

Bioinformatics: from data to biological knowledge

Dena Leshkowitz, Ph.D
Bioinformatics Unit
Hebrew University

Agenda

- Introduction to Bioinformatics
- Some Biological terms
- Genome and Genes
- Gene Expression Profiling
- Exploring gene function
- Databases

Traditional Biology

- Small team working on a specialized topic
- Well defined experiment to answer precise work

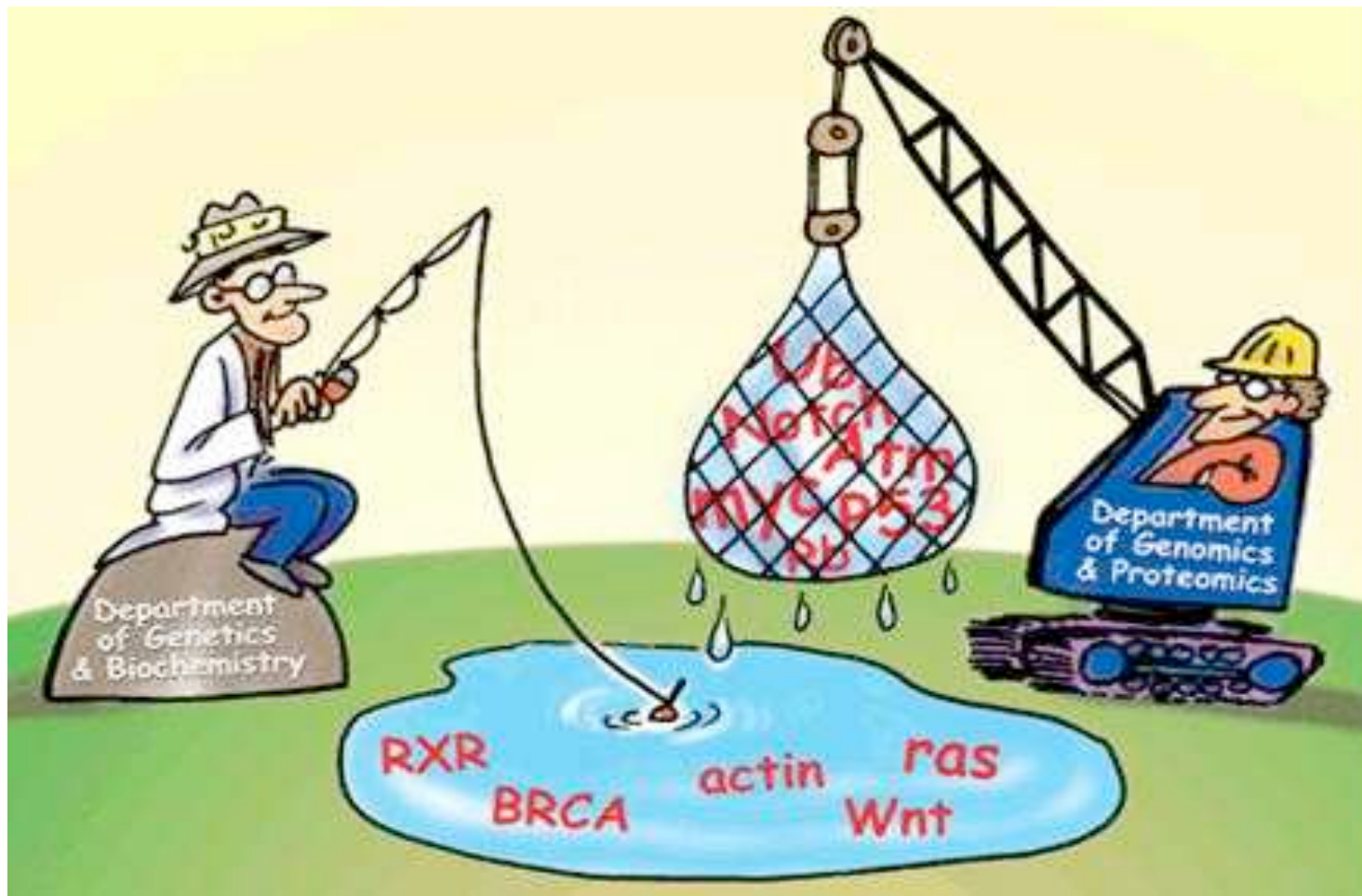


New High-Throughput Biology

Large international teams using cutting edge technology - automation



Traditional Biology Versus High-Throughput



High Throughput Biology

DNA Sequencing
(Automation 1996)

```
...ACGTGACTGAGGACCGTG  
CGACTGAGACTGACTGGGT  
CTAGCTAGACTACGTTTTA  
TATATATATACGTCGTCGT  
ACTGATGACTAGATTACAG  
ACTGATTTAGATACCTGAC  
TGATTTTAAAAAATATT...
```

DNA Microarrays =
DNA Chips (1995)



Each producing hundred giga bytes of information

Biology is becoming an information science

What Do We Do With All This Data?

- Design data structures to represent this information
- Develop databases and software to store the data, enter new data and query the data
- Develop fast applications to analyze the data (data mining)

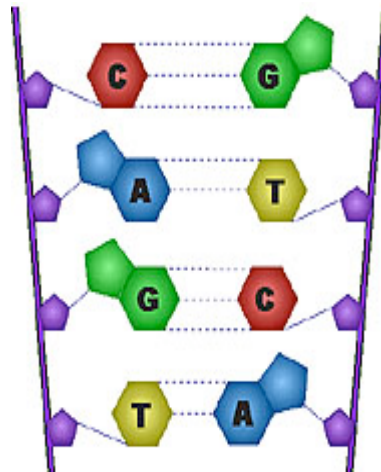
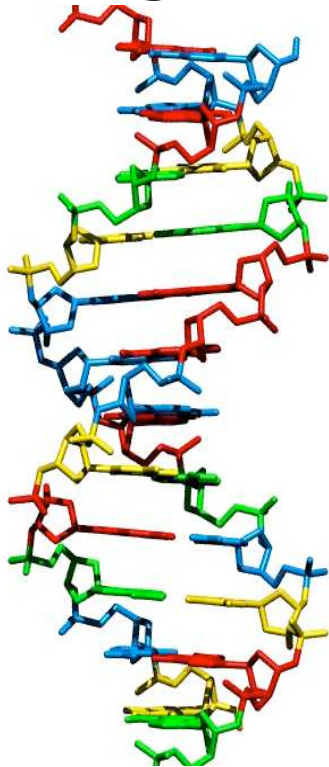
Bioinformatics Definition

- The application of computer technology and software applications, to gather, store, analyze and integrate biological information

Lincoln Stein: "Biologists using computers, or the other way around."

DNA

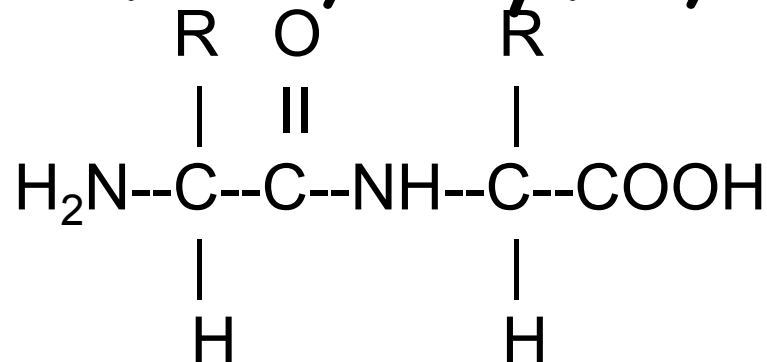
Two strands of linked nucleotides with one of the four bases adenine (A), thymine (T), guanine (G) and cytosine (C).



