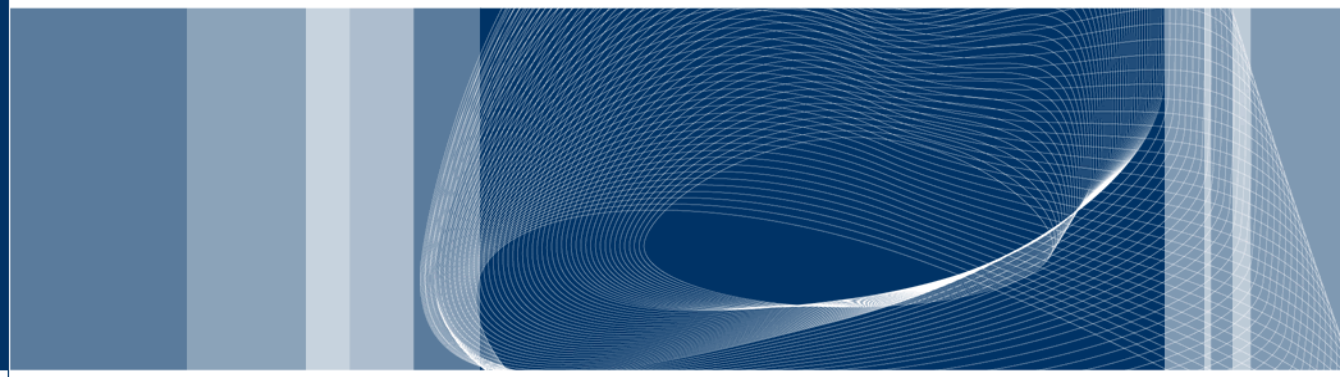
POLITECNICO DI MILANO



Data Mining for Biological Data Analysis

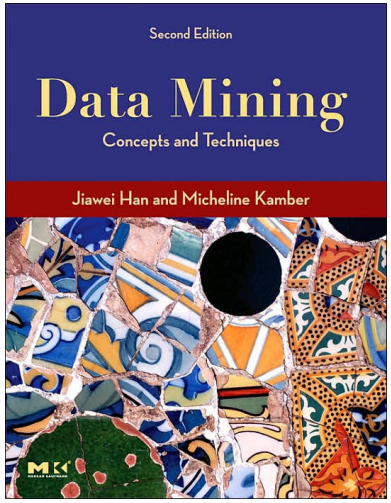
Data Mining and Text Mining (UIC 583 @ Politecnico di Milano)

Daniele Loiacono

References

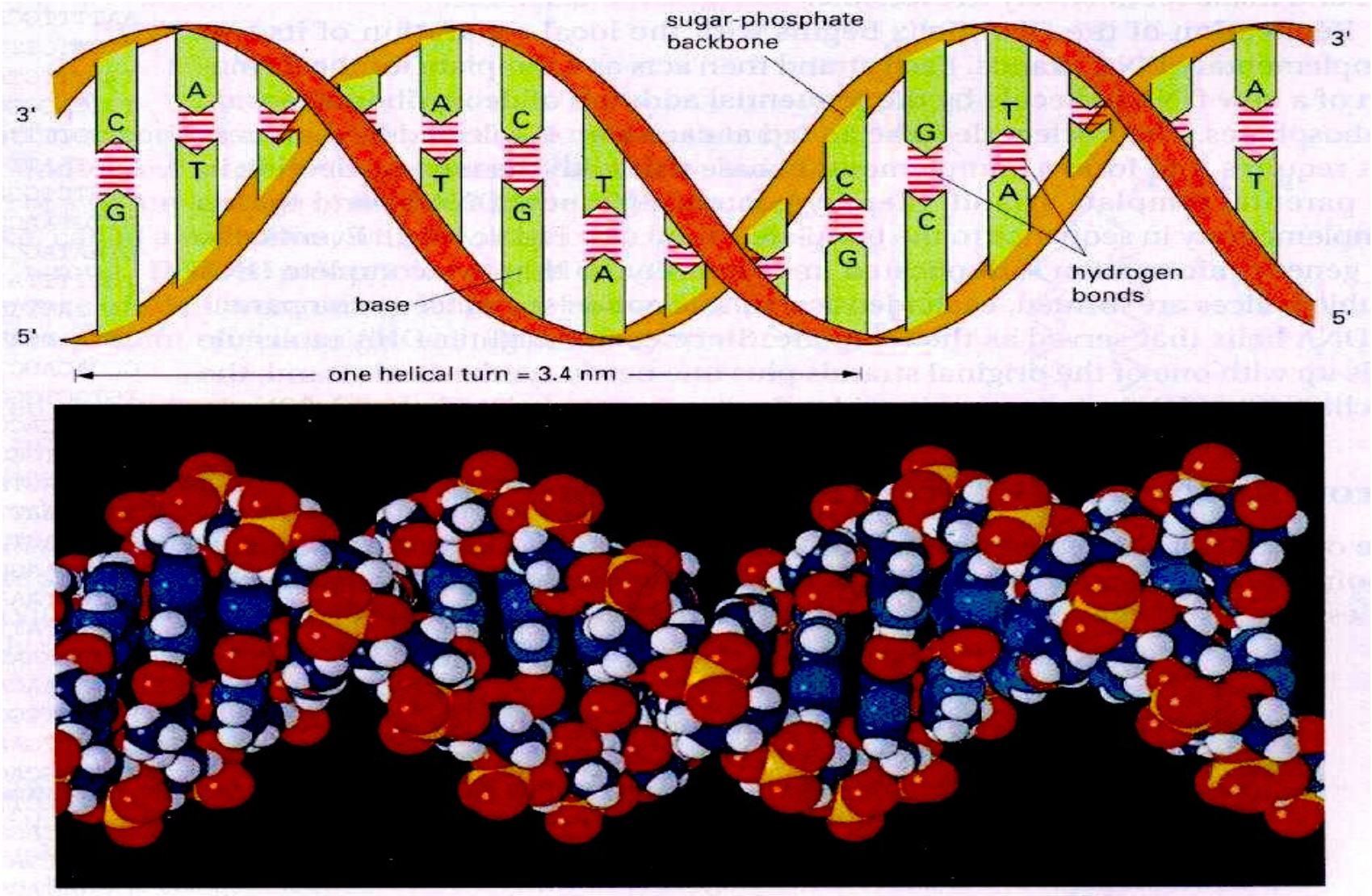
- **Data Mining Course** by *Gregory-Platesky Shapiro* available at www.kdnuggets.com

