACGCCGTGGCGCACGTGG
ACGACATGCCCAACGCGCTGTCCGCCCTGAGCGACCTGCACGCGCACAAGCTTCGGGTGGACCCG
GTCAACTTCAAGCTCCTAAGCCACTGCCTGCTGGTGACCCTGGCCGCCCACCTCCCCGCCGAGTTC
ACCCCTGCGGTGCACGCCTCCCTGGACAAGTTCCTGGCTTCTGTGAGCACCGTGCTGACCTCCAAA
TACCGTTAAGCTGGAGCCTCGGTGGCCATGCTTCTTGCCCCTTGGGCCTCCCCCCAGCCCCTCCTC
CCCTTCCTGCACCCGTACCCCCGTGGTCTTTGAATAAAGTCTGAGTGGGCGGC

A cDNA sequence (reading frame)

>gi|14456711|ref|NM_000558.3| **Homo sapiens hemoglobin, alpha** 1 (HBA1), mRNA

ACTCTTCTGGTCCCCACAGACTCAGAGAGAACCCACCATGGTGCTGTCTCCTGCCGACAAGACCA
ACGTCAAGGCCGCCTGGGGTAAGGTCGGCGCGCACGCTGGCGAGTATGGTGCGGAGGCCCTGGA
GAGGATGTTCCTGTCCTTCCCCACCACCAAGACCTACTTCCCGCACTTCGACCTGAGCCACGGCTC
TGCCCAGGTTAAGGGCCACGGCAAGAAGGTGGCCGACGCGCTGACCAACGCCGTGGCGCACGTG
GACGACATGCCCAACGCGCTGTCCGCCCTGAGCGACCTGCACGCGCACAAGCTTCGGGTGGACCC
GGTCAACTTCAAGCTCCTAAGCCACTGCCTGCTGGTGACCCTGGCCGCCCACCTCCCCGCCGAGTT
CACCCCTGCGGTGCACGCCTCCCTGGACAAGTTCCTGGCTTCTGTGAGCACCGTGCTGACCTCCAA
ATACCGTTAAGCTGGAGCCTCGGTGGCCATGCTTCTTGCCCCTTGGGCCTCCCCCCAGCCCCTCC
TCCCCTTCCTGCACCCGTACCCCCGTGGTCTTTGAATAAAGTCTGAGTGGGCGGC

A protein sequence

>gi|4504347|ref|NP_000549.1| **alpha 1 globin [Homo sapiens]**

MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHGKKVA
DALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFL
ASVSTVLTSKYR

And, a whole genome…

ACTCTTCTGGTCCCCACAGACTCAGAGAGAACCCACCATGGTGCTGTCTCCTGCCGACAAGACCAACGTC
AAGGCCGCCTGGGGTAAGGTCGGCGCGCACGCTGGCGAGTATGGTGCGGAGGCCCTGGAGAGGATGTT
CCTGTCCTTCCCCACCACCAAGACCTACTTCCCGCACTTCGACCTGAGCCACGGCTCTGCCCAGGTTAAG
GGCCACGGCAAGAAGGTGGCCGACGCGCTGACCAACGCCGTGGCGCACGTGGACGACATGCCCAACGC
GCTGTCCGCCCTGAGCGACCTGCACGCGCACAAGCTTCGGGTGGACCCGGTCAACTTCAAGCTCCTAAG
CCACTGCCTGCTGGTGACCCTGGCCGCCCACCTCCCCGCCGAGTTCACCCCTGCGGTGCACGCCTCCCT
GGACAAGTTCCTGGCTTCTGTGAGCACCGTGCTGACCTCCAAATACCGTTAAGCTGGAGCCTCGGTGGCC
ATGCTTCTTGCCCCTTGGGCCTCCCCCCAGCCCCTCCTCCCCTTCCTGCACCCGTACCCCCGTGGTCTTT
GAATAAAGTCTGAGTGGGCGGCACTCTTCTGGTCCCCACAGACTCAGAGAGAACCCACCATGGTGCTGTC
TCCTGCCGACAAGACCAACGTCAAGGCCGCCTGGGGTAAGGTCGGCGCGCACGCTGGCGAGTATGGTGC
GGAGGCCCTGGAGAGGATGTTCCTGTCCTTCCCCACCACCAAGACCTACTTCCCGCACTTCGACCTGAGC
CACGGCTCTGCCCAGGTTAAGGGCCACGGCAAGAAGGTGGCCGACGCGCTGACCAACGCCGTGGCGCA
CGTGGACGACATGCCCAACGCGCTGTCCGCCCTGAGCGACCTGCACGCGCACAAGCTTCGGGTGGACCC
GGTCAACTTCAAGCTCCTAAGCCACTGCCTGCTGGTGACCCTGGCCGCCCACCTCCCCGCCGAGTTCACC
CCTGCGGTGCACGCCTCCCTGGACAAGTTCCTGGCTTCTGTGAGCACCGTGCTGACCTCCAAATACCGTT
AAGCTGGAGCCTCGGTGGCCATGCTTCTTGCCCCTTGGGCCTCCCCCCAGCCCCTCCTCCCCTTCCTGCA
CCCGTACCCCCGTGGTCTTTGAATAAAGTCTGAGTGGGCGGCACTCTTCTGGTCCCCACAGACTCAGAGA
GAACCCACCATGGTGCTGTCTCCTGCCGACAAGACCAACGTCAAGGCCGCCTGGGGTAAGGTCGGCGCG
CACGCTGGCGAGTATGGTGCGGAGGCCCTGGAGAGGATGTTCCTGTCCTTCCCCACCACCAAGACCTACT
TCCCGCACTTCGACCTGAGCCACGGCTCTGCCCAGGTTAAGGGCCACGGCAAGAAGGTGGCCGACGCGC
TGACCAACGCCGTGGCGCACGTGGACGACATGCCCAACGCGCTGTCCGCCCTGAGCGACCTGCACGCGC
ACAAGCTTCGGGTGGACCCGGTCAACTTCAAGCTCCTAAGCCACTGCCTGCTGGTGACCCTGGCCGCCCA
CCTCCCCGCCGAGTTCACCCCTGCGGTGCACGCCTCCCTGGACAAGTTCCTGGCTTCTGTGAGCACCGTG
CTGACCTCCAAATACCGTTAAGCTGGAGCCTCGGTGGCCATGCTTCTTGCCCCTTGGGCCTCCCCCCAGC
CCCTCCTCCCCTTCCTGCACCCGTACCCCCGTGGTCTTTGAATAAAGTCTGAGTGGGCGGCGCCGTGGCG
CACGTGGACGACATGCCCAACGCGCTGTCCGCCCTGAGCGACCTGCACGCGCACAAGCTTCGGGTGGAC
CCGGTCAACTTCAAGCTCCTAAGCCACTGCCTGCTGGTGACCCTGGCCGCCCACCTCCCCGCCGAGTTCA
CCCCTGCGGTGCACGCCTCCCTGGACAAGTTCCTGGCTTCTGTGAGCACCGTGCTGACCTCCAAATACCG
TTAAGCTGGAGCCTCGGTGGCCATGCTTCTTGCCCCTTGGGCCTCCCCCCAGCCCCTCCTCCCCTTCCTG

