生物信息中的统计模型

2015年春

有关信息

- 邓明华
  - Email: dengmh@pku.edu.cn
  - Office: 北京大学理科一号楼1579E
  - Phone: 62767562, 13522856599

- 课程网页

考评

- 平时作业：20%
- 期末Project：80%

以下PPT部分来源于
Zhaohui Qin(University of Emory)
2011年北京大学统计中心暑期课程第一讲

Bioinformatics

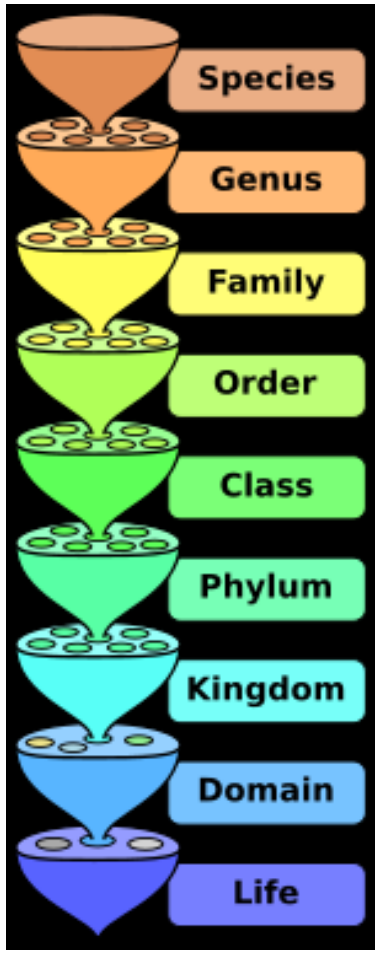
Bioinformatics is the research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, archive, analyze, or visualize such data;

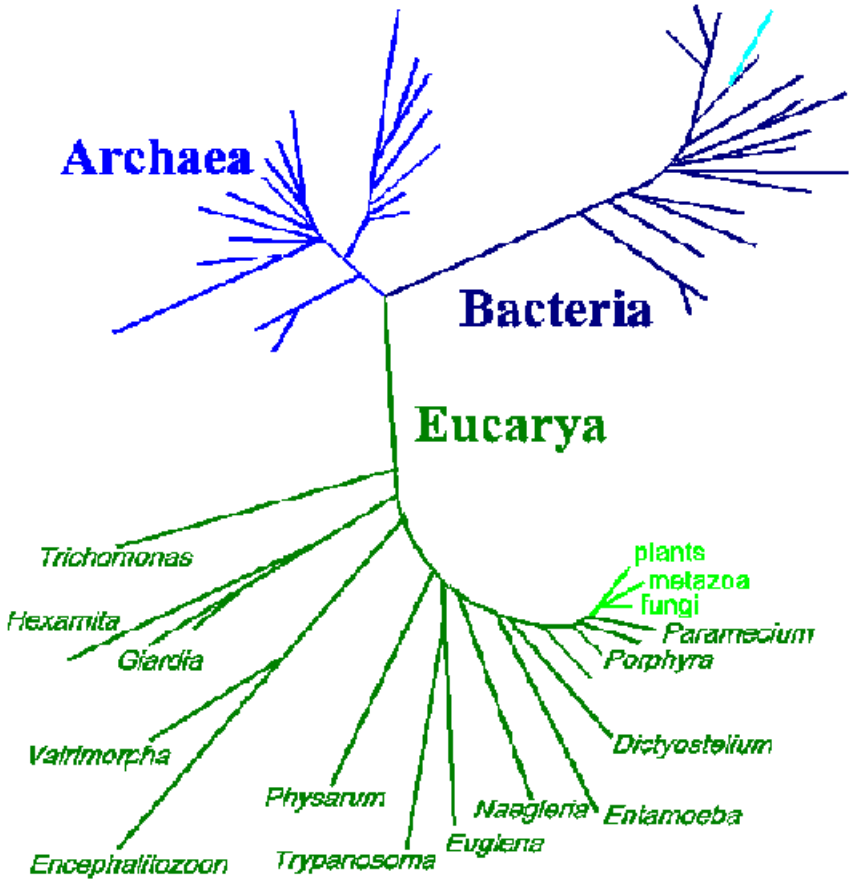
Computational biology is the development and application of data-analytical and theoretical methods, mathematical modeling and computational simulation techniques to the study of biological, behavioral, and social systems.

Bioinformatics

- Overlay of **Biology**, **Computer science** and **Statistics**.
- Topics:
    - Sequence alignment
    - Protein folding
    - Gene finding
    - Functional annotation
    - Network inference

Tree of Life





modified from N.R. Pace, ASM News 62:464, 1996

Molecules of Life

- DNA
- RNA
- Protein

DNA

- Deoxyribonucleic acid(脱氧核糖核酸)
- Consist of four nucleotides
  - A Adenine(腺嘌呤)
  - C Cytosine(胞嘧啶)
  - G Guanine(鸟嘌呤)
  - T Thymine(胸腺嘧啶)

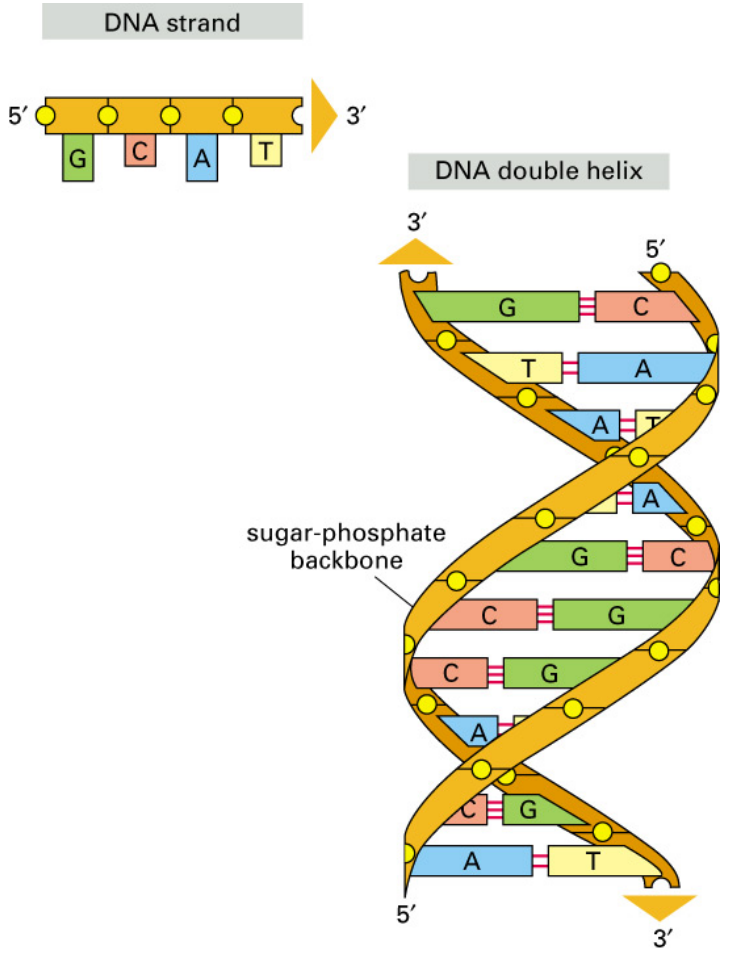
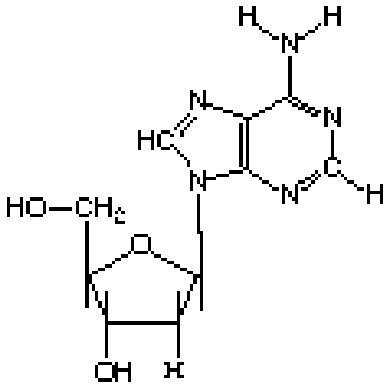
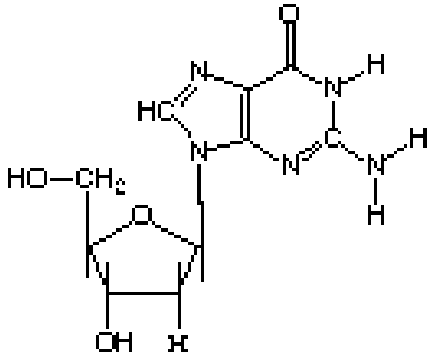


Figure 4–3 part 2 of 2. Molecular Biology of the Cell, 4th Edition.

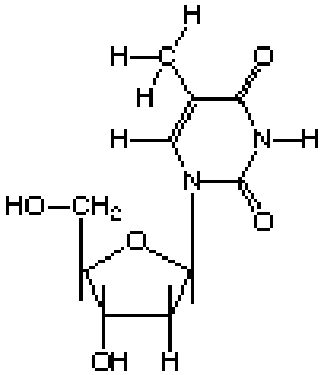
DNA



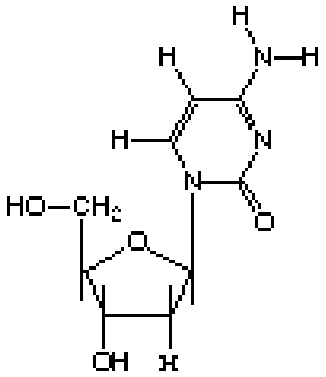
Adenine



Guanosine



Thymine



Cytosine

Purines (嘌呤)

Pyrimidines(嘧啶)