- Jiawei Han and Micheline Kamber, "Data Mining: Concepts and Techniques", The Morgan Kaufmann Series in Data Management Systems (Second Edition)
  - Chapter 8



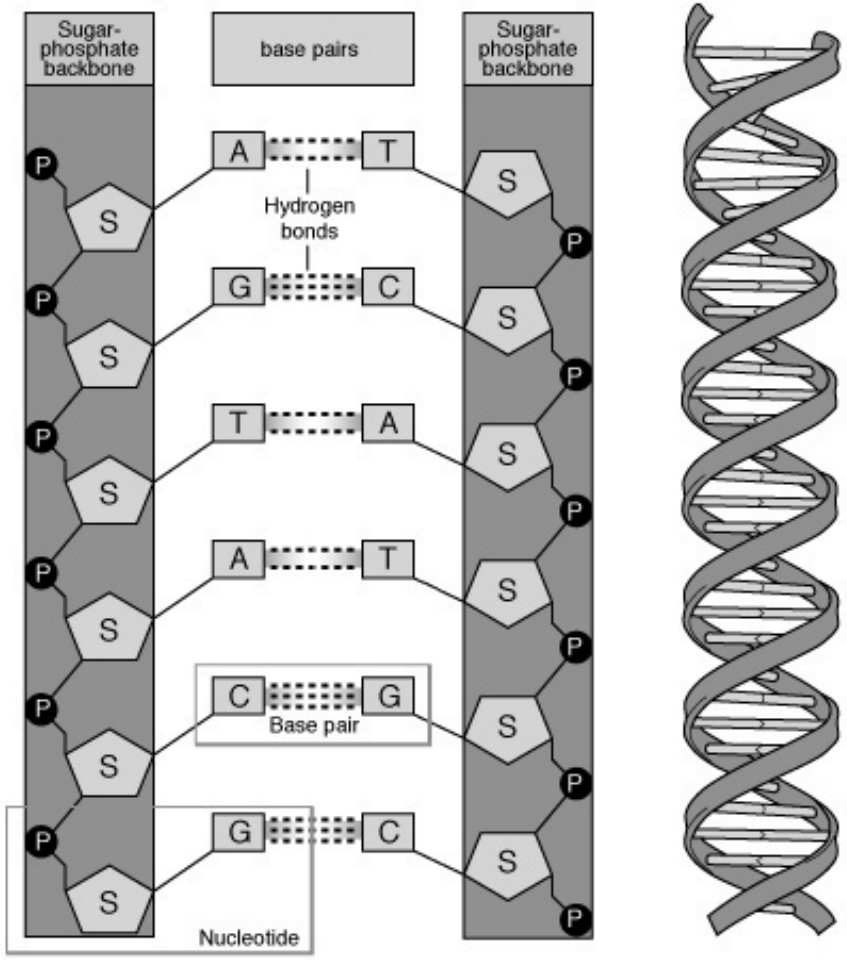
Introduction to Biology

The DNA


sugar-phosphate backbone
3'
5'
base
hydrogen bonds
one helical turn = 3.4 nm

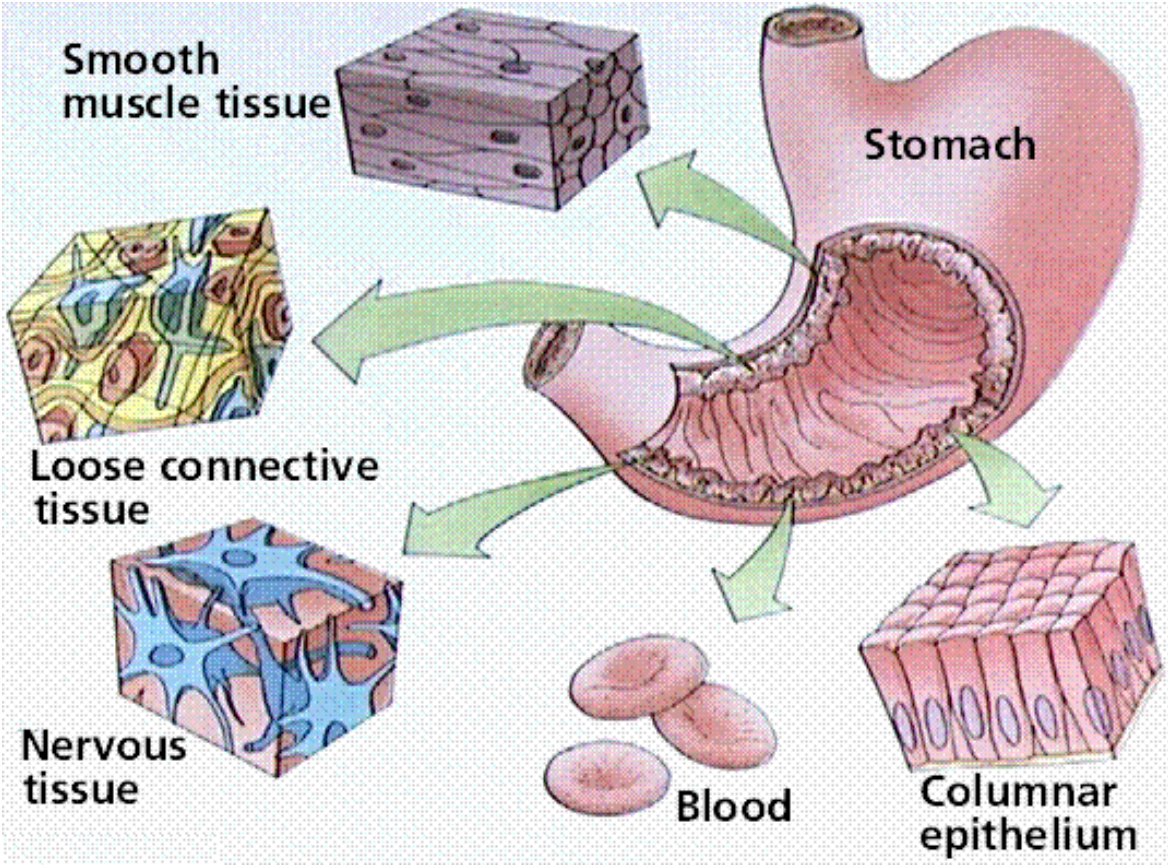
DNA components

- Four nucleotide types:
  - Adenine
  - Guanine
  - Cytosine
  - Thymine
- Hydrogen bonds:
  - A-T
  - C-G



Different cell types

- All cells of an organism contain the same DNA content (and the same genes) yet there is a variety of cell types.