How big are whole genomes?

*E. coli* 4.6 x $10^6$ nucleotides

- Approx. 4,000 genes

Yeast 15 x $10^6$ nucleotides

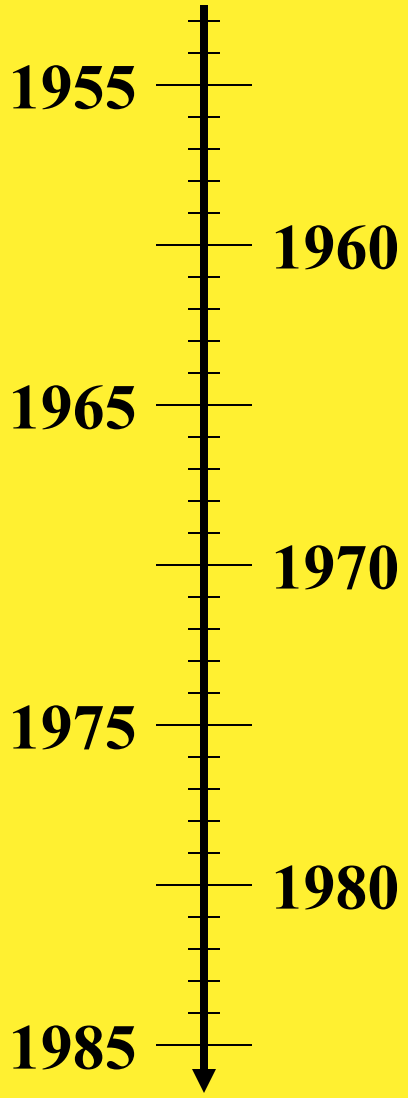
- Approx. 6,000 genes

Human 3 x $10^9$ nucleotides

- Approx. 30,000 genes

Smallest human chromosome 50 x $10^6$ nucleotides

# From DNA to Genome




Watson and Crick DNA model

Sequence alignment

PDB (Protein Data Bank)