Base Pairing

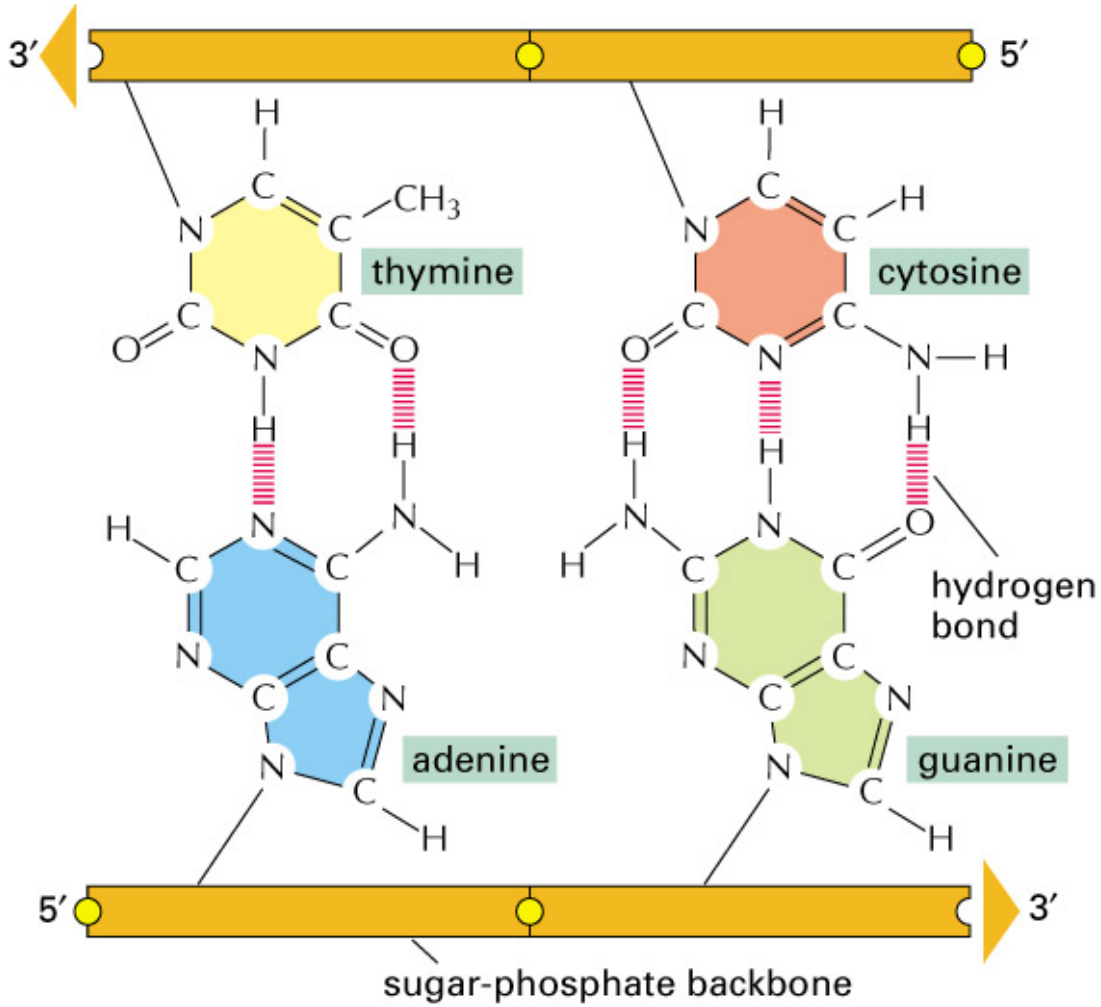
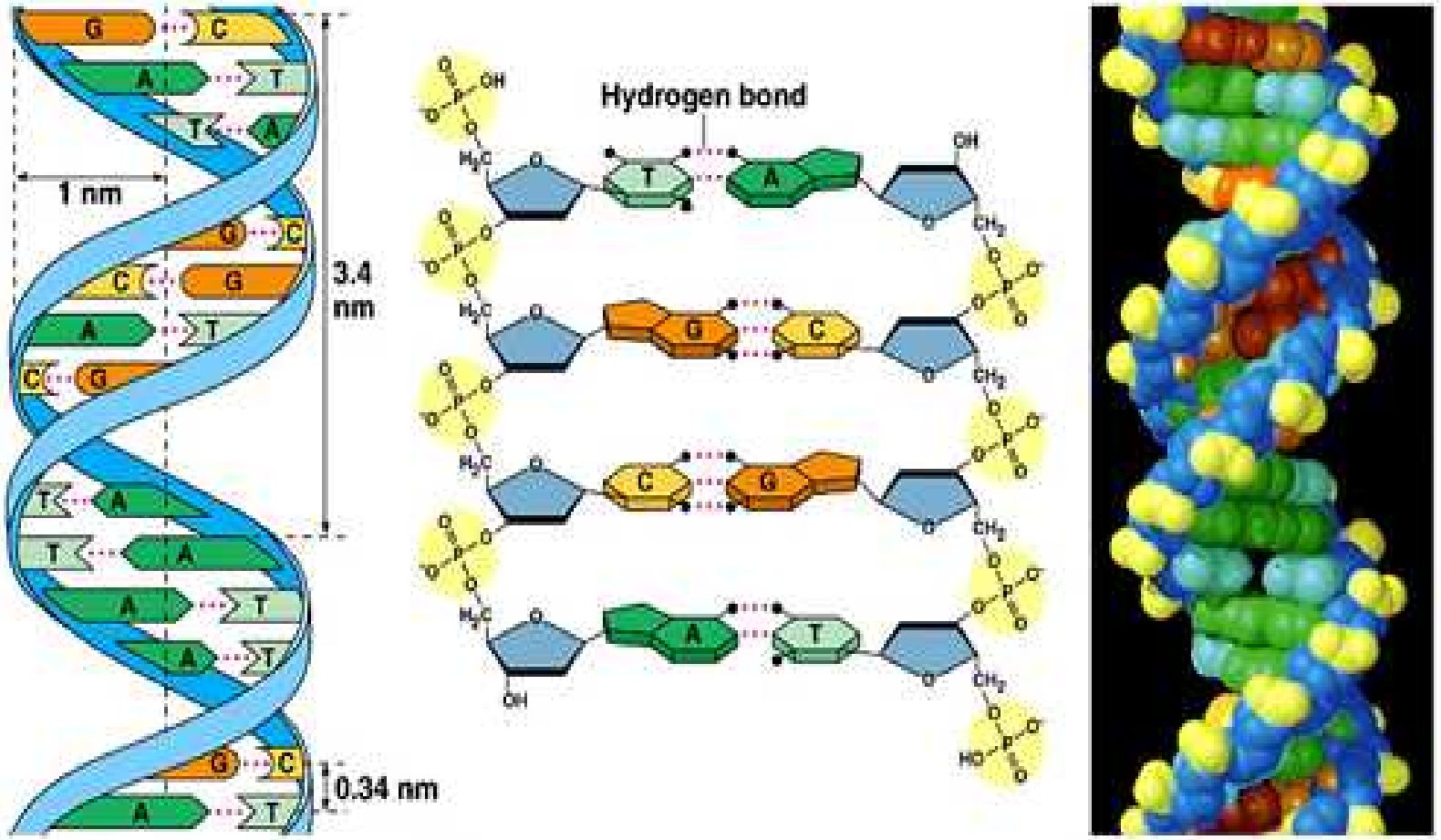
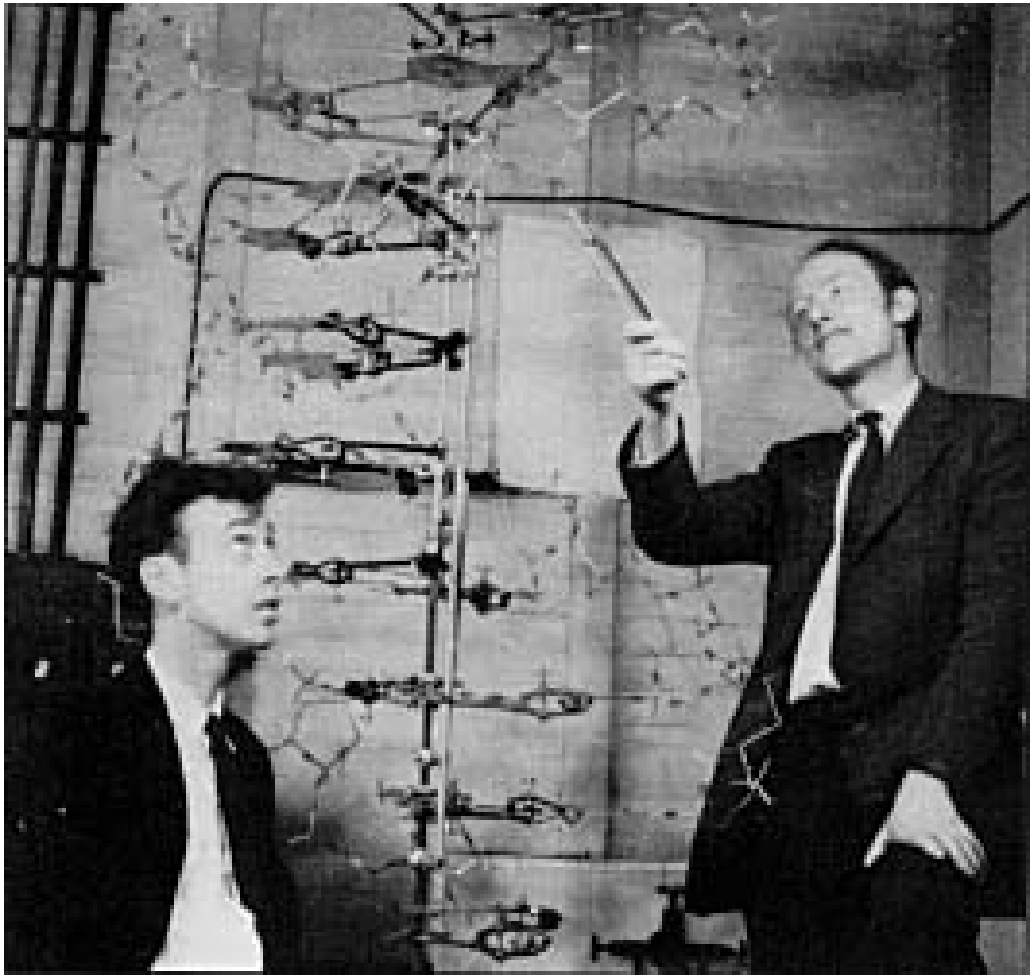


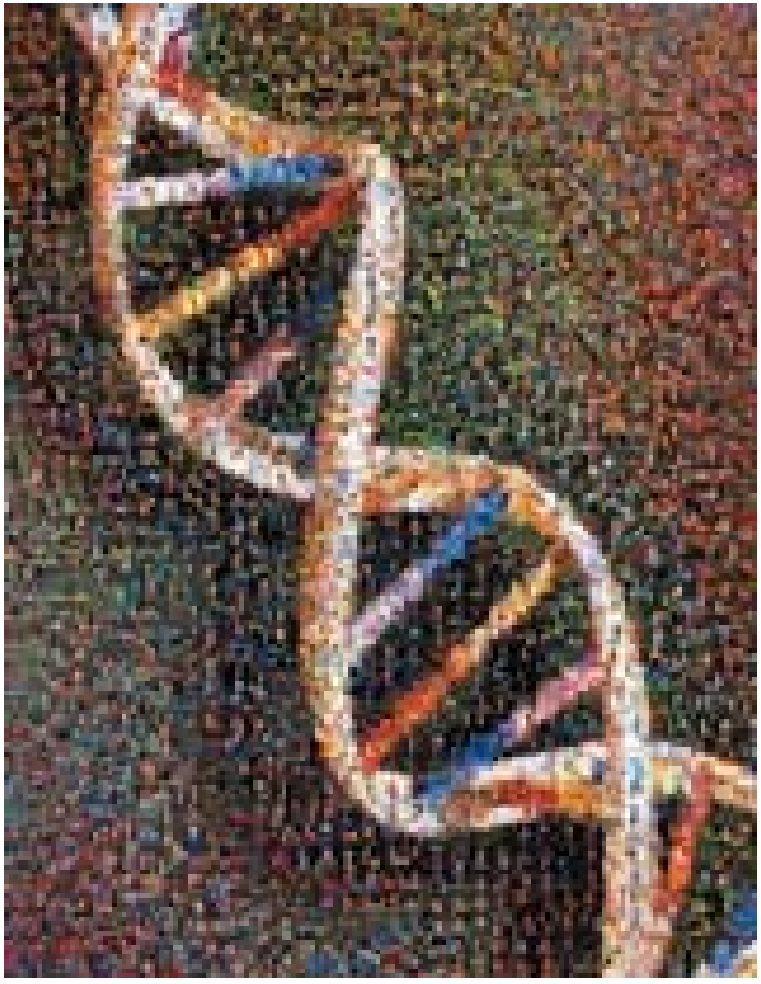
Figure 4–4. Molecular Biology of the Cell, 4th Edition.