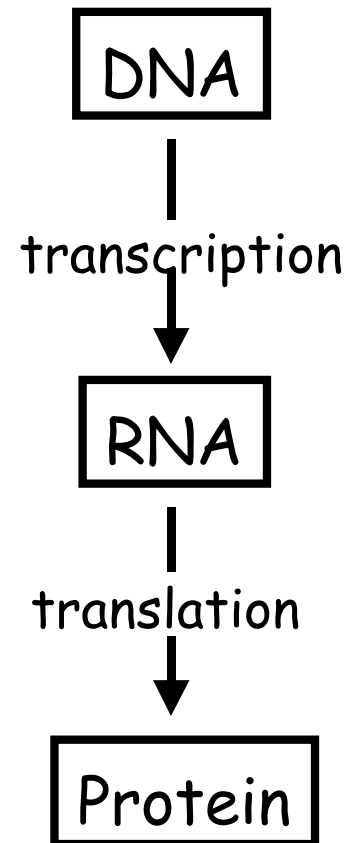
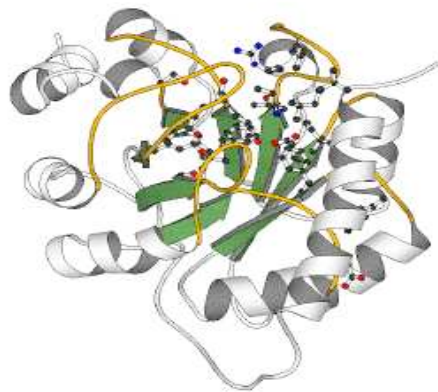
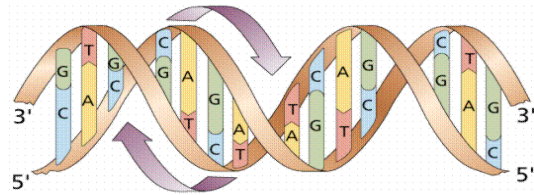
```
...ACGTGACTGAGGACCGTG  
CGACTGAGACTGACTGGGT  
CTAGCTAGACTACGTTTTA  
TATATATATACGTCGTCGT  
ACTGATGACTAGATTACAG  
ACTGATTTAGATACCTGAC  
TGATTTTAAAAAATATT...
```

Protein

- Polypeptide: made up of a sequence of amino acids.
- Folds to form a complex 3-D structure.
- Has unique functions in the cell. Examples are hormones, enzymes, and antibodies.



Gene Expression- Central Dogma



CCTGAGCCAACTATTGATGAA

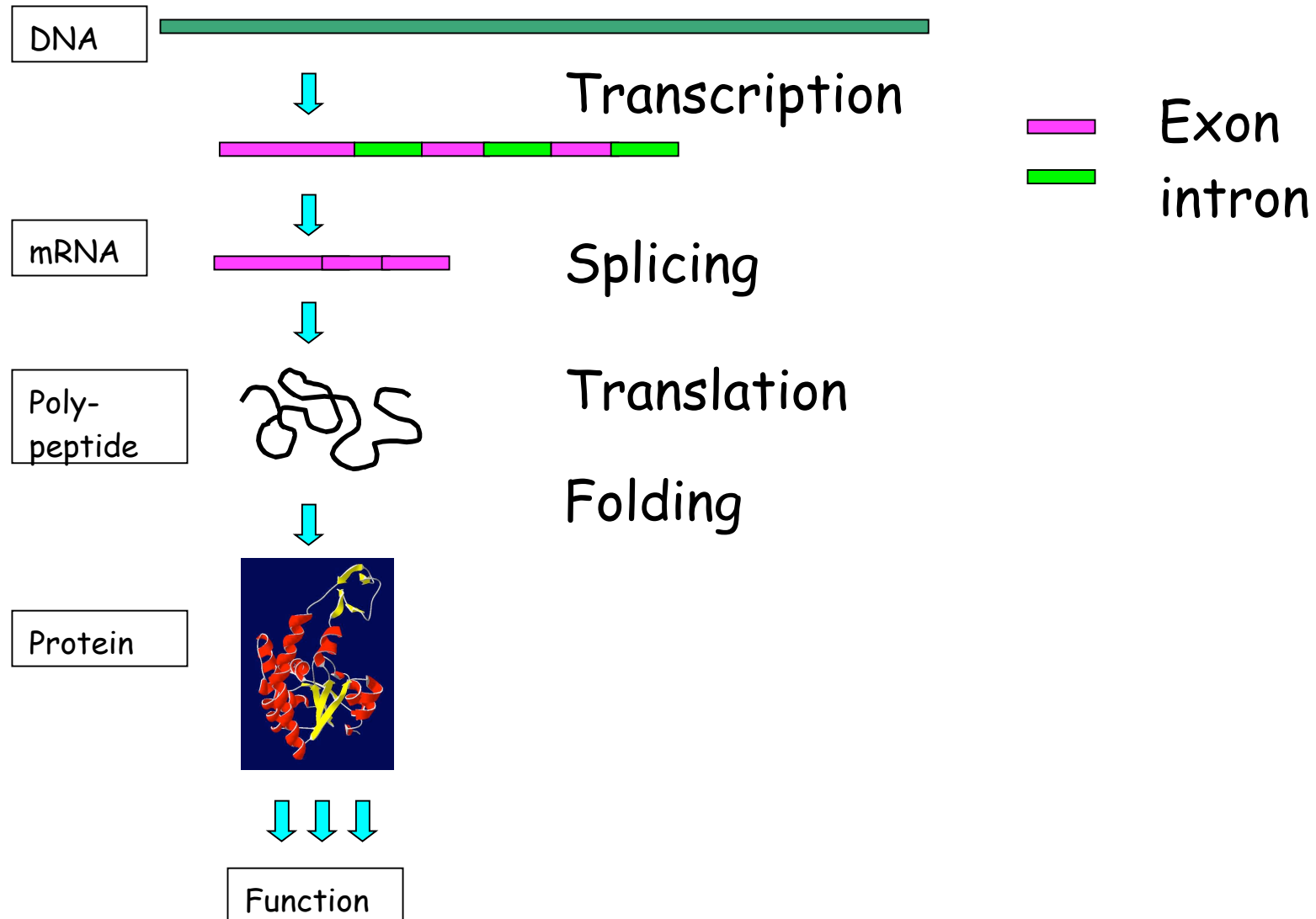


CCUGAGCCAAACUAUUGAUGAA



PEPTIDE

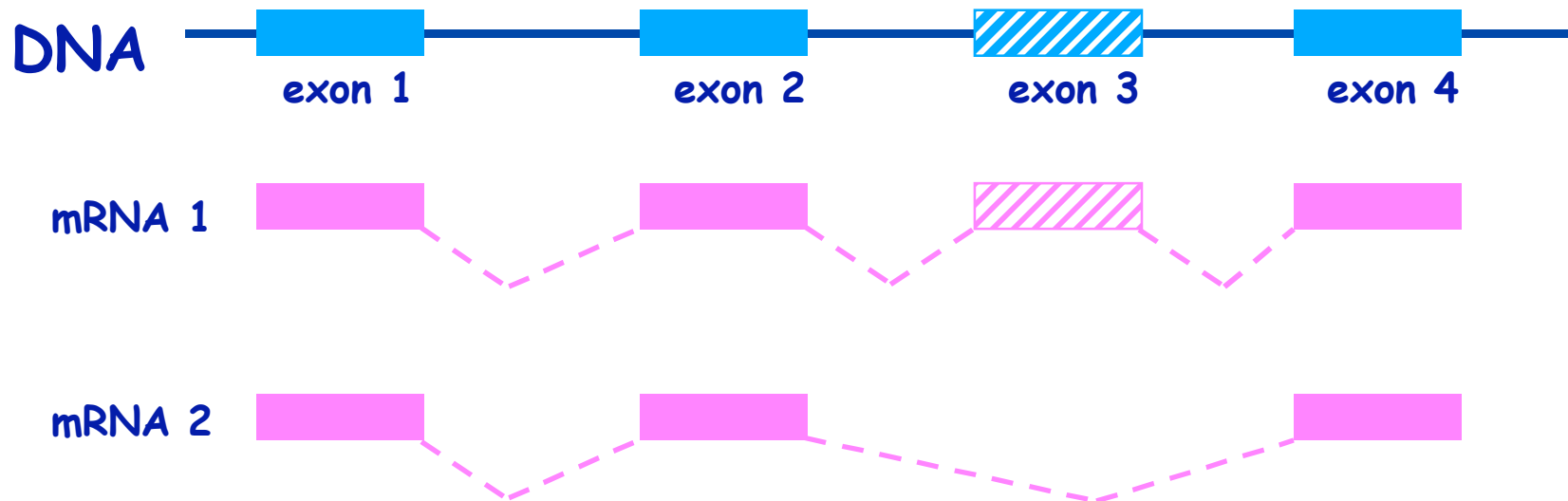
Gene Expression- Splicing



One Gene - Many Products...