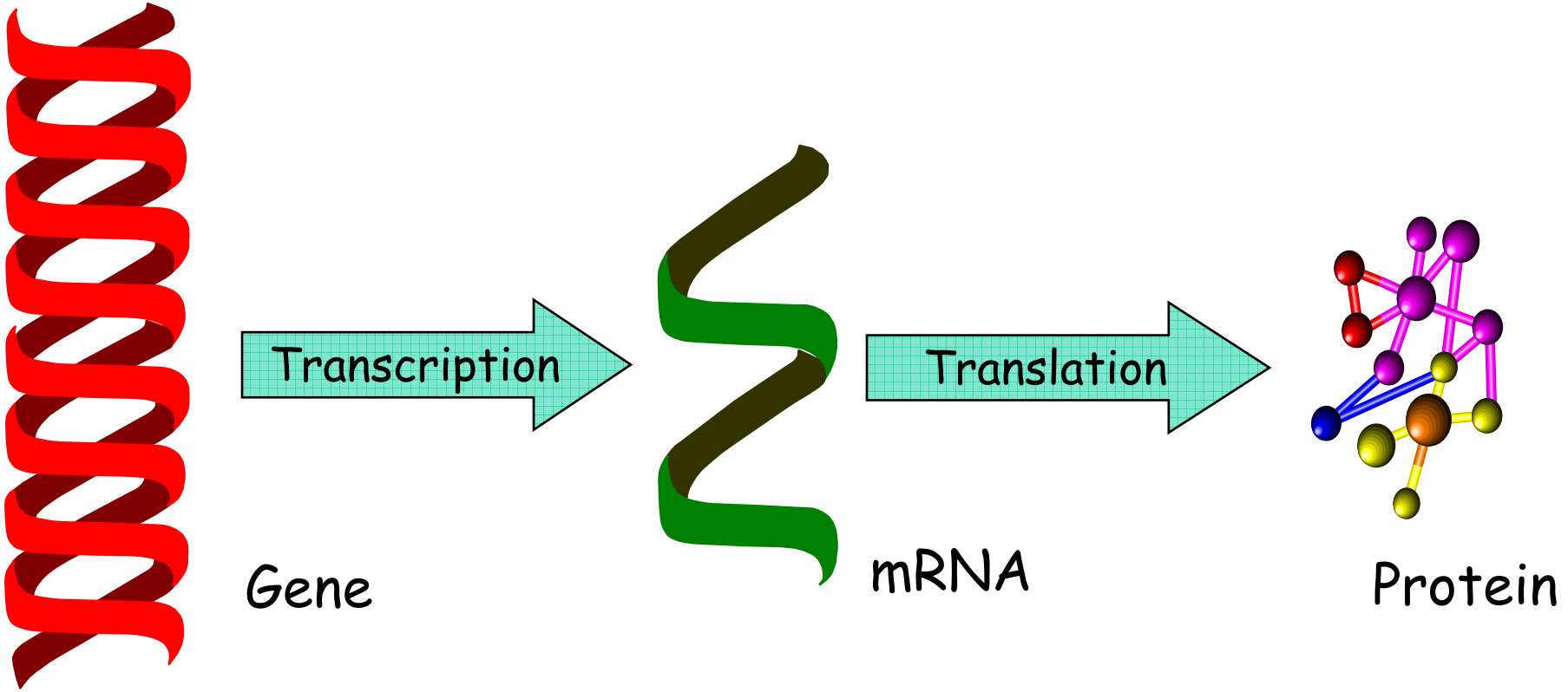


So, how does the cell use DNA ?

The "Central Dogma"

- DNA contains thousands of particular segments called **genes**
- Genes contain "instructions" for making **proteins**
- In order to be executed these "instructions" have to be transcribed into **mRNA** (similar to DNA, with Uracil instead of Thymine)
- Proteins are defined by a sequence of **amino acids** (20 types)
- There are almost one million of proteins that act alone or in complexes to perform many cellular functions

Gene expression



Cells express **different** subset of the genes in different **tissues** and under different **conditions**

Muscle, nervous, blood ...

Disease, mutation...

Genomic and Proteomic

- **Thousands** of genes (~25K in human DNA) function in a **complicated** and **orchestrated** way that creates the mystery of life.
- **Genomic** studies the **functionality** of specific genes, their relations to **diseases**, their **associated proteins** and their participation in **biological processes**
- **Proteins** (~1M in human organism) are responsible for many regulatory functions in cells, tissues and organism
- **Proteome**, the collection of proteins produced, evolves dynamically during time depending on environmental signals.
- **Proteomic** studies the sequences of proteins and their functionalities

Data Mining of Biological Data (1)

- Semantic integration of genomic and proteomic databases
  - Data produced by different labs need to be integrated
  - Data mining can be used to perform data cleaning, integration, object reconciliation to merge heterogeneous databases
- Alignment of nucleotide/protein sequences
  - Build phylogenetic trees
  - Similarity search
  - Difference search
- Protein structure analysis
  - 3D structure of proteins heavily affects their functionalities
  - Prediction of protein structures
  - Discovery of regularities

Data Mining of Biological Data (2)

- Association and path analysis of gene sequences
  - Analysis of gene associations in diseases
  - Discovery of sequential patterns of genes correlated to different stages of diseases
- Visualization
  - Support to knowledge discovery
  - Interactive data explortation