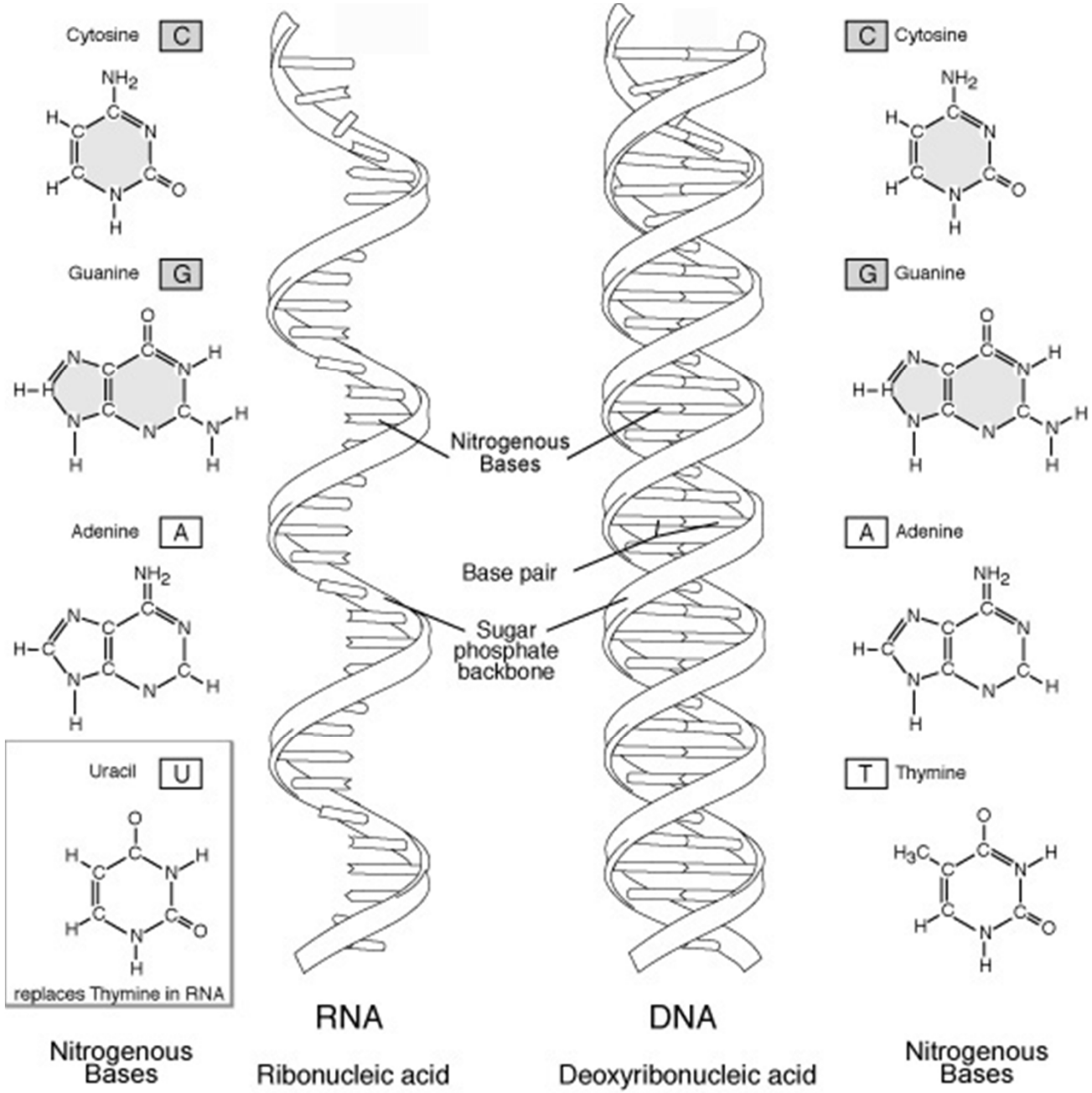
Structure of DNA



DNA Structure



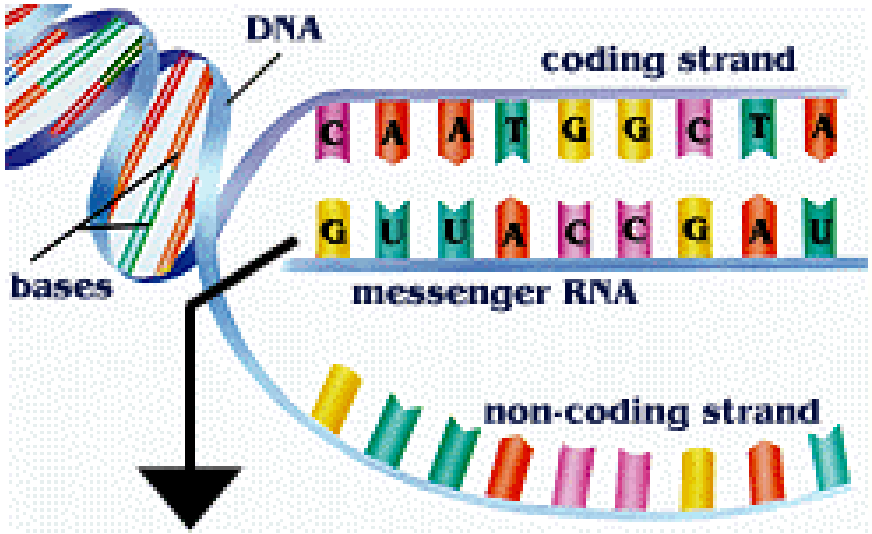



Cytosine C
Guanine G
Adenine A
Uracil U
replaces Thymine in RNA
Nitrogenous Bases
Nitrogenous Bases
Base pair
Sugar phosphate backbone
RNA
Ribonucleic acid
DNA
Deoxyribonucleic acid
C Cytosine
G Guanine
A Adenine
T Thymine
Nitrogenous Bases

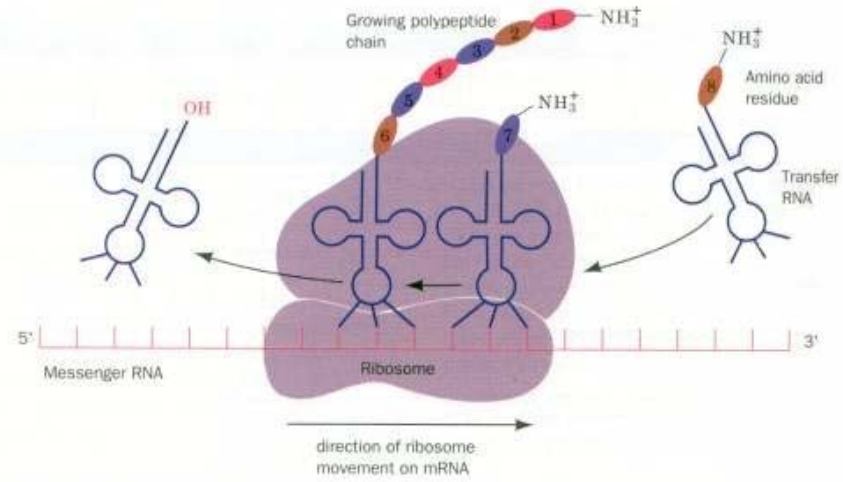
RNA

- Ribonucleic acid (核糖核酸)
  - mRNA : Messenger RNAs, code for proteins
  - rRNA: Ribosomal RNAs, form the basic structure of the ribosome and catalyze protein synthesis.
  - tRNA: Transfer RNAs, central to protein synthesis as adaptors between mRNA and amino acids.

mRNA and tRNA



Transcription



Translation

Proteins

Main building blocks and functional molecules, take up ~20% of eukaryotic cell's weight.

- Structural proteins
- Enzymes
- Antibodies
- Transmembrane proteins

Genetic Code

Instructions

1. Find the letter of the first base of the codon in the column at the left.
2. Go across this row until you are in the column headed by the letter of the second base.
3. Then find the third base, marked at the far right of the table.
4. In each box, the words in black are the amino acids specified by the codons in blue.

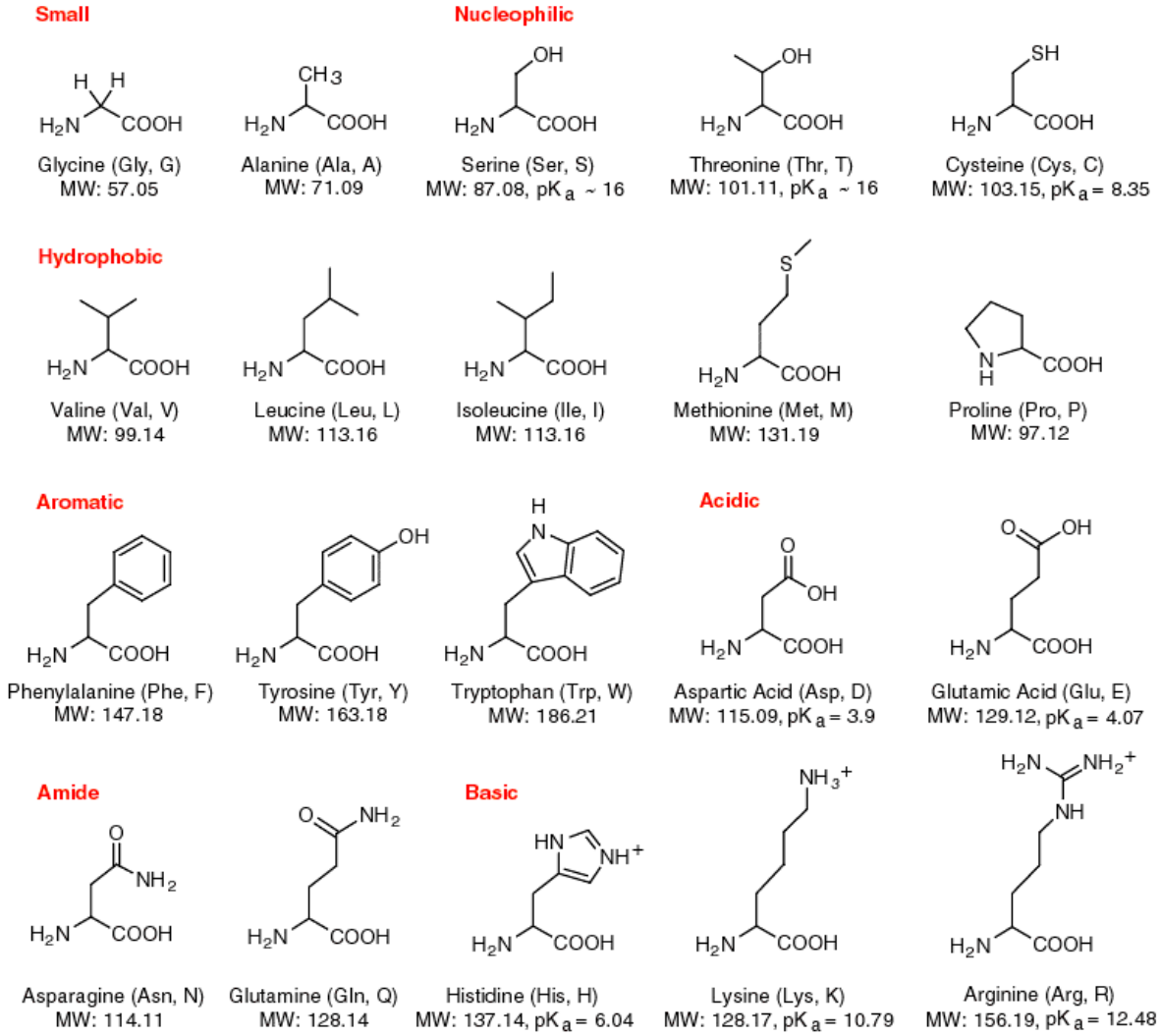
First Base (5′)	Second Base: U	Second Base: C	Second Base: A	Second Base: G	Third Base (3′)
U	UUU, UUC Phenylalanine UUA, UUG Leucine	UCU, UCC, UCA, UCG Serine	UAU, UAC Tyrosine UAA STOP† UAG STOP†	UGU, UGC Cysteine UGA STOP† UGG Tryptophan	U C A G
C	CUU, CUC, CUA, CUG Leucine	CCU, CCC, CCA, CCG Proline	CAU, CAC Histidine CAA, CAG Glutamine	CGU, CGC, CGA, CGG Arginine	U C A G
A	AUU, AUC, AUA Isoleucine AUG Methionine (START)*	ACU, ACC, ACA, ACG Threonine	AAU, AAC Asparagine AAA, AAG Lysine	AGU, AGC Serine AGA, AGG Arginine	U C A G
G	GUU, GUC, GUA, GUG Valine	GCU, GCC, GCA, GCG Alanine	GAU, GAC Aspartic acid GAA, GAG Glutamic acid	GGU, GGC, GGA, GGG Glycine	U C A G