Alternative Splicing

~60% of the human genes



The Era of the 'omes

Common lexicon defining cell populations

- **Genome, 1932:**

entire DNA sequence of an organism

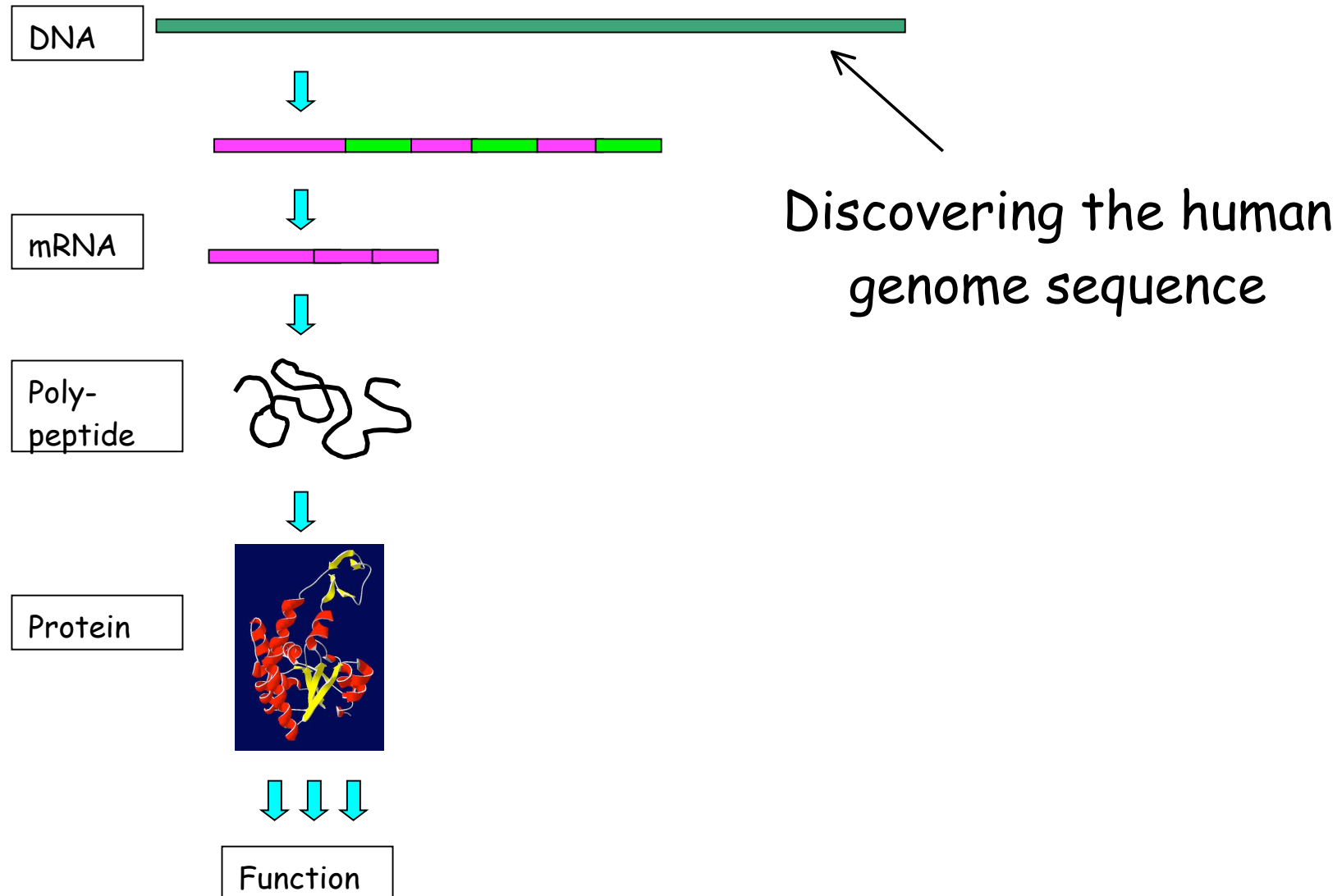
- **Proteome, 1995:**

entire set of proteins

- **Transcriptome, 1997:**

entire set of mRNA

Why do we need Bioinformatics?



2000

The Human Genome Project Complete Assembly



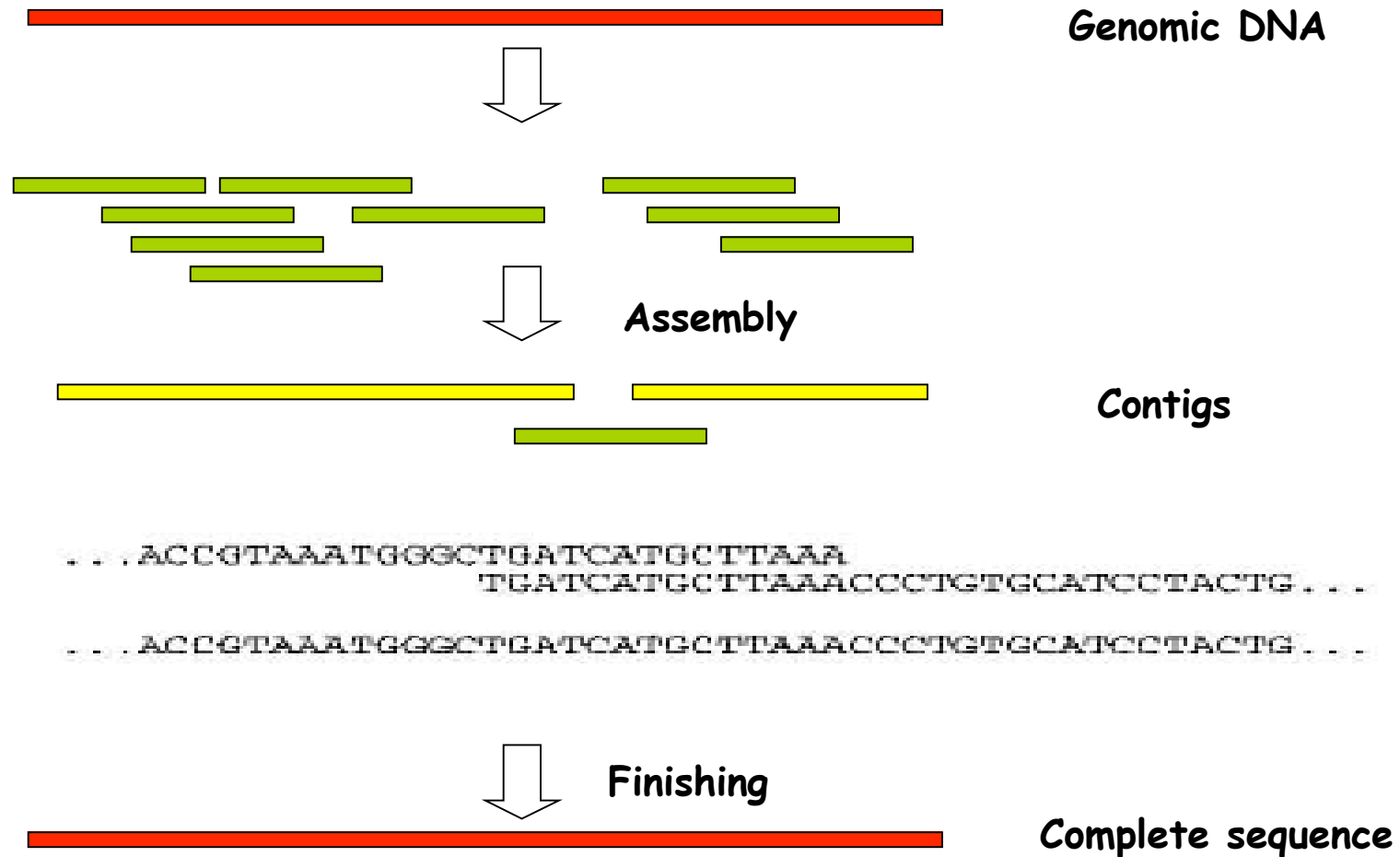
http://www.genomenewsnetwork.org/timeline/2000_human.shtml

The Human Genome Project



The Bioinformatic Challenge -
assembling all the pieces

Assembling the Human Genome (3,200,000,000 bases)