GenBank database

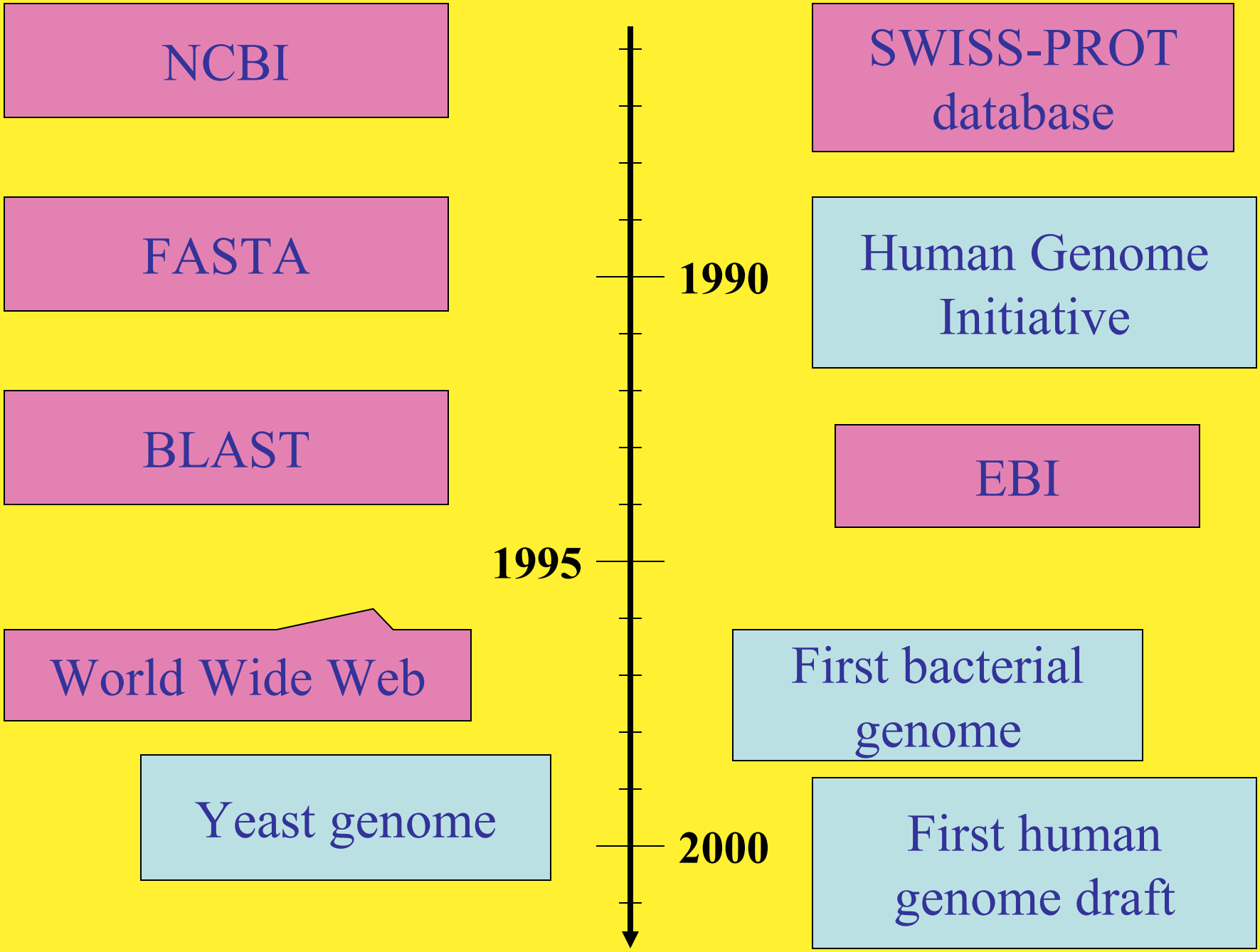
Sanger sequences insulin protein

Dayhoff's Atlas

ARPANET (early Internet)

Sanger dideoxy DNA sequencing

PCR (Polymerase Chain Reaction)


NCBI
SWISS-PROT database
FASTA
1990
Human Genome Initiative
BLAST
EBI
1995
World Wide Web
First bacterial genome
Yeast genome
2000
First human genome draft