*The codon AUG initiates synthesis of a polypeptide and calls for methionine as the first amino acid.

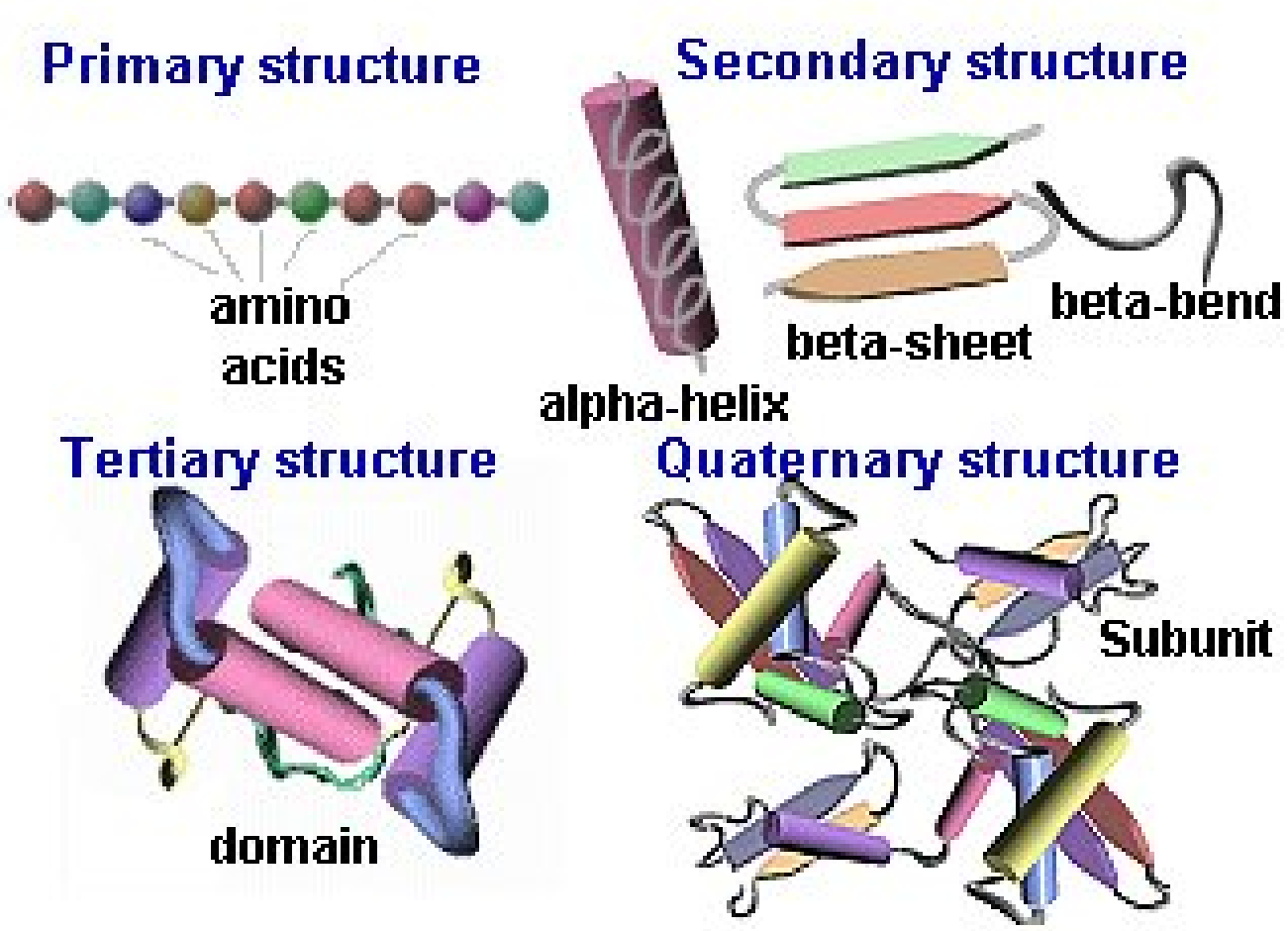
†The three STOP codons signal positions where the ribosome stops reading and terminates the polypeptide chain.

Amino Acids

Amino Acids are compounds that contain both an amino group (-NH3+) and a carboxy acid group (-COO-). 19 of the 20 amino acids fit this rule, and in addition, have a single carbon atom between these two groups, and a variable extension off this carbon atom. The exception is proline.