The Human Genome Project

- The new challenges

- Predicting the genes
- Discovering variation between individuals

- The end of the beginning

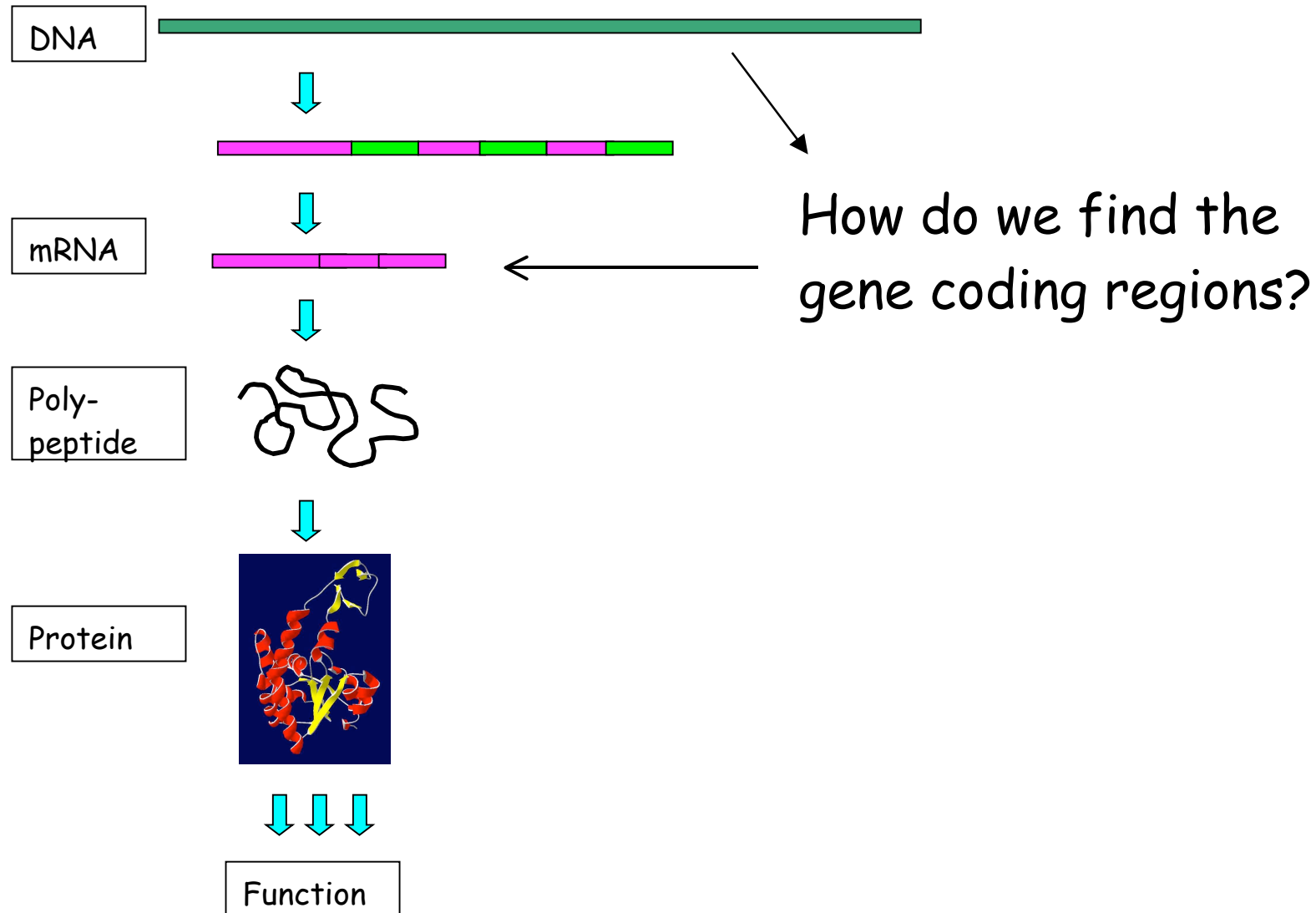
"The true payoff from the HGP will be the ability to better diagnose, treat and prevent disease, and most of those benefits to humanity still lie ahead"

(Francis Collins, *Genome Res*, May 2001)

Genomes Statistics

- Number of completed genomes
 - 1000 viruses
 - 100 Bacteria
 - ~50 eukaryotic species
- In comparison: "the sea may support as many as 2 million different bacteria" (NCBI- coffee break)

Why do we need Bioinformatics?



Where are the genes?

[illegible]

> HSKIIIBE, Human gene for casein kinase II subunit beta (EC 2.7.1.37).
ggggctgagatgtaaattagaggagctggagaggagtgccttcagagtttgggttgctttaagaaagggg
gggtccgaattctcccggtgggtggagggccgaatgtgggaggagaccagaggcagggaagga
gaacttgagcttactgacactgttcttttctagctgacgtgaatgagctcagaggagggtgtc
ctggatttccctgggtctgtgggctccgtggcaatgaattcttctgagttctcttcaacctcc
ctacttgccagcttcacatatcttcccaccagacgttcttccacatatccacttctacactgttctct
aaagcttttatgggagagagtgtagggtgaactaggagagacacaagtaacttctgctgagttgggagt
agaaacaagcacaacagatgcagttgtgttgatgataaggcatcacttagagcattttgcccagggtcaa
agatgaggattttgatatgggttccctcttgggttccatgtcctgacaggtggatgaagactacatcca
ggacaaatttaactcttactggactcaatgagcaggtccctcactatcgacaagctctagacatgatctt
ggacctggagcctgggtgaggcacccctcagggttggtttgtgtgtgtgcgtgcactattttctcttcaa
atctctattcacttgccctgaattttgccaaatttcccttgggtctcttgatttctttaacccccaaattca
tgctttattttgatccctccacctgactcttgtctagttttgtgacgtatatcacttgttctcatgtttt
tgcaagggtcagaagcccagggttcttgggtcccatgcccagatgttggatggggtaaggcccaaaagta
gggtgctaggcaaaactgaatagcccgccagccctggatatgggcaggggcacctaggaaagctgaaaaaca
agtagttgcatttggccgggctgtgggttcagatgaagaactggaagacaaccccaaccagagtgacctg
attgagcaggccagccgagatgctttatggattgatccacgcccgtacatccttaccacccgtggcatc
gcccagatgggtgaggcctctctgctcctacctgcctcctcttgagcagtaagagacacagggttccctgca
gcaagaagtcatgtttaagccctgtttaaggaagctagctgagaagaggggaagaaccccagaacttgg
ccctgccctaatttggaagaaaggcaacacagaagtttgagagcccattctagtcacagagaagggggcct
ctggacagagtgggaaggagtgcgcacagagtgggtatgggttgggctgcgaaggagggttgccctctct
ttacatctacctgccaaaccccttccattgtattcacctcagttggaaaagtaccagcaaggagactttg
gttactgtcctcgtgtgtactgtgagaaccagccaatgcttccattGgtgagtggtgaagaaggga
ggaaagcacctgtgtggcagttcttatgggaaggagtggggctcaacacattggagcctgagtcctgagg
ggaggttaggttaggaatagggggatacctggcctgctgagttctggctgtctcccaggccttccagacat
cccagggtgaagccatgggtgaagctctactgccccaaagtgcattggatgtgtacacaccaagtcatcaag
acaccatcacacggatggcgccctacttcggcactgggttccctcacatgctcttcatgggtgcattccga
gtacgggcccaagagaaactgccaaaccagtttgtgccaggtaggagagcaggagaggtcattaagggtca
aaggaaaggcccaagatccccagagagggggaggacagggcattggcccttcttgagggtctgcttctcc
cagaatcaggggcatctccctgctgagtgactgtgggaaagttatttgattatctgtgcttgagttacct
tattgtagaatgttcttgagctgagaagtgggaaccaacagaggctttagctctgagcaggtccatagag
gagctcagggtggggagggtgggaatgcagggtgactggcagggcctggatggggctcatgctgctgcctct
ctgacctctgccctggcctaggctctacggtttcaagatgagggcctaccagctgcagctcca
agccgccagcaacttcaagagcccagtcacagacgatccgctccccccacctgtcctgcagtc
tttgtcttttcccttcttttttgccacctttcaggaatgggttttagtttaaattaaagga
gtcgttatcgtgggtgggaatatgaaataaagtagaagaaaaggccatgagctagttctgctgggtgcttgc
ggaaaggggggtggagcgtggccatggaaatcgggctccacggcccagggtatgg