DNA Microarray Analysis

Microarray Data Analysis

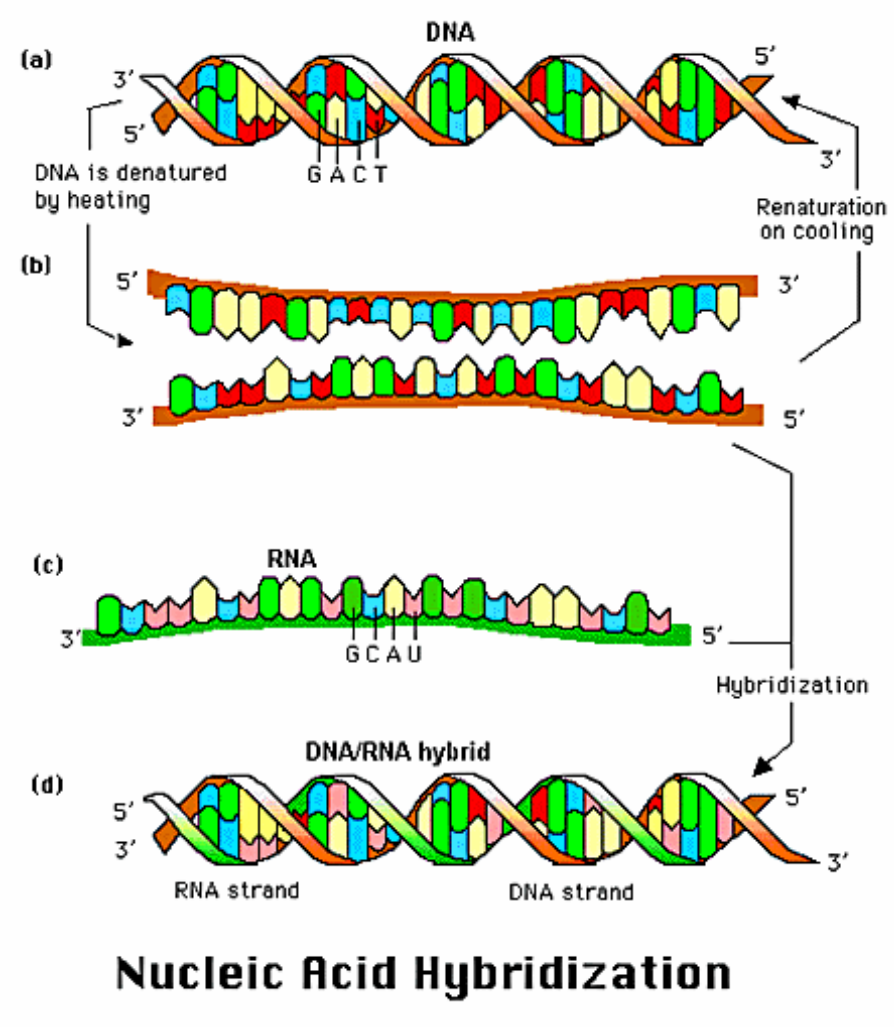
- The DNA Microarray is a technology that allows the analysis of the gene expression levels in samples collected
- Such an analysis has many potential applications
  - Earlier and more accurate diagnostics
  - New molecular targets for therapy
  - Improved and individualized treatments
  - Fundamental biological discovery (e.g. finding and refining biological pathways)
- Examples
  - Molecular diagnosis of leukemia, breast cancer, ...
  - Discovery that genetic signature strongly predicts outcome
  - A few new drugs, many new promising drug targets

Motivation for DNA Microarrays

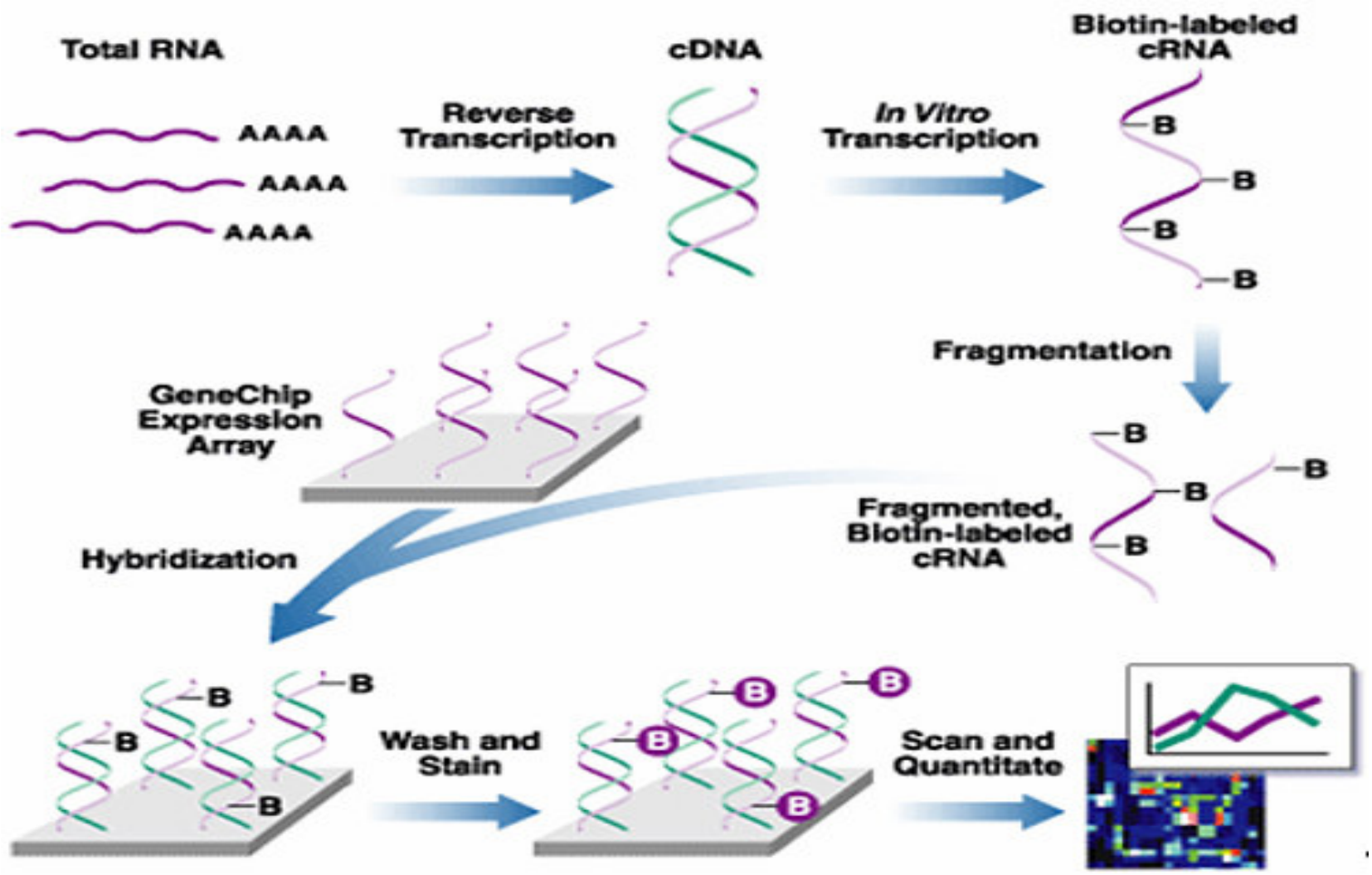
- **Traditional methods** in molecular biology generally work on a **”one gene in one experiment”** basis, which means that the throughput is very limited and the **”whole picture”** of genes function is hard to obtain
- “In early 1997, scientists never envisioned looking at more than 25 to 50 gene-expression levels simultaneously. Today everybody tells us that they want to look at the whole genome.” – Kreiner, Affymetrics
- With a technology for **simultaneously** analyzing the expression levels of **large numbers** of genes we can:
  - Study the behavior of **co-regulated gene networks**.
  - Look for groups of **genes involved in a particular biological process** or in a specific disease by identifying genes whose expression levels change under certain circumstances.
  - Detecting changes in gene expression level in order to have clues on its **product function**.
  - Compare **normal** organism and **mutant** RNA transcription profiles.