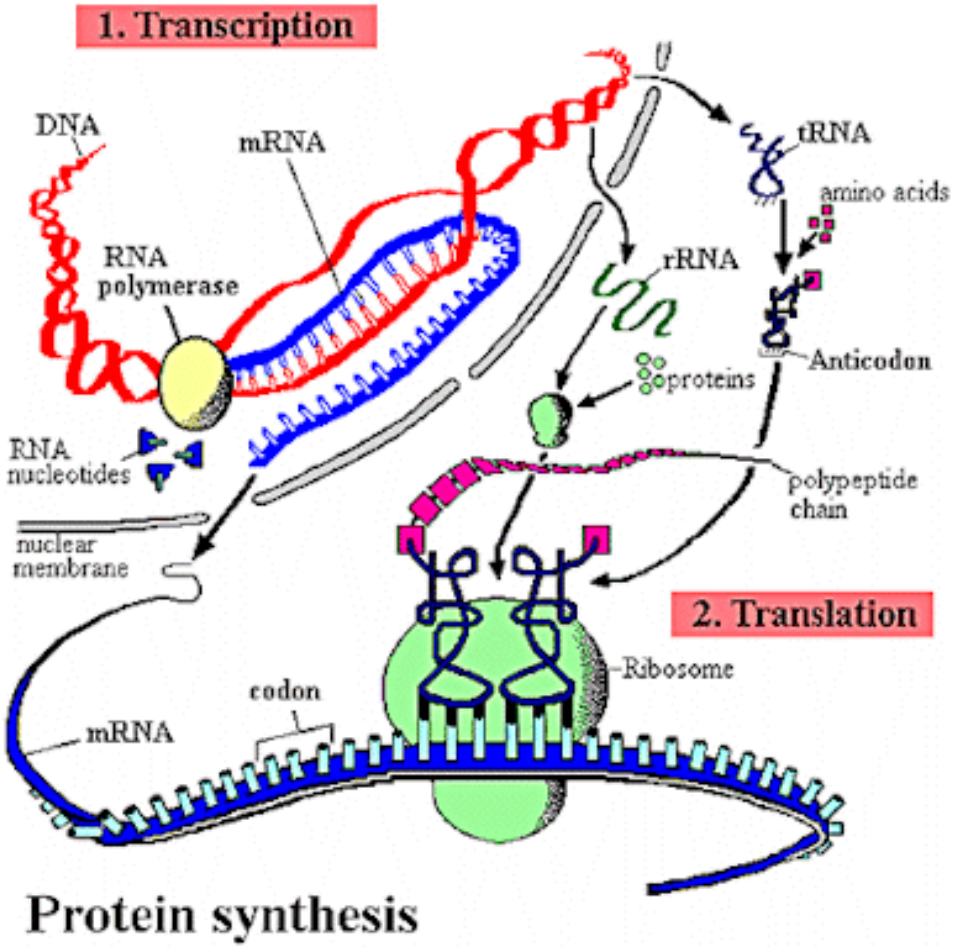


Protein Structure



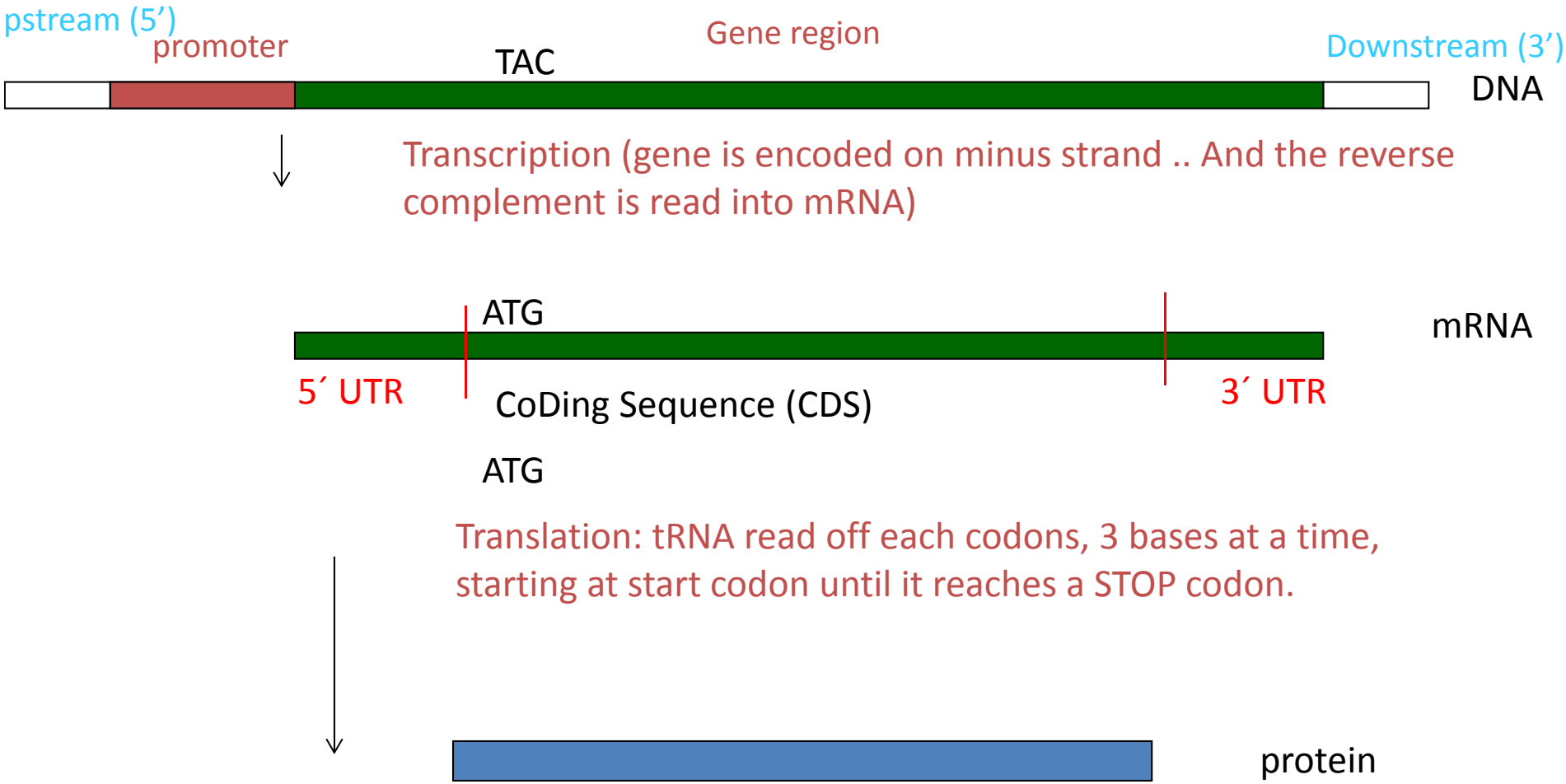
The Central Dogma

DNA→RNA→Protein

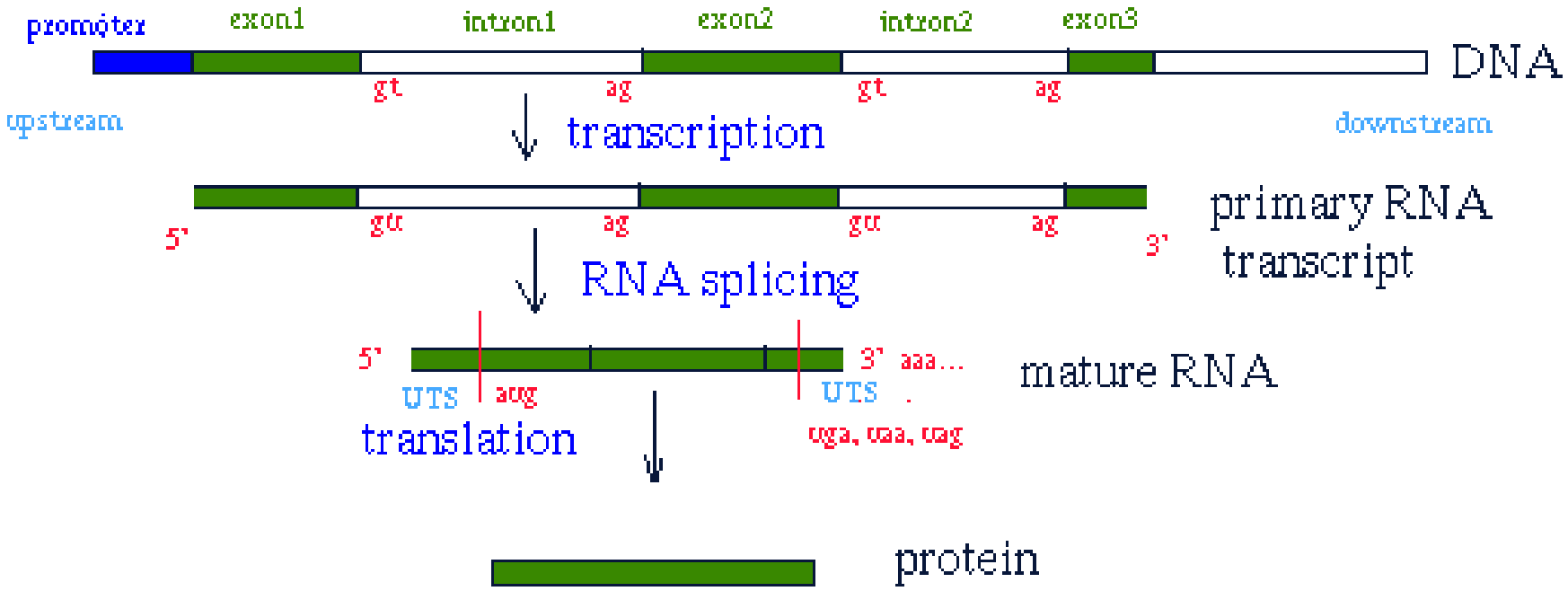


Prokaryotic Genes

Prokaryotes (intronless protein coding genes)



Eukaryotic Genes



Transcription

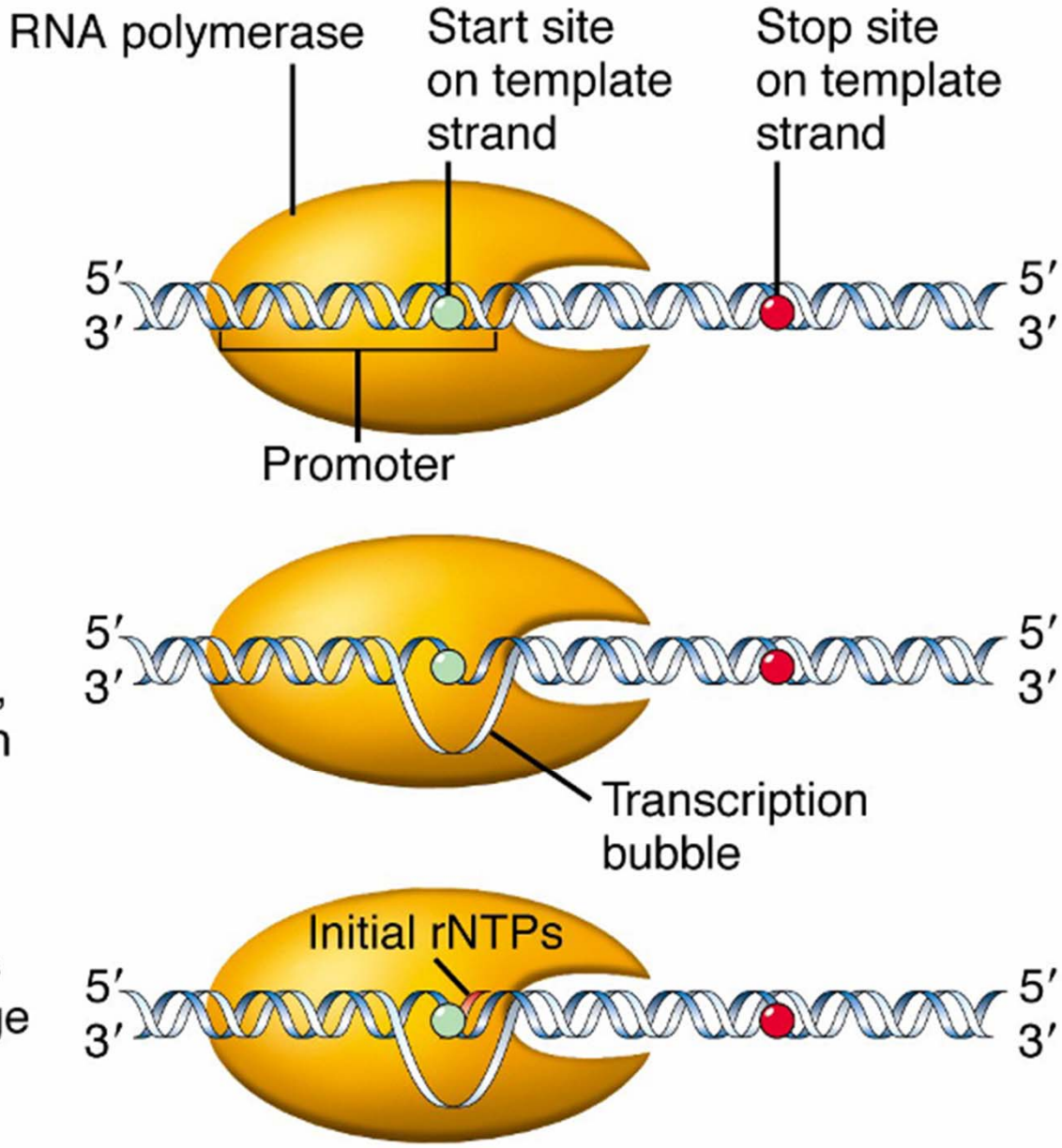
- Process of copying DNA to RNA
- Differs from DNA synthesis in that only one strand of DNA, the *template strand*, is used to make mRNA
- Does not need a primer to start
- Can involve multiple RNA polymerases
- Divided into 3 stages
  - Initiation
  - Elongation
  - Termination

INITIATION

1 Polymerase binds to promoter sequence in duplex DNA. "Closed complex"

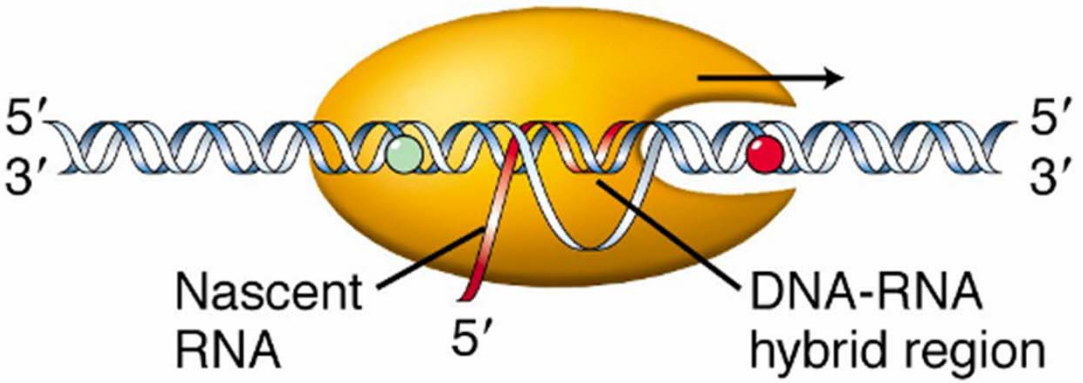
2 Polymerase melts duplex DNA near transcription start site, forming a transcription bubble. "Open complex"