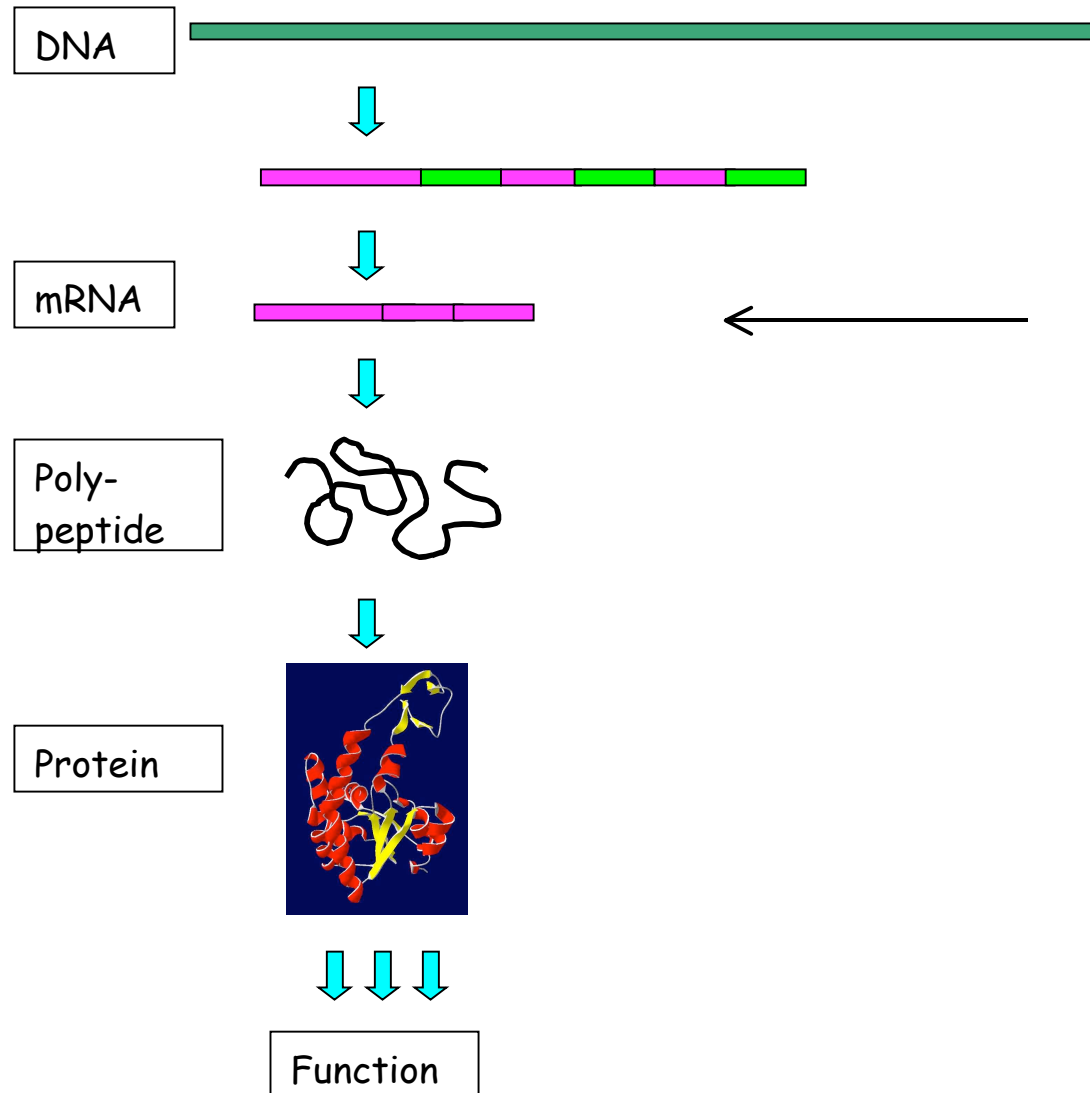
Human Gene Prediction Challenges

- Genes can be difficult to identify
 - Alternative splicing
 - 1-2 % coding sequence
 - Pseudogenes
- How many genes are there in the human genome?
 - 30 to 38 thousand
 - Most are gene predicted
 - There is a high rate of problems in the predictions.
- An information problem that is still not solved

Genomes & Genes numbers

Species	Genome Size	Number of genes
<i>Bacteriophage lambda</i> (virus)	5×10^4	67
<i>Caenorhabditis elegans</i> (worm)	1×10^8	~18,000
<i>Homo sapiens</i> (human)	3×10^9	~30,000

Why do we need Bioinformatics?

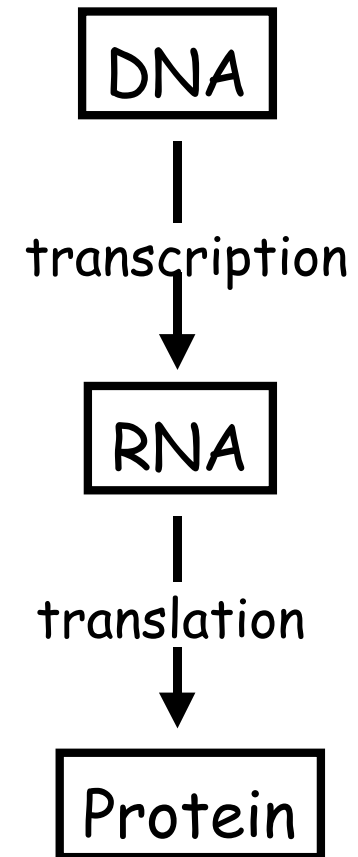


What is the amount of the various mRNAs in a cell ?

Gene Expression Varies

- The cells in our body have the same genome
- The genes expressed in a cell vary due to:
 - Cell type
 - External conditions

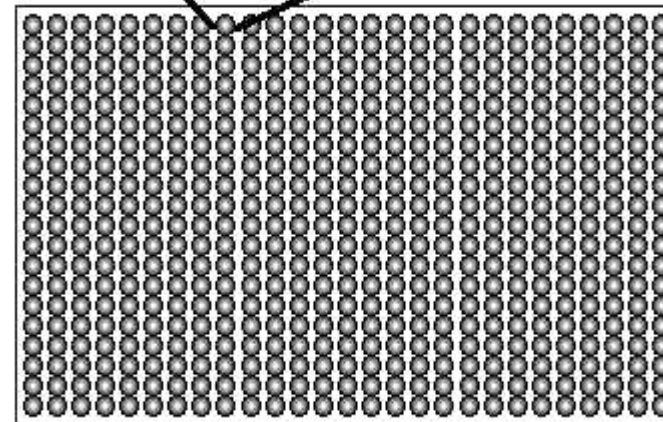
Measuring Gene Expression:
Microarrays



DNA Microarrays

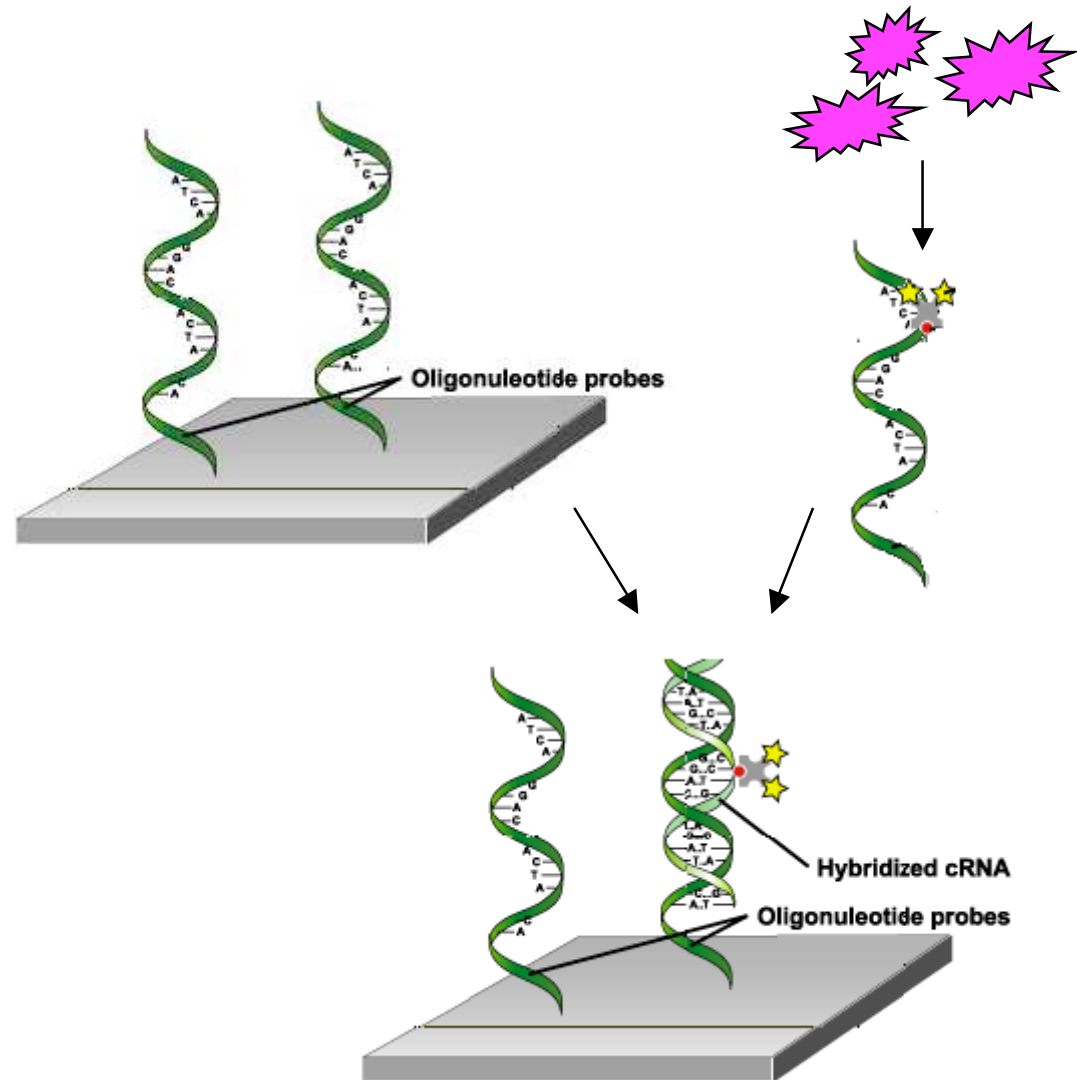
- A 2D array of DNA sequences from thousands of genes
- Each spot has many copies of same gene