# 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*).

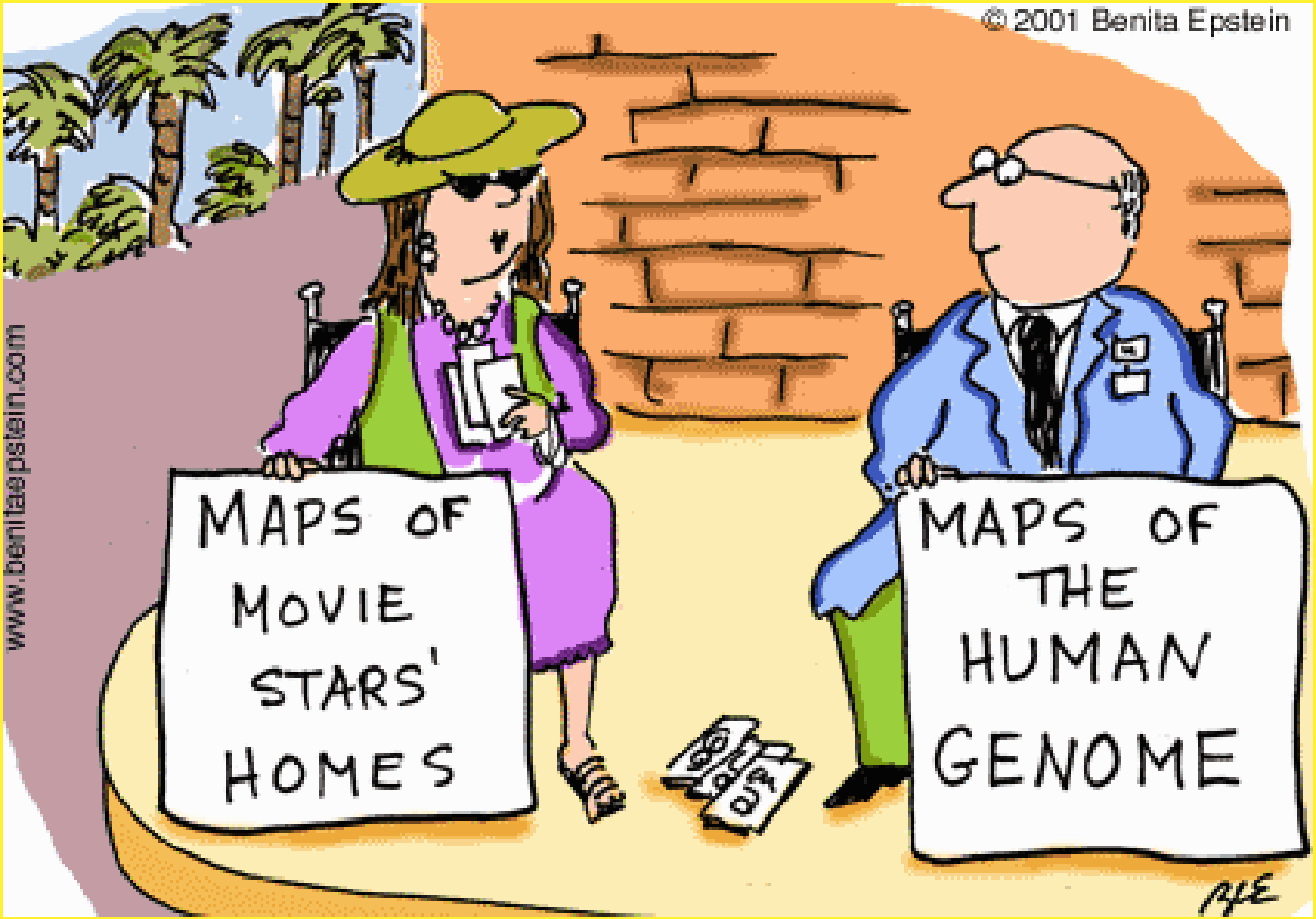
The Protein DataBank followed in 1972 with a collection of ten X-ray crystallographic protein structures. The SWISSPROT protein sequence database began in 1987.