3 Polymerase catalyzes phosphodiester linkage of two initial rNTPs.



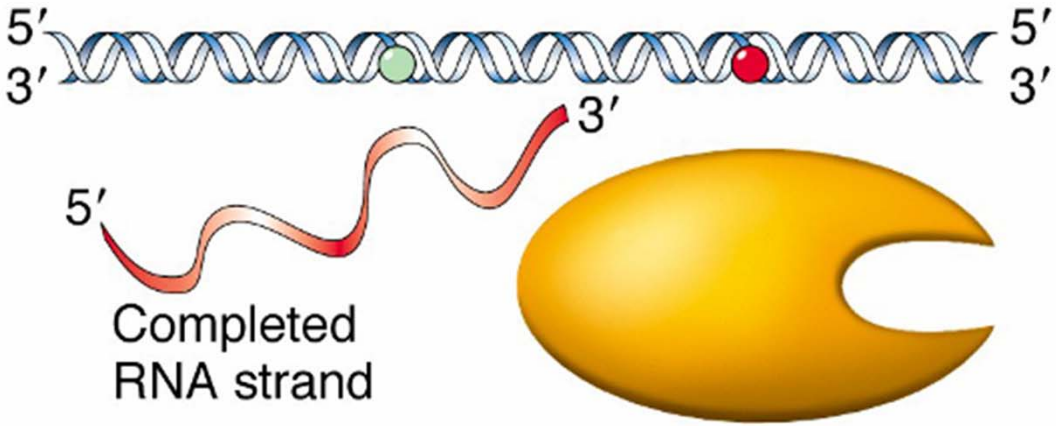
ELONGATION

4 Polymerase advances 3′ → 5′ down template strand, melting duplex DNA and adding rNTPs to growing RNA.

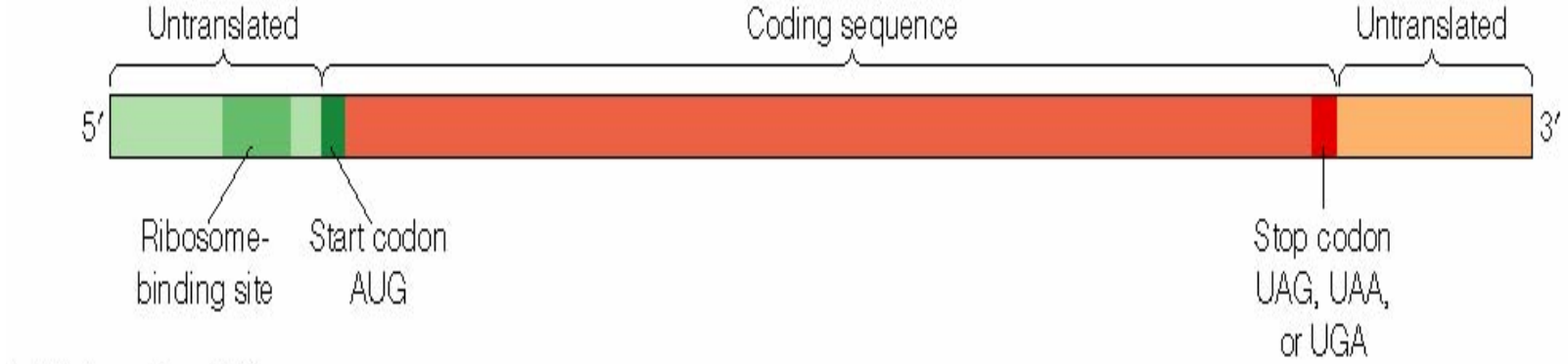


TERMINATION

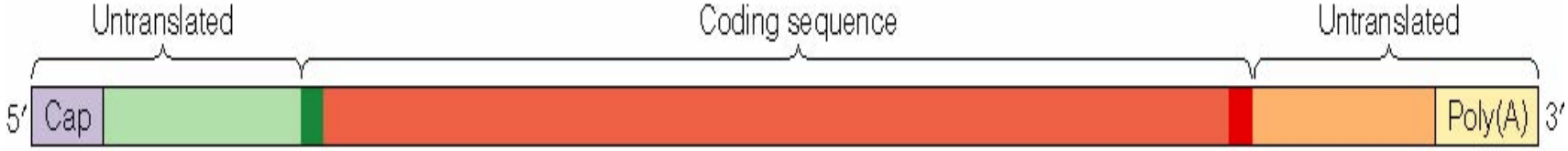
5 At transcription stop site, polymerase releases completed RNA and dissociates from DNA.



Transcription: The final product

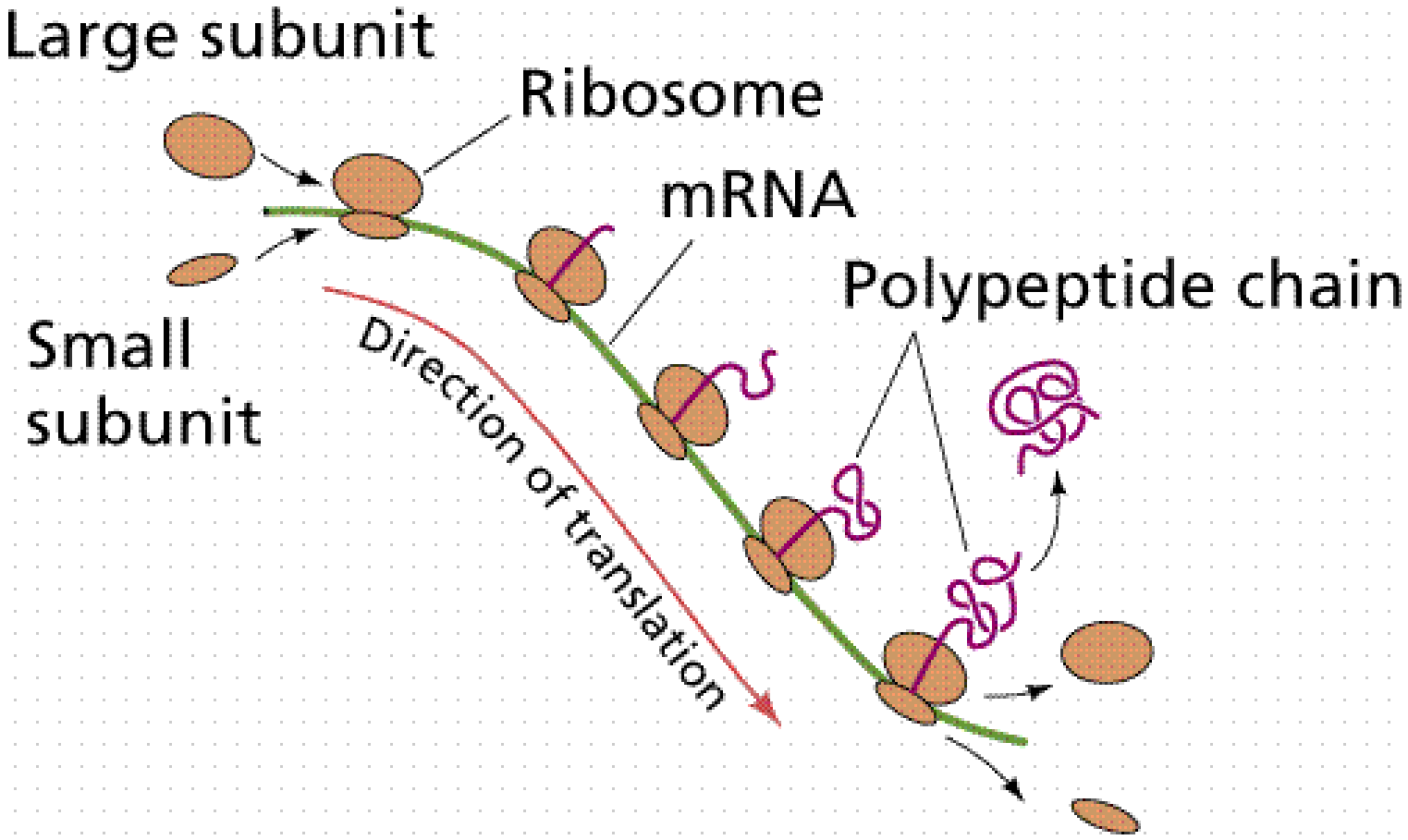


(a) Prokaryotic mRNA



(b) Eukaryotic mRNA

Translation



课程安排

- 生物背景和课程简介
- 概率论基础
- Hidden Markov Model (HMM)及其应用
  - Markov Chain
  - HMM理论
  - HMM和基因识别 (Topic I)
  - HMM和序列比对 (Topic II)
- 进化树的概率模型 (Topic III)
- Motif finding中的概率模型 (Topic IV)
  - EM algorithm
  - Markov Chain Monte Carlo (MCMC)
- 基因表达数据分析 (Topic V)
  - 聚类分析-Mixture model
  - Classification-Lasso Based variable selection
- 基因网络推断 (Topic VI)
  - Bayesian网络
  - Gaussian Graphical Model
- 基因网络分析 (Topic VII)
  - Network clustering
  - Network Motif
  - Markov random field (MRF)
- Dimension reduction及其应用 (Topic VIII)

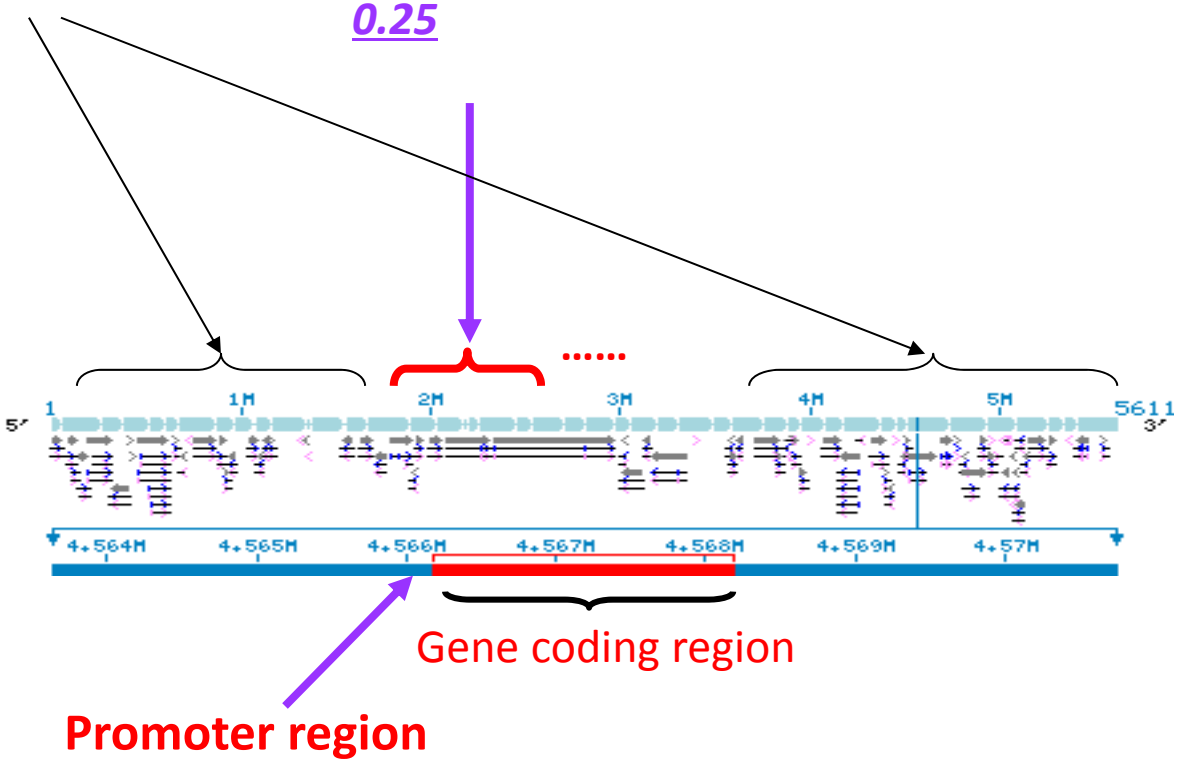
Topic I: Sequence's Feature Detection

- Problem I: CpG island finding
- Problem II: Gene finding (promoter prediction, Splicing site prediction, Translation Initial Site Prediction etc.)
- Hidden Markov Model is a powerful method for these problems

什么是CpG岛？

CG-poor regions: P(CG) ~ 0.07!