The technology: hybridization

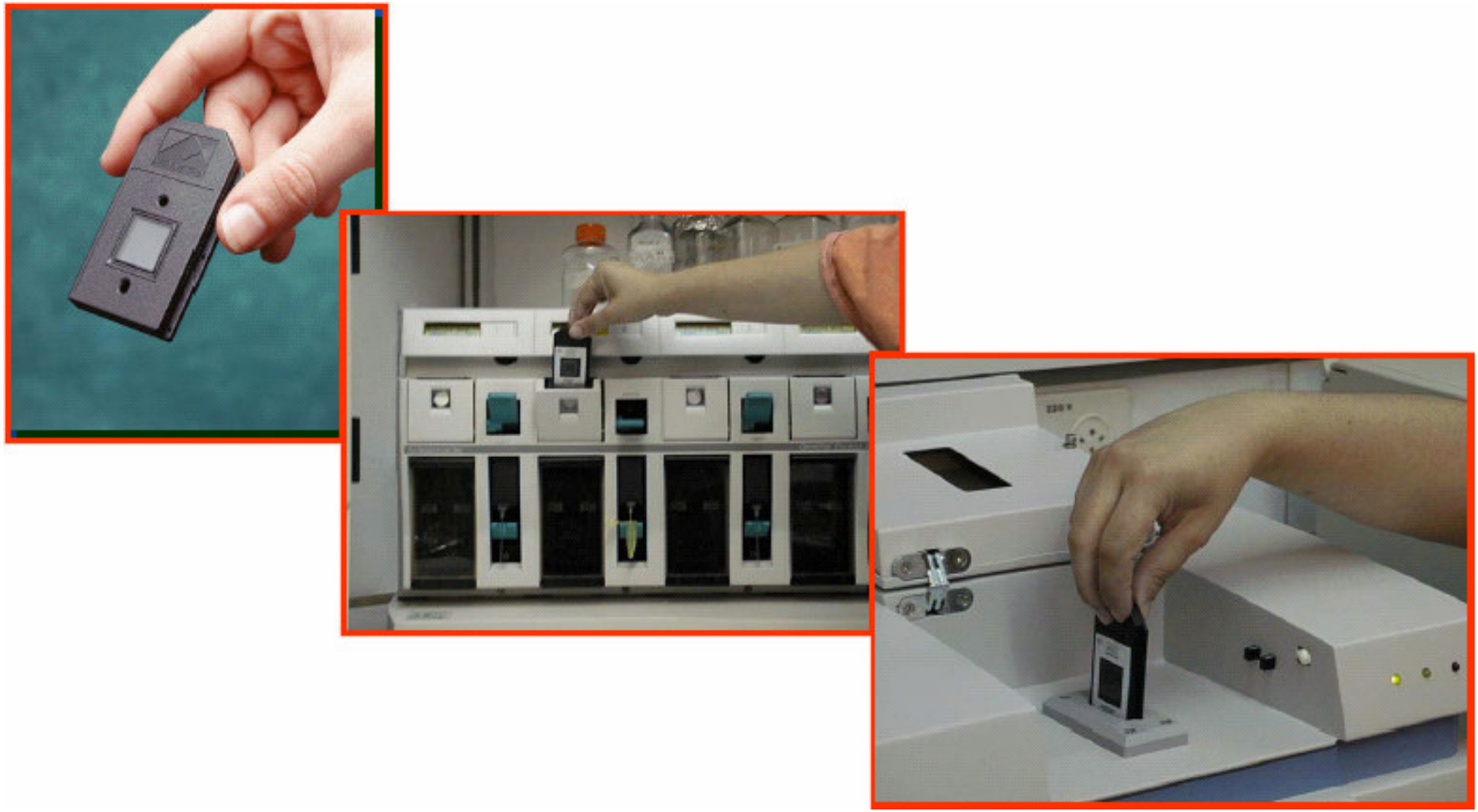
- DNA double strands form by "gluing" of complementary single strands.
- RNA transcript, introduced during the renaturation process, competes with the coding DNA strand and forms double-stranded DNA/RNA hybrid molecule