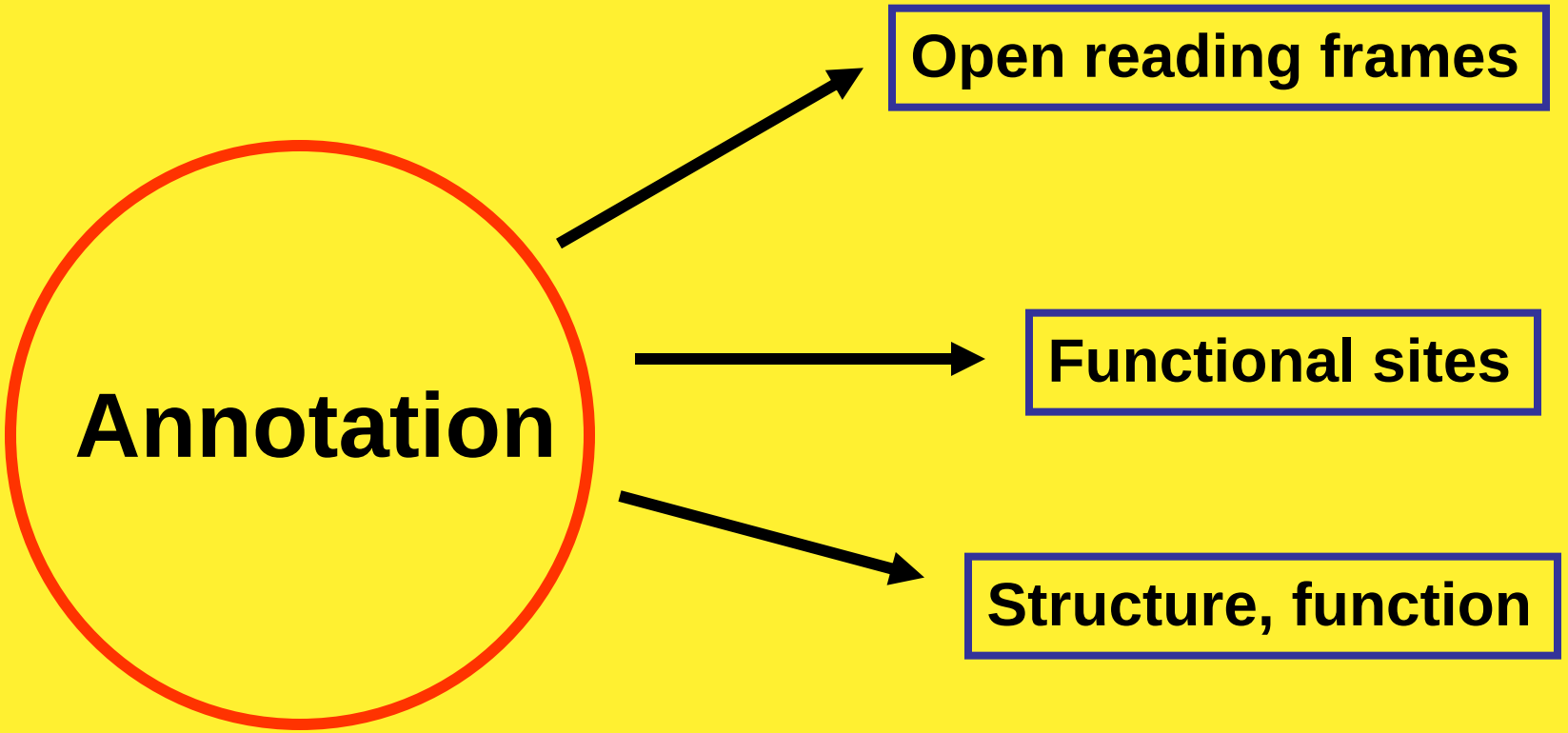
# Complete Genomes

**as of August 2011:**

| | |
|---|---|
| **Eukaryotes** | **37** |
| **Prokaryotes** | **1708** |
| **Total** | **1745** |