CG-rich region: P(CG) ~ 0.25



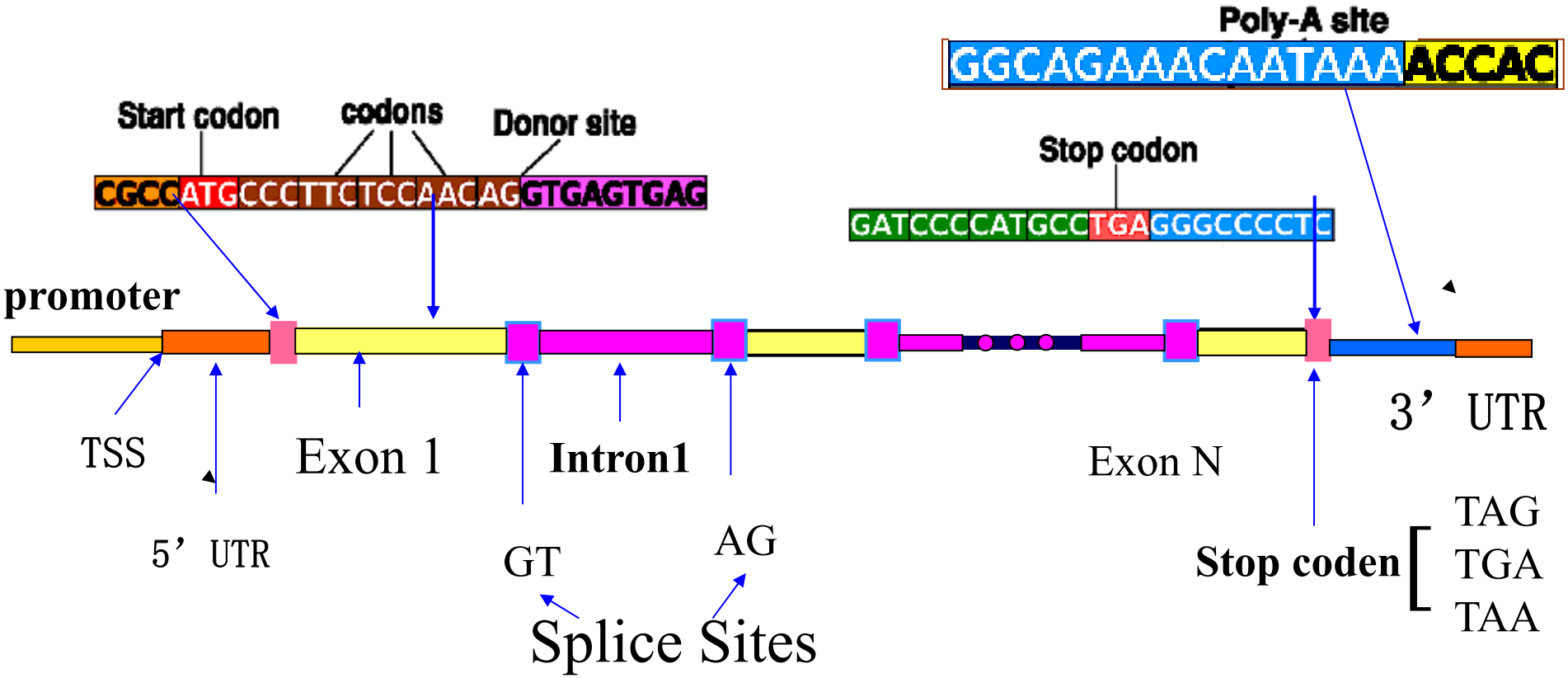
CpG岛的生物学意义

- CpG-rich regions are associated with genes which are *frequently transcribed*.
- Helps to understand gene expression related to *location* in genome.

CpG岛的生物学意义

- CpG-rich regions are associated with genes which are *frequently transcribed*.
- Helps to understand gene expression related to *location* in genome.

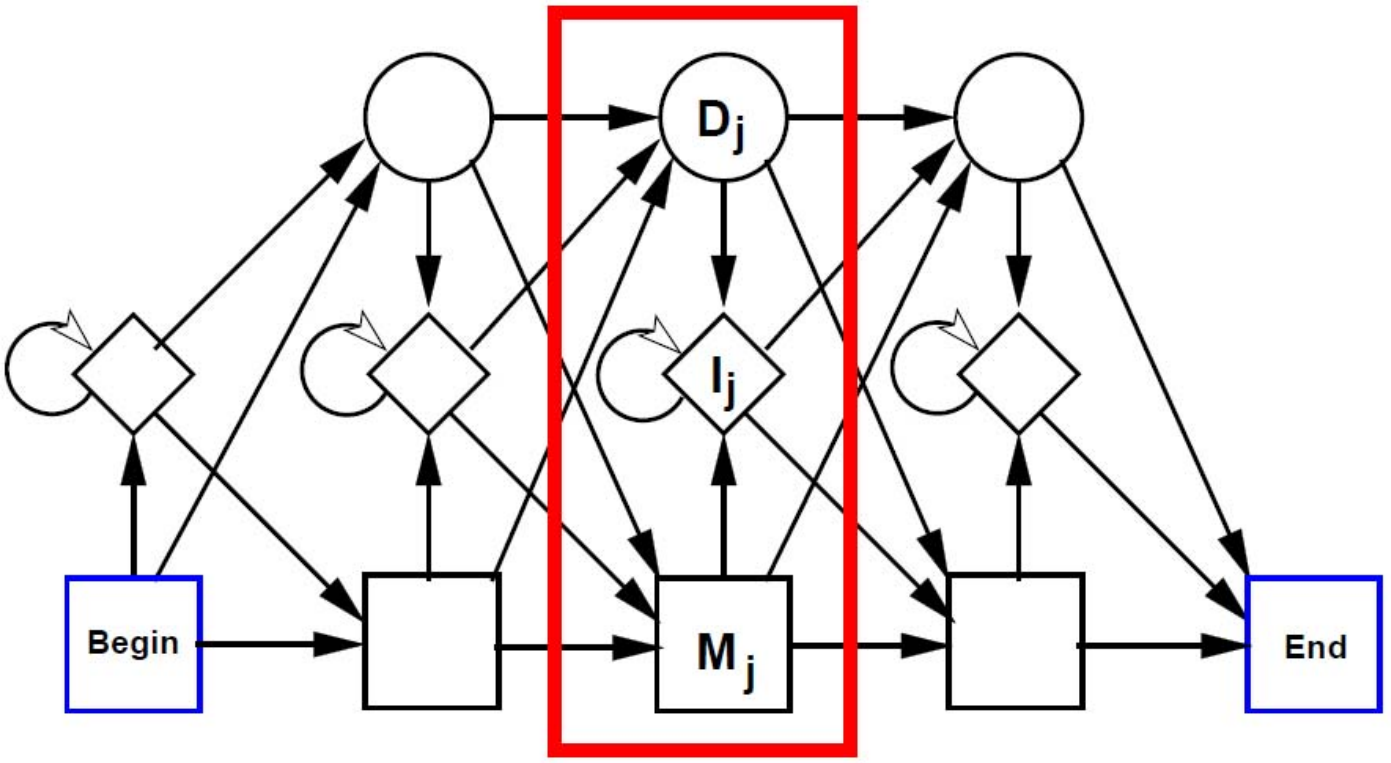
基因的结构


Poly-A site
GGCAGAAACAATAAAACCAC
Start codon
codons
Donor site
CGCCATGCCCTTCTCCAACAGGTGAGTGAG
Stop codon
GATCCCCATGCCTGAGGGCCCCTC
promoter
TSS
5’ UTR
Exon 1
Intron1
GT
AG
Splice Sites
Exon N
3’ UTR
Stop coden
TAG
TGA
TAA

Topic II: Multiple Alignment

	X	X	.	.	.	X
bat	A	G	-	-	-	C
rat	A	-	A	G	-	C
cat	A	G	-	A	A	-
gnat	-	-	A	A	A	C
goat	A	G	-	-	-	C
	1	2	.	.	.	3