TCCTTTCGG	AACGGTTGGC	GTCTGCGCAC	GGCGGTGTGG	GGCATGACAT
GCCGCCCCAG	GAACAACCCC	GACACGGCTT	TAAGCCTCTC	AAATCGCTGT
AGACATCATC	TTTACGTGCT	TGGCTTGCCC	TGCCACCATT	AGGGCTGPTC
CCGCGACGAC	TCGCCATTCA	ACCTCAGTCC	TTCCGGTTGA	GCGAGTGGGT
CGCGCGCAAG	GTGCGAATGG	GTGCGCGCGA	AAGTGTGCG	CTGGCTGTAT
TATATGCTGC	CTATAGCGAG	ACTAACGACC	CACACTTTCA	CACAAGGATT
TCCCGCTAAT	GGGTACCTCG	CGTCAGGACC	TTGACGCAAG	CGCGCCTTCG
GTTGGCCCCA	AGCTTGCTAG	GACTACTTAT	CTTGAGCTCA	TTTAACATCC
CGGCGCCTCT	CCGGGAGCGG	TCGTGCGGAA	GAAGTCAAAC	CCGGAACGGC
GTTGACAAAG	CGTGGAGACA	TCGATACCTC	TGTGTCAGCG	GCCACAAATC



Performing a DNA Microarray Experiment

- Extract mRNA from a sample
- Label mRNA
- Allow labeled mRNA to react with DNA on the microarray
- Measure amount of label per spot



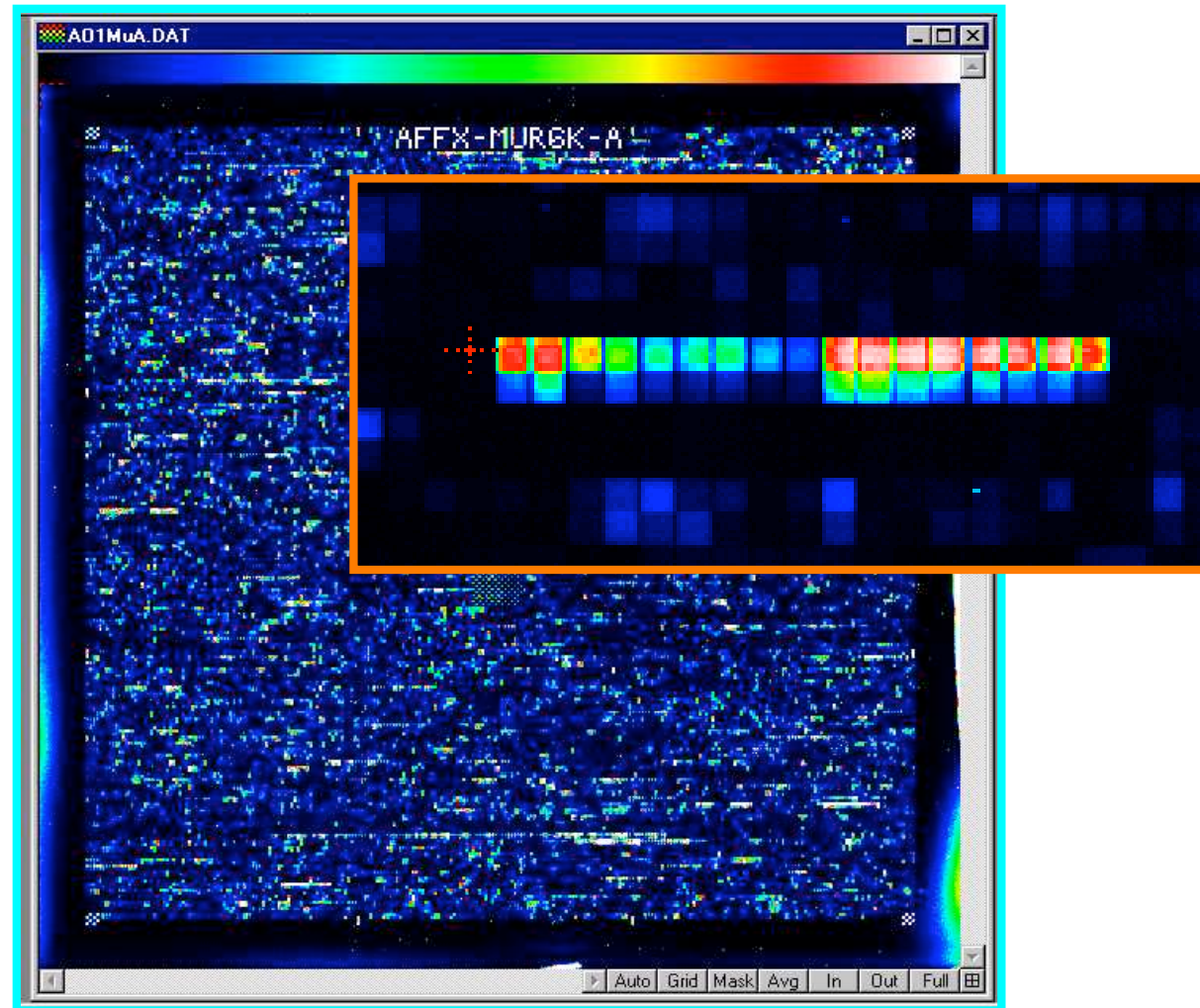
DNA Microarray Image

Affymetrix

Genechip

- 500,000 probes on 1.3cm²

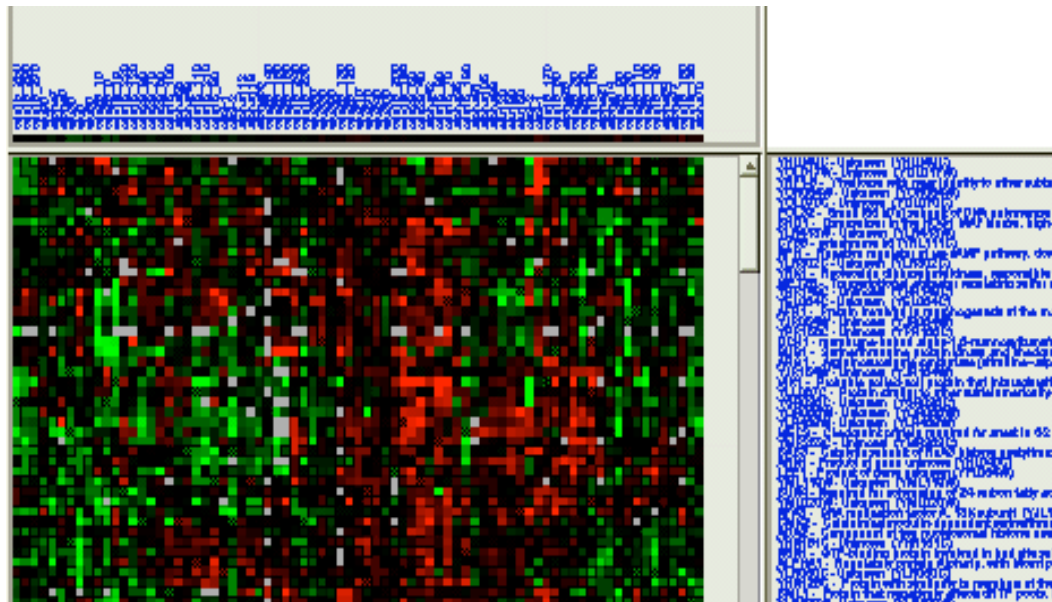
- Two chips for the whole transcriptome



Goal of Microarray Experiments

Measure level of gene expression across many different conditions:

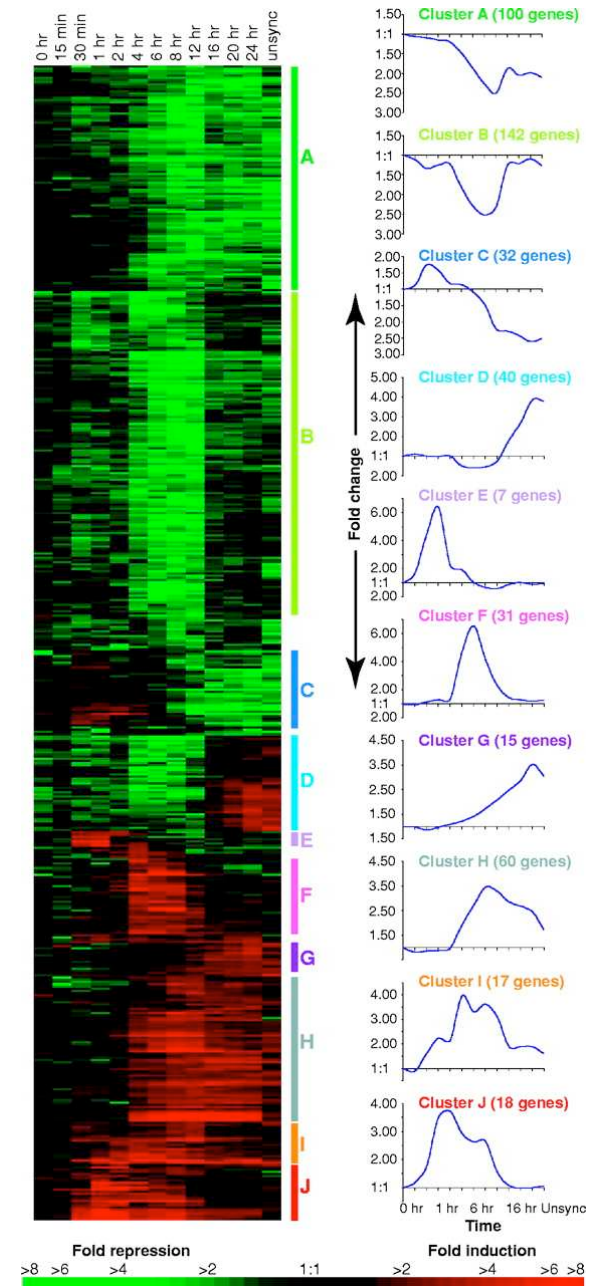
Expression Matrix M : $\{\text{conditions}\} \times \{\text{genes}\}$
conditions



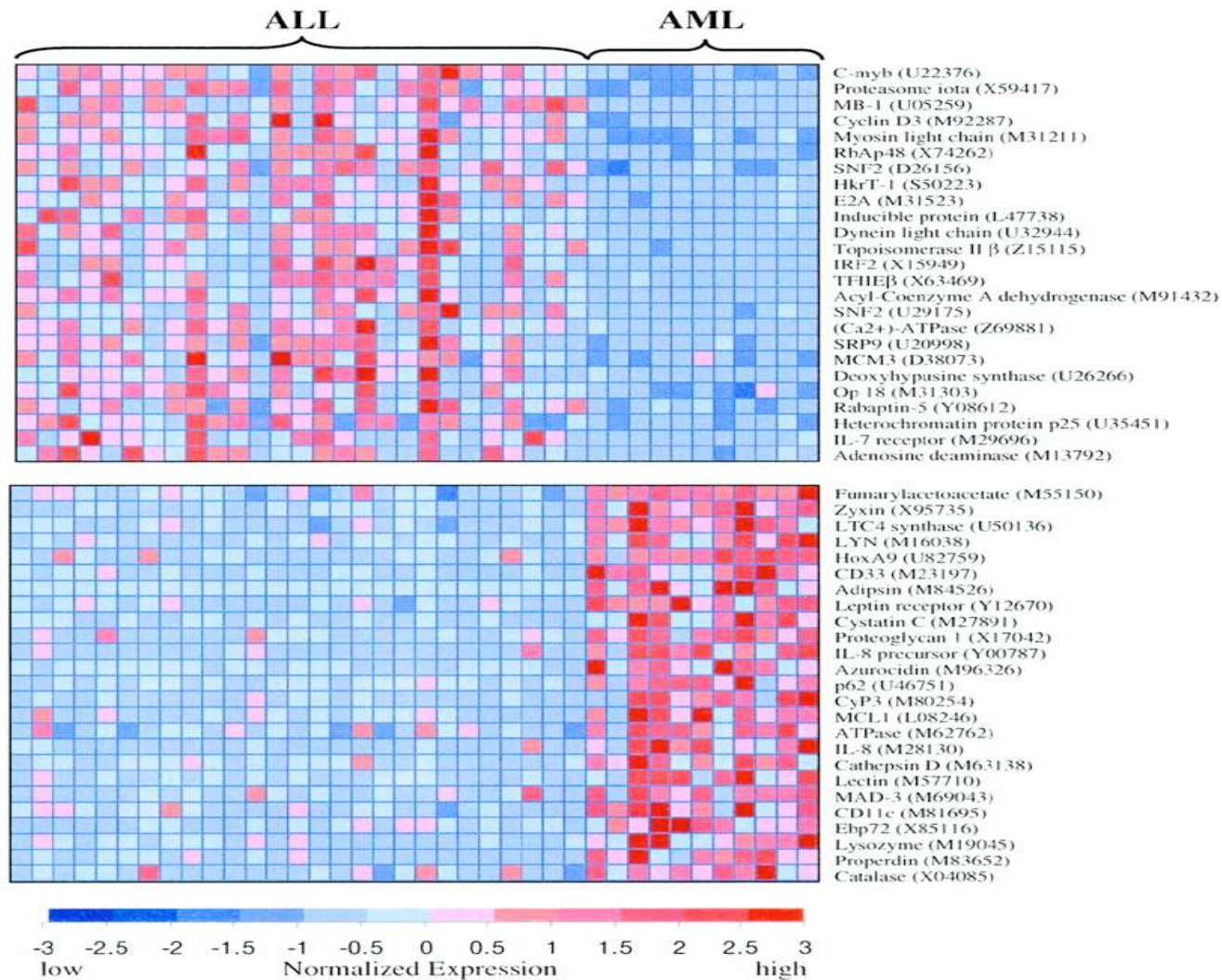
How to make sense of the data?

Clustering

- We attempt to look for groups in the data.
What are the common 'patterns' of gene expression in my dataset?
- Grouping can be done on the basis of some notion of 'similarity' of expression