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




Annotation
Open reading frames
Functional sites
Structure, function

**CCTGACAAATTCGACGTGCGGCATTGCATGCAGACGTGCATG**

**CGTGCAAATAATCAATGTGGACTTTTCTGCGATTATGGAAGAA**

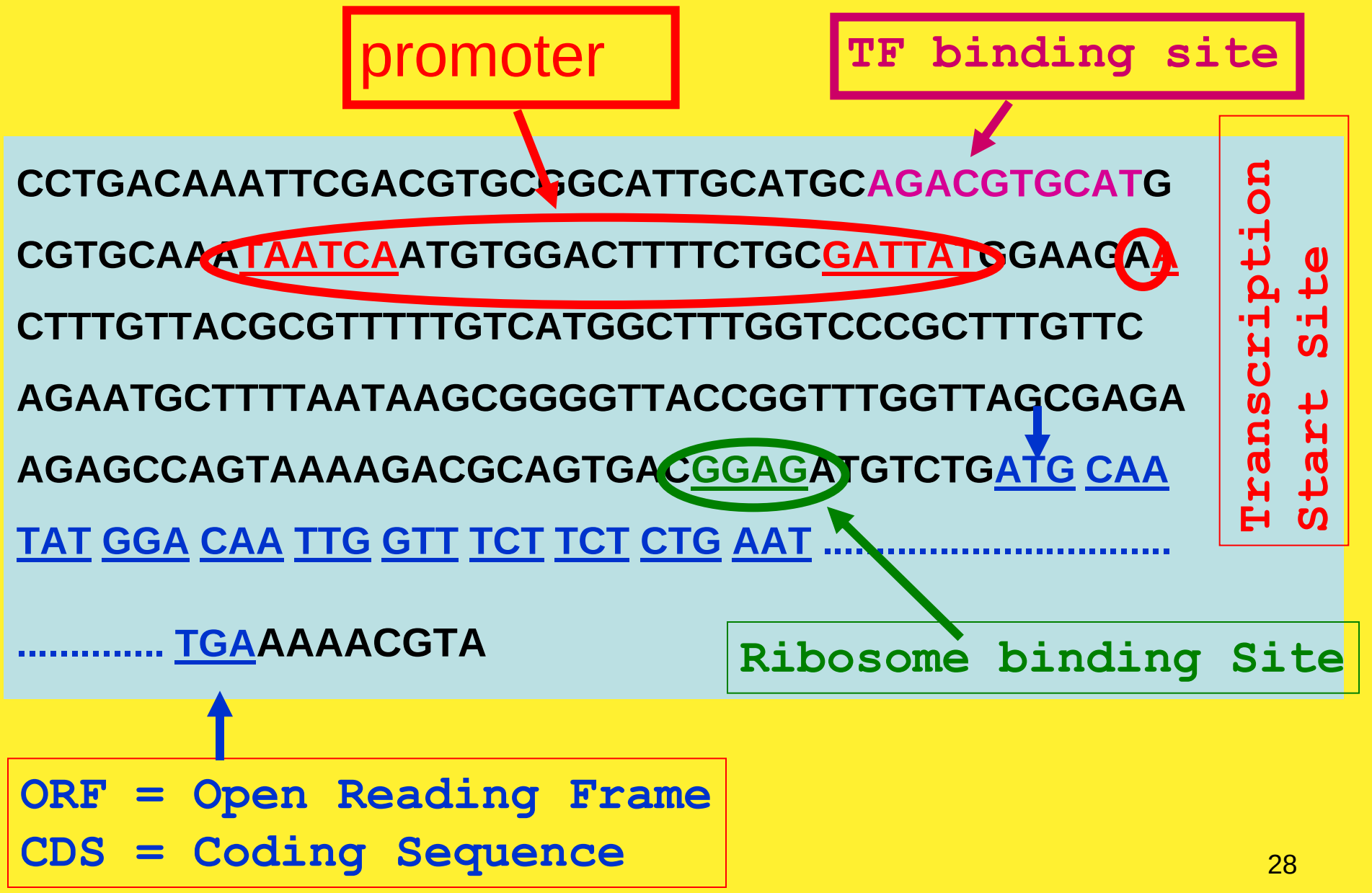
**CTTTGTTACGCGTTTTTGTCATGGCTTTGGTCCCGCTTTGTTC**

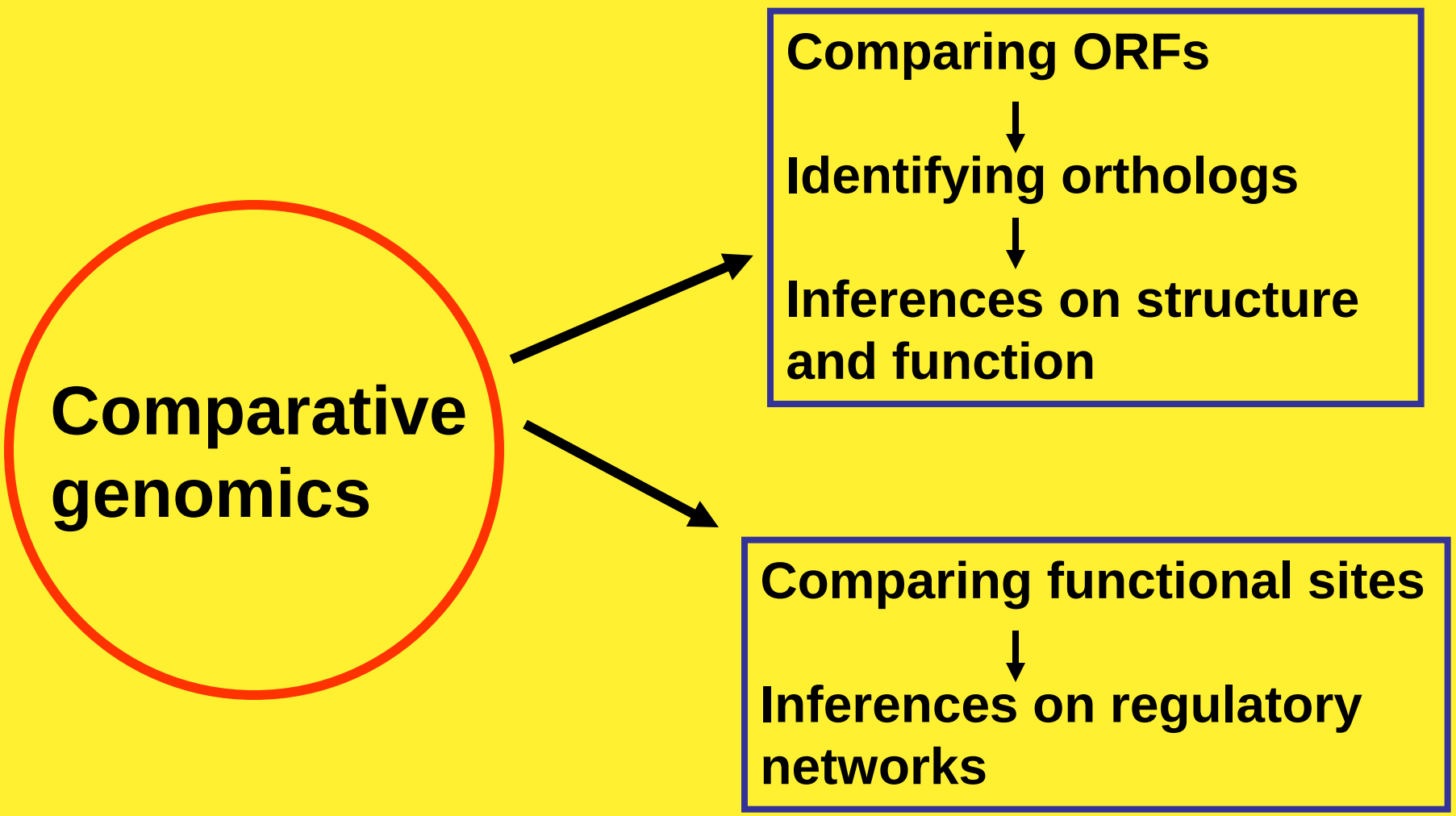
**AGAATGCTTTTAATAAGCGGGGTTACCGGTTTGGTTAGCGAGA**

**AGAGCCAGTAAAAGACGCAGTGACGGAGATGTCTGATG CAA**

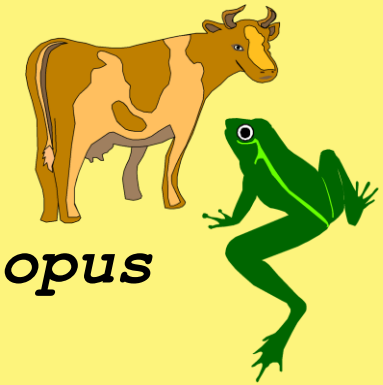
**TAT GGA CAA TTG GTT TCT TCT CTG AAT .....**