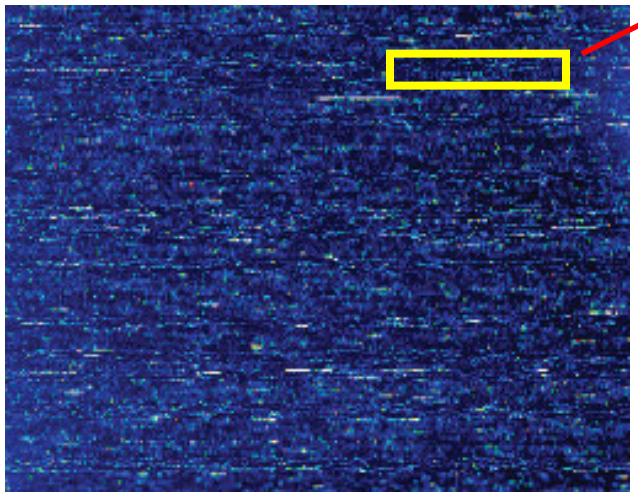
The technology: the whole picture



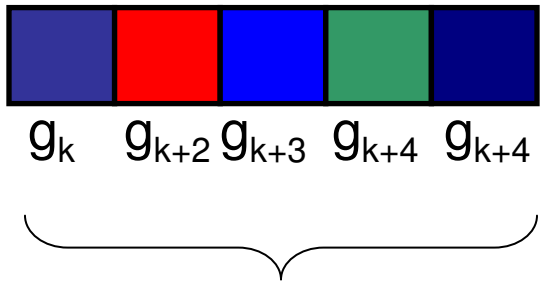
The technology: Affymetrix Chip



The technology: scanning



An observation



Genes expression levels

Microarray Data Analysis Types

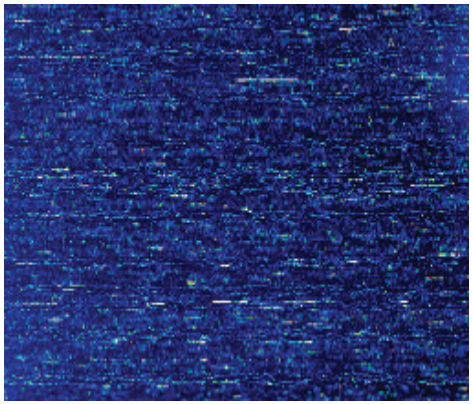
- Gene Selection
  - Find genes for therapeutic targets (new drugs)
- Classification (supervised)
  - Identify disease
  - Predict outcome / select best treatment
- Clustering (unsupervised)
  - Find new biological classes / refine existing ones
  - Exploration (discovery of unknown classes)

Challenges

- Main challenges
  - Few samples (usually < 100) but many features (usually genes > 1000)
  - High probability of finding **false positives,** that are knowledge discovered due to random noise
  - Models discovered need to be explainable to biologists
- Main steps
  - **Data preparation**
  - **Feature selection**
  - **Apply a classification methods**
  - **Tuning parameters with crossvalidation**

Preparing data

- Microarray data is translated in a **n x p table**, where n is the number of observations and p is the number of genes tested
- Each element <i,j> of the table is the expression level of gene j in the observation I
- Thresholds and transformations are applied to data
- Genes with a not significant variability through the whole dataset are excluded



	Gene 1	Gene 2	Gene 3
Sample 1	104	3208	40
Sample 2	32	1095	41

Genes Selection

- Most learning algorithms look for non-linear combinations of features
  - Can easily find ***spurious*** combinations given few records and many genes ("false positives problem")
- Classification accuracy improves if we first reduce number of genes by a linear method
  - e.g. T-values of mean difference

$$\frac{(Avg_1 - Avg_2)}{\sqrt{(\sigma_1^2 / N_1 + \sigma_2^2 / N_2)}}$$

- Select the top N genes from each class

Genes Selection: Randomization Approach

Class	...	Gene i	...
Healthy	...	3208	...
Sick	...	56	...
Sick	...	80	...
Healty	...	4560	...
...	...	...	...