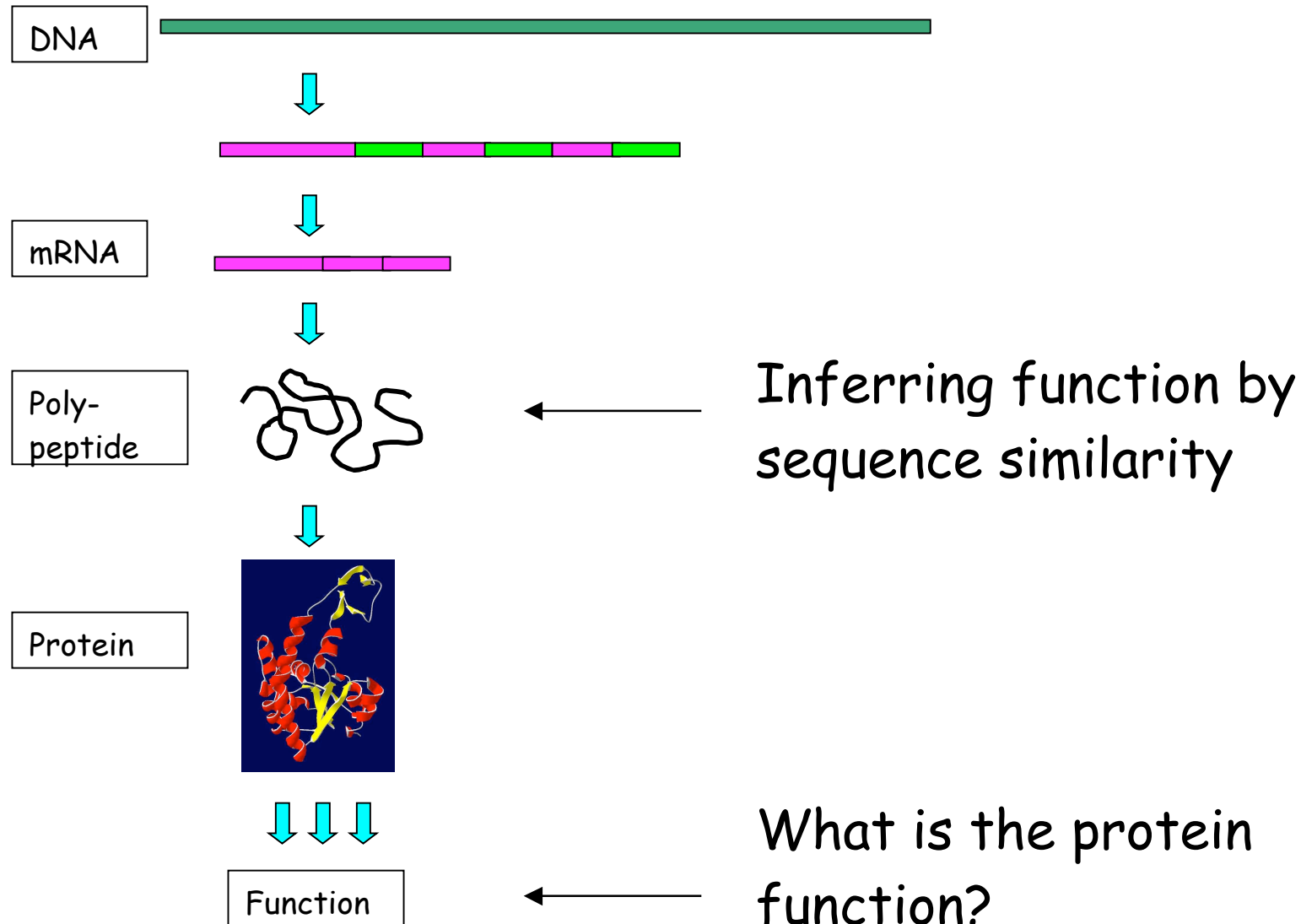


Classification



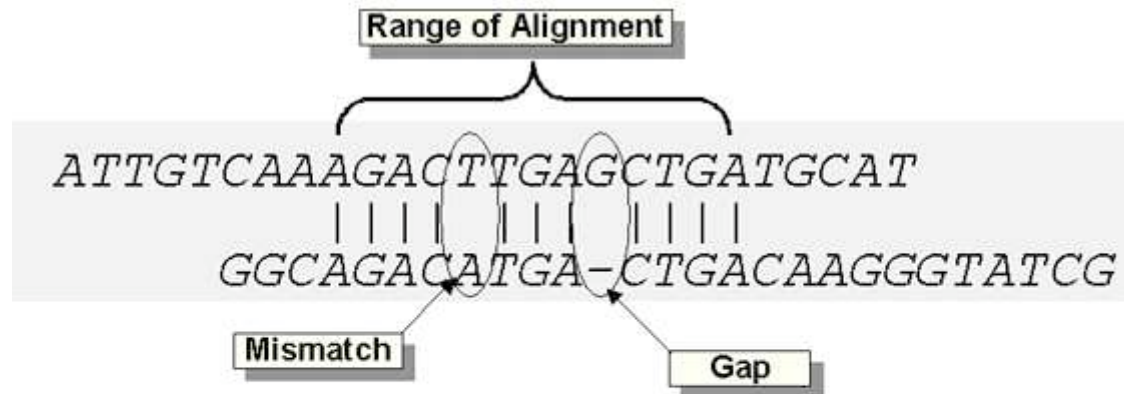
http://www.blc.arizona.edu/courses/181gh/rick/human_genetics/picture5.html

Why do we need Bioinformatics?



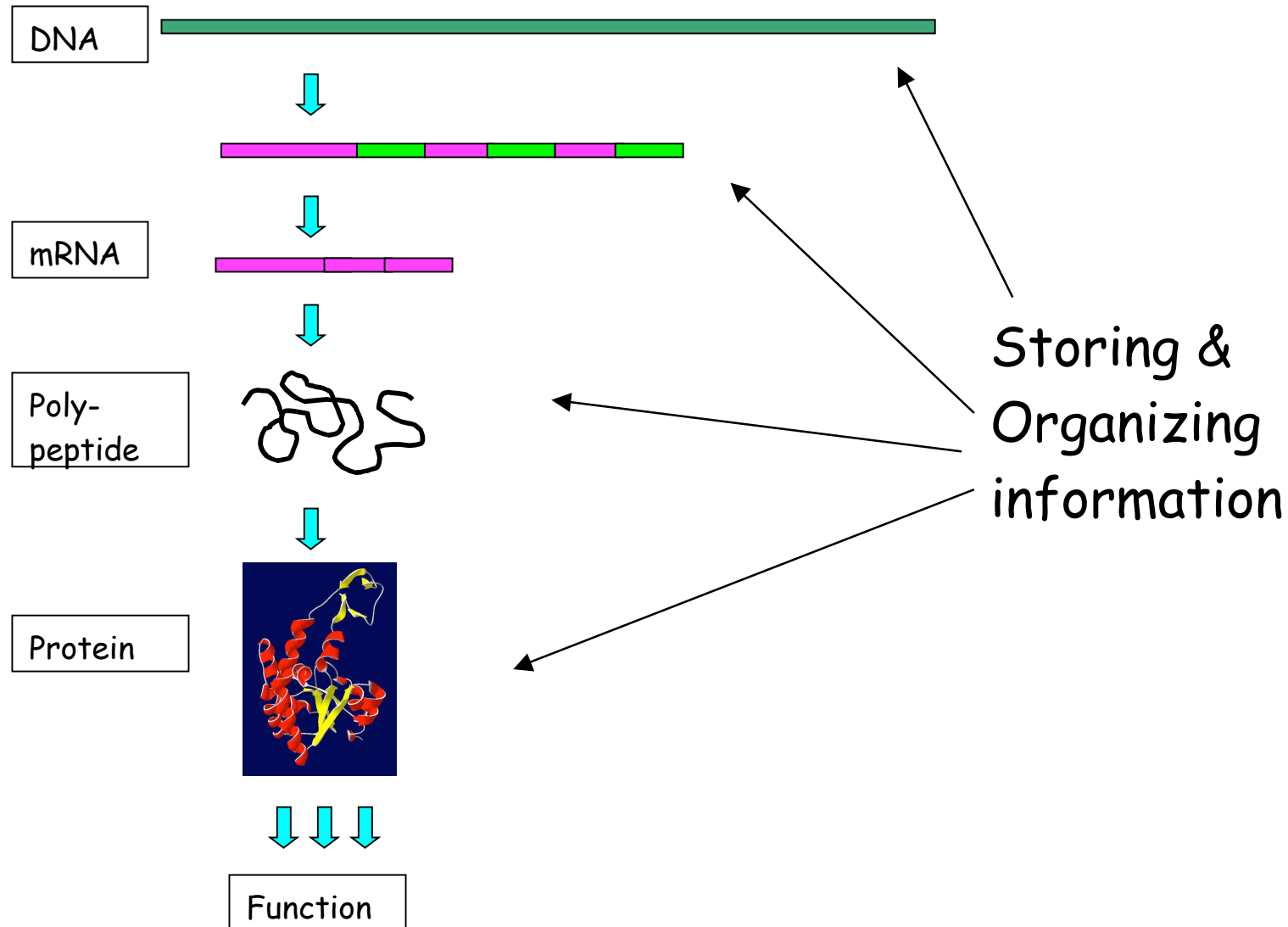
Sequence Alignment

Definition: a process of lining-up 2 or more sequences with maximum conservation



$$S = \sum (\text{identities, mismatches}) - \sum (\text{gap penalties})$$

Why do we need Bioinformatics?

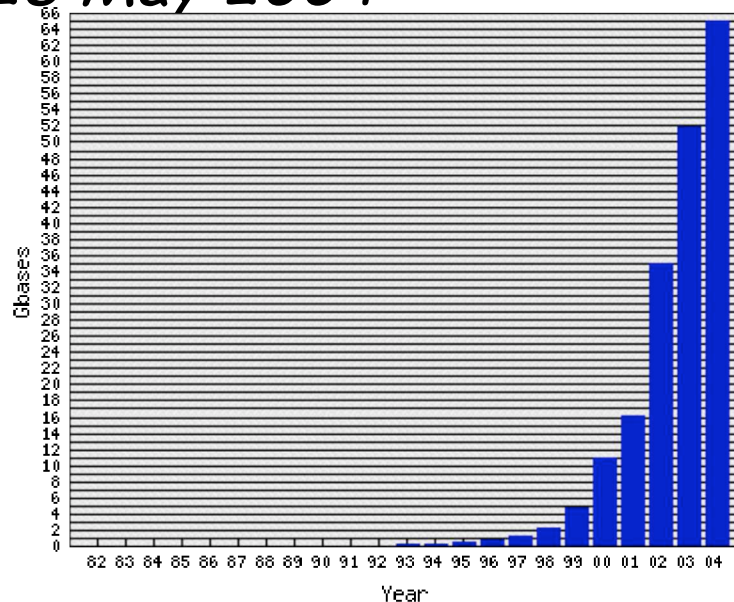


EMBL Database Growth

Total nucleotides 65,592,536,942 bases

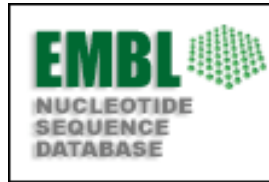
Number of entries 39,342,329

28 May 2004



<http://www3.ebi.ac.uk/Services/DBStats/>

Sequence Databases



- NCBI (USA)
- EMBL (Europe)
- DDBJ (Japan)

The Challenge Today

Integrating omics data to
understand the biological
process (networks)



The End