**..... TGAAAAACGTA**


promoter
TF binding site
CCTGACAAATTCGACGTGCGGCATTGCATGCAGACGTGCATG
CGTGCAAATAATCAATGTGGACTTTTCTGCGATTATCGGAAGAA
CTTTGTTACGCGTTTTTGTCATGGCTTTGGTCCCGCTTTGTTC
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


Comparative genomics
Comparing ORFs
Identifying orthologs
Inferences on structure and function
Comparing functional sites
Inferences on regulatory networks

# Alignment preproinsulin

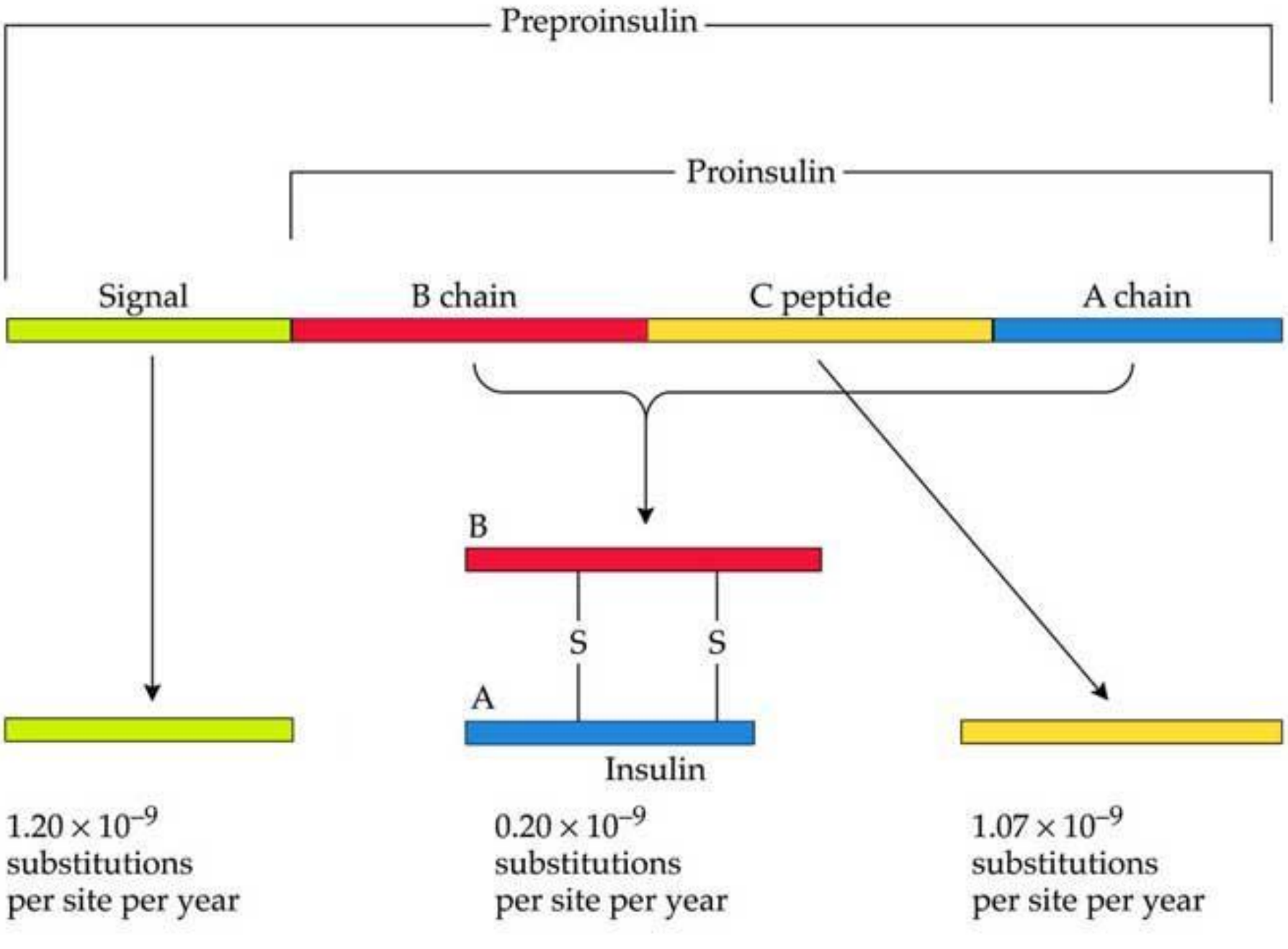


```
Xenopus   MALWMQCLP-LVLVLLFSTPNTEALANQHL
Bos       MALWTRLRPLLALLALWPPPPARAFVNQHL
          **** :   * *.*: *:..* :. *:****

Xenopus   CGSHLVEALYLVCGDRGFFYYPKIKRDIEQ
Bos       CGSHLVEALYLVCGERGFFYTPKARREVEG
          ****************:***** ** :*::*

Xenopus   AQVNGPQDNELDG-MQFQPQEYQKMKRGIV
Bos       PQVG---ALELAGGPGAGGLEGPPQKRGIV
          .**.      ** *       *     *****

Xenopus   EQCCHSTCSLFQLENYCN
Bos       EQCCASVCSLYQLENYCN
          **** *.***:*******
```