T-Value

- Is T-Value outcome due to chance ?

Genes Selection: Randomization Approach

Class	…	Gene i	…
Healthy	…	3208	…
Sick	…	56	…
Sick	…	80	…
Healty	…	4560	…
…	…	…	…

T-Value 1

T-Value 2

…

T-Value M

- Is T-Value outcome due to chance ?
- Randomization approach
  - Generate M random permutations of the class columns
  - Compute T-values for each permutation and for each gene
  - How frequent a big T-value occurs for a random permutation?
  - Keep genes with high T-value and desired significance
- Limitations
  - Genes are assumed independent
  - Randomization is a conservative approach

Genes Selection: Wrapper Approach

- Generate several models and evaluate them
  - Apply T-Values to identify the top N genes
  - Evaluate (with crossvalidation) the accuracy of the model learned using all the subset of genes selected
  - Choose the simplest model that reaches the best performance
- Issues
  - Computationally expensive
  - Validation sets used in the genes selection process cannot be used to assess the final performance of the model!

Classification Methods

- Decision Trees/Rules
  - Model easy to understand
  - Find smallest gene sets, but not robust
  - Poor performance
- Neural Nets
  - Work well for reduced number of genes
  - Model is difficult to understand
- K-nearest neighbor
  - Good results for small number of genes, but no model
- Naïve Bayes
  - Simple, robust, but ignores gene interactions
- Support Vector Machines (SVM)
  - Good accuracy, does own gene selection, but hard to understand
- …

Biological Sequence Alignment

Alignment of biological sequence (1)

- Given two or more input biological sequences, identify similar sequences with long conserved subsequences
- Sequences can be either nucleotides (DNA/RNA) or amino acids (proteins)
  - Nucleotides align with if they are identical
  - Amino acids align if identical or if one can be derived from the other
- Tasks
  - Pairwise sequence alignment
  - Multiple sequence alignment
- Applications
  - Discovering phylogentic trees
  - Similarity searches

Alignment of biological sequence (2)

- **Substitution matrix** is used to define
  - cost of substitutions
  - cost of insertions and deletions
- Cost is inversely proportional to the probability that a substitution/insertion/deletion occurred
- Gaps ("—") can be used to indicate positions where it is preferable not to align two symbols
- The introduction of a gap ("—") is usually associated to a negative cost (**penalty)**

Example

- Align the following sequences:

HEAGAWGHEE
PAWHEAE

- Evaluate the following alignments according to the substitution matrix provided and the a gap penalty of -8

	A	E	G	H	W
A	5	-1	0	-2	-3
E	-1	6	-3	0	-3
H	-2	0	-2	10	-3
P	-1	-1	-2	-2	-4
W	-3	-3	-3	-3	15

H	*E*	*A*	*G*	*A*	*W*	*G*	*H*	*E*	–	*E*	
		|			|		|	|		|	
P	–	*A*	–	–	*W*	–	*H*	*E*	*A*	*E*	
-2	-8	+5	-8	-8	+15	-8	+10	+6	-8	+6	**= 0**

H	*E*	*A*	*G*	*A*	*W*	*G*	*H*	*E*	–	*E*	
				|	|		|	|		|	
–	–	*P*	–	*A*	*W*	–	*H*	*E*	*A*	*E*	
-8	-8	-1	-8	+5	+15	-8	+10	+6	-8	+6	**= +1**

Pairwise sequence alignment

- Two major approaches
  - **Local alignment**, works on segments and merge them
  - **Global alignment**, works on entire sequence
- Global alignment approaches search for the optimal alignment starting from optimal subsequences
- *Needleman-Wunsch* and *Smith-Waterman* algorithms exploit **dynamic programming** to find the optimal solution
- Both these algorithm have a **computational complexity** that is **quadratic** w.r.t. **sequences length**!
- Local alignment approaches (e.g. BLAST and FASTA) may be not able to find the best alignment but are more suitable to deal with long sequencess

BLAST

- BLAST breaks the sequences in small fragments called **words**
- A word is a **k-tuple** of elements (typically 11 nucleotides or 3 amino acids)
- BLAST first builds an **hash tables** of **neighborhood** words, that are closely matching
- A closeness threshold is used and statistics is applied to define how significant the matches are
- Starting from a fragment, the alignment is extended in both the direction by choosing the best scoring matches
- BLAST has computationally complexity linear w.r.t. to the sequence length
- Several specialized versions of BLAST have been introduced
  - Protein similarity searches (BLASTP)
  - Variable word size (BLASTN)
  - Discontiguous alignments (MEGABLAST)

Multiple Sequence Alignment Methods

- Is important both in phylogenetic analysis and in the discovery of protein structures
- Multiple alignment is computationally more challenging
- **Freng-Doolittle alignment**
  - Performs the pairwise alignments
  - Merge them following a guide tree generated with a hierarchical clustering approach
- Hidden Markov Models
  - More sophisticated probabilistic approach to represent statistical regularities in the sequences