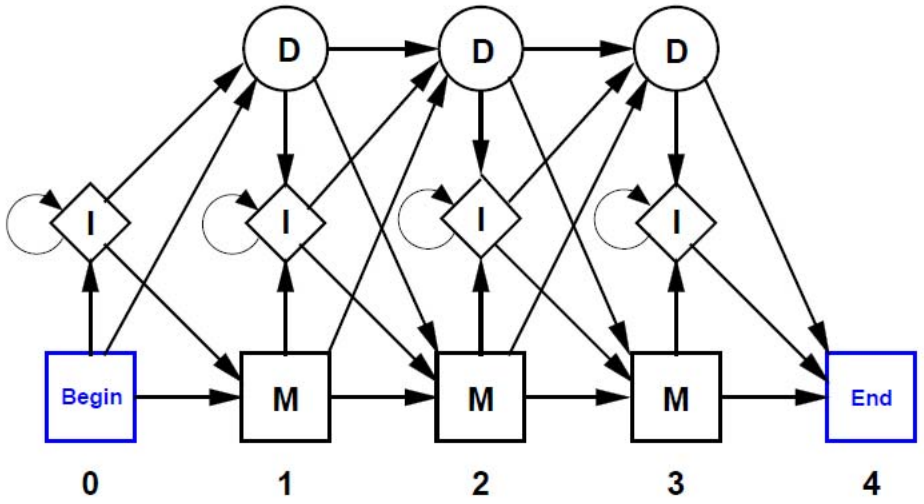
Profiled HMM



Transition structure of a profile HMM

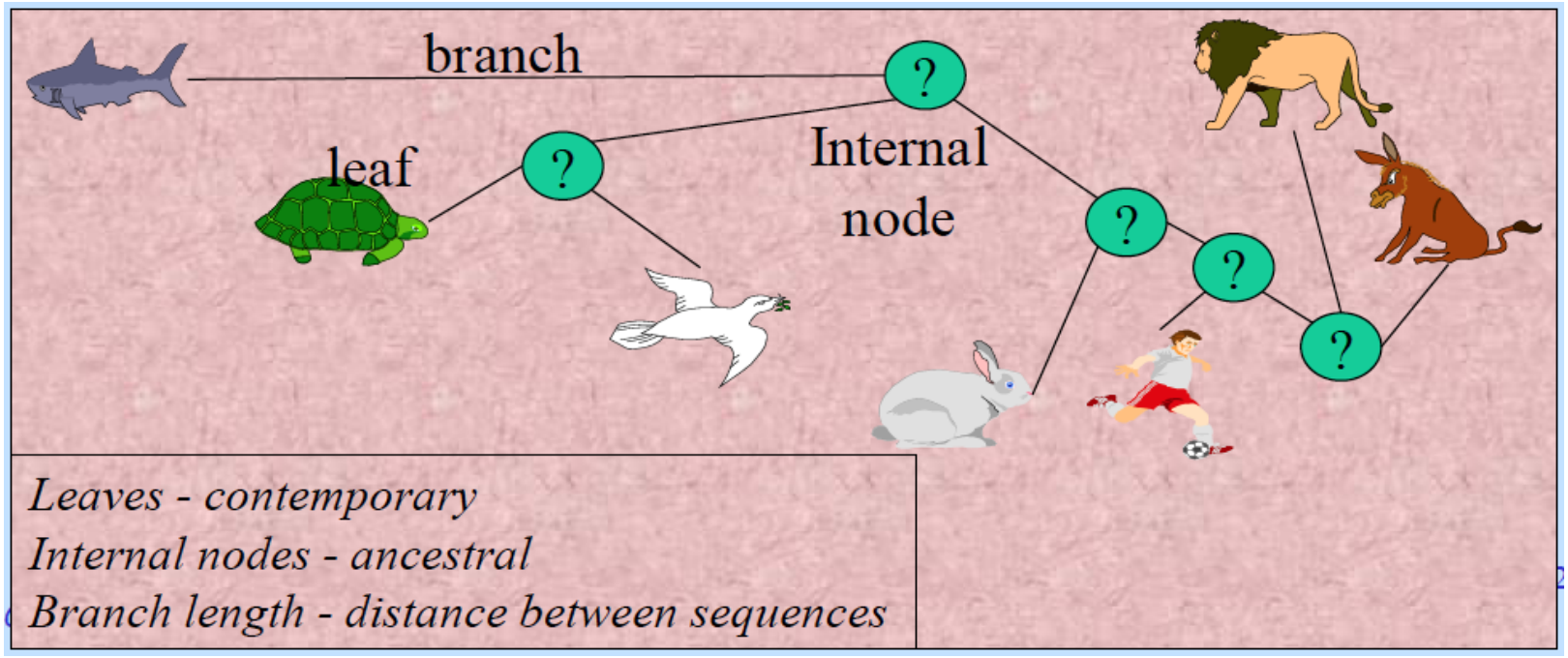
Example of Profile HMM

	X	X	.	.	.	X
bat	A	G	-	-	-	C
rat	A	-	A	G	-	C
cat	A	G	-	A	A	-
gnat	-	-	A	A	A	C
goat	A	G	-	-	-	C
	1	2	.	.	.	3



Topic III: Tree of lifes

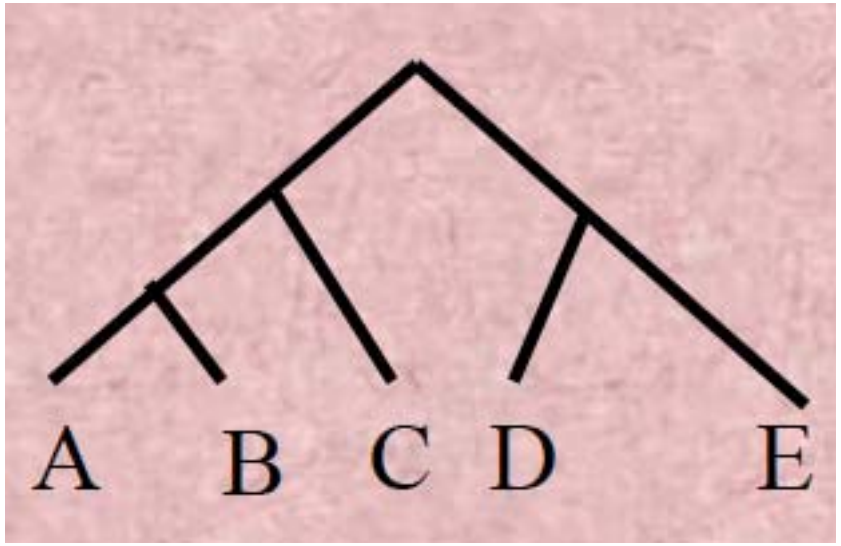
- Phylogeny: the ancestral relationship of a set of species
- Represented by a phylogenetic tree



Inferring a phylogenetic tree

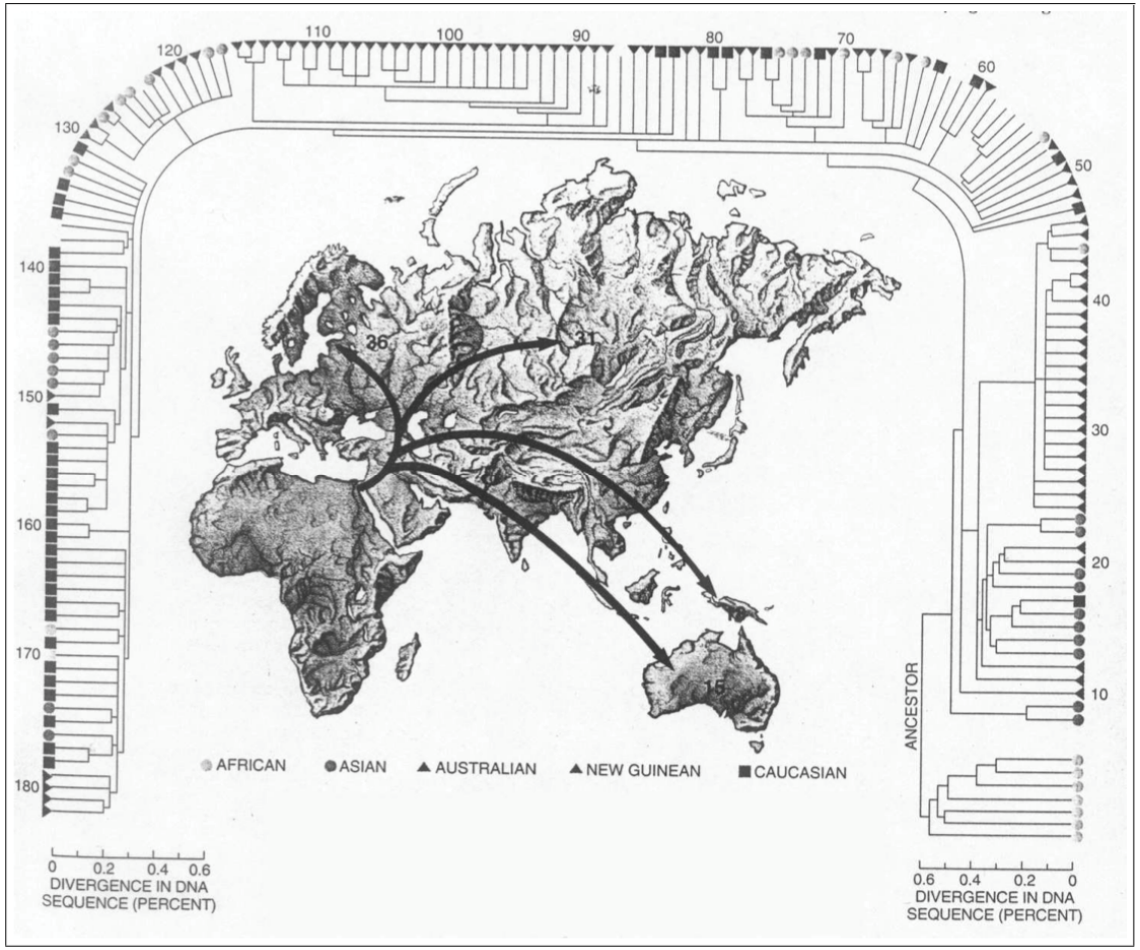
- Classical: morphological characters
- Modern: molecular sequences

```
A: CAGGTA
B: CAGACA
C: CGGGTA
D: TGCACT
E: TGCGTA
```



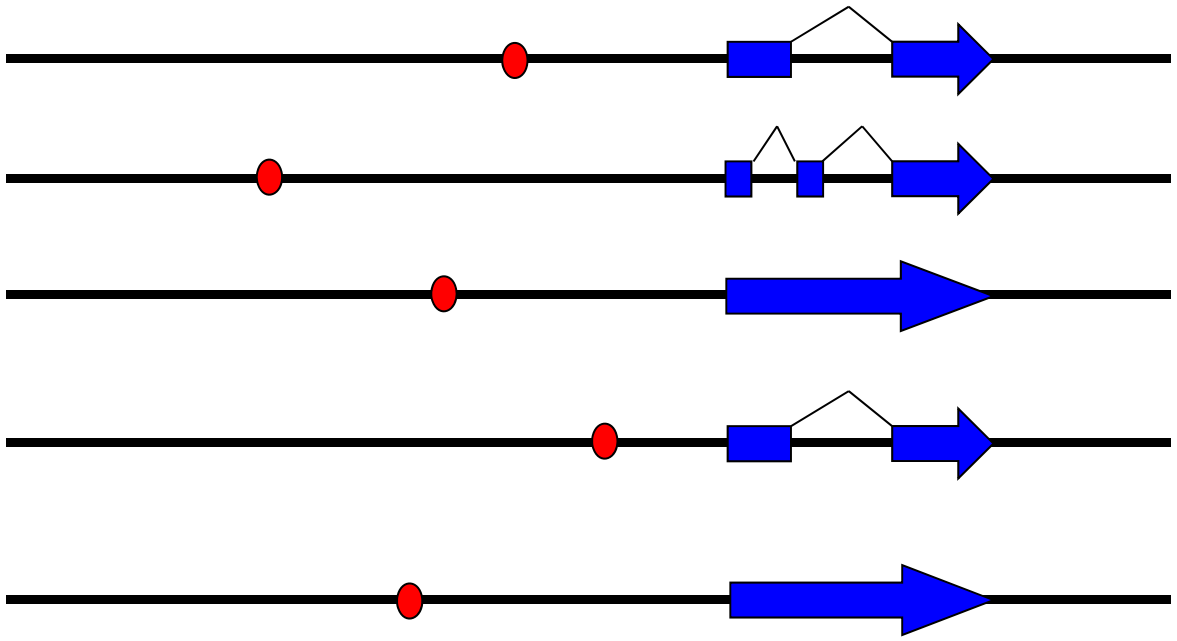
- Approaches: probabilistic model, bootstrap

An example: Out of Africa



Topic IV: Motif Finding