Preproinsulin
Proinsulin
Signal
B chain
C peptide
A chain
B
S
S
A
Insulin
1.20 × 10−9
substitutions
per site per year
0.20 × 10−9
substitutions
per site per year
1.07 × 10−9
substitutions
per site per year

**Functional genomics**

**Genome-wide profiling of:**
- **mRNA levels**
- **Protein levels**

↓

**Co-expression of genes and/or proteins**

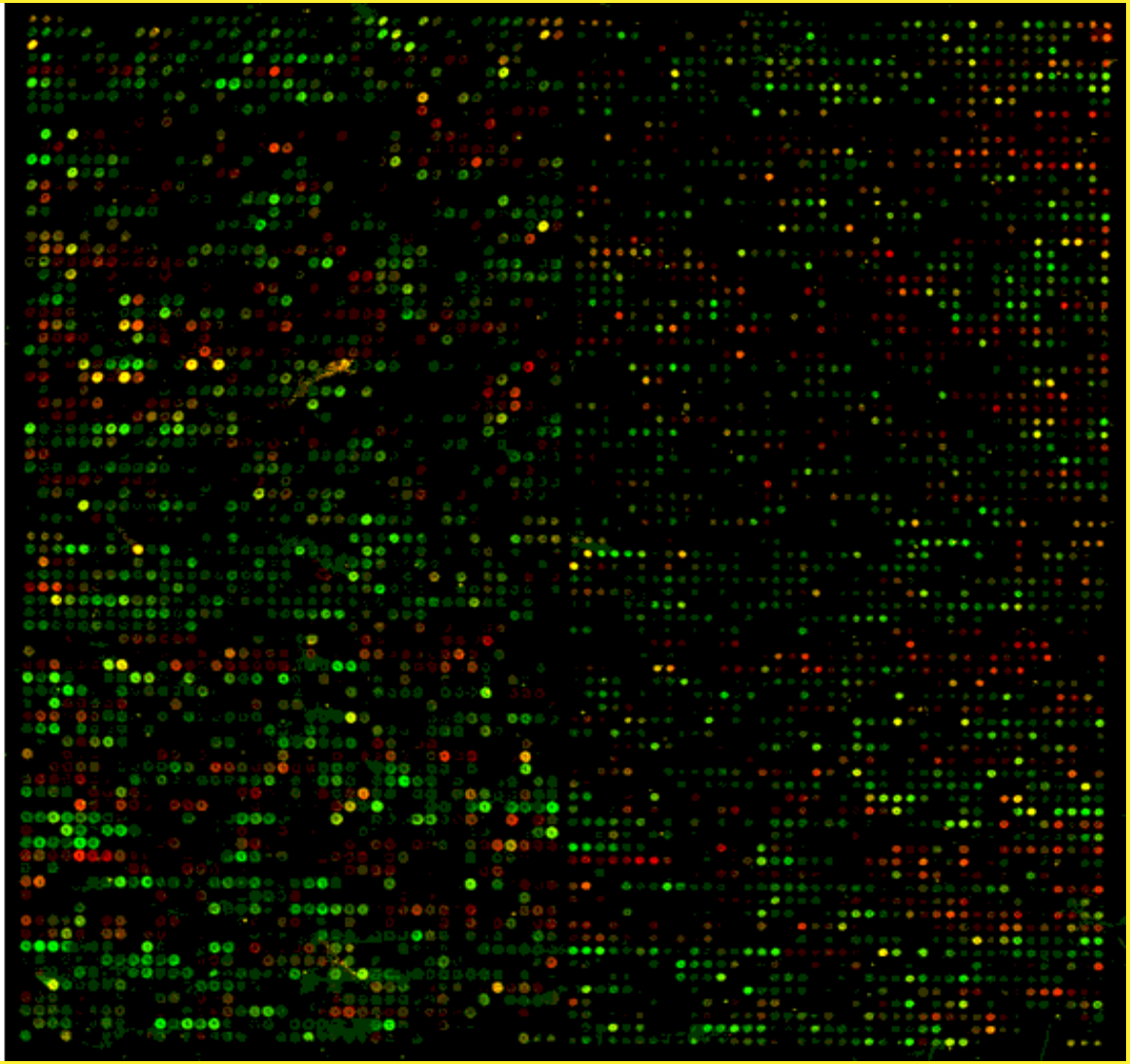
**Identifying protein-protein interactions**

↓

**Networks of interactions**

# 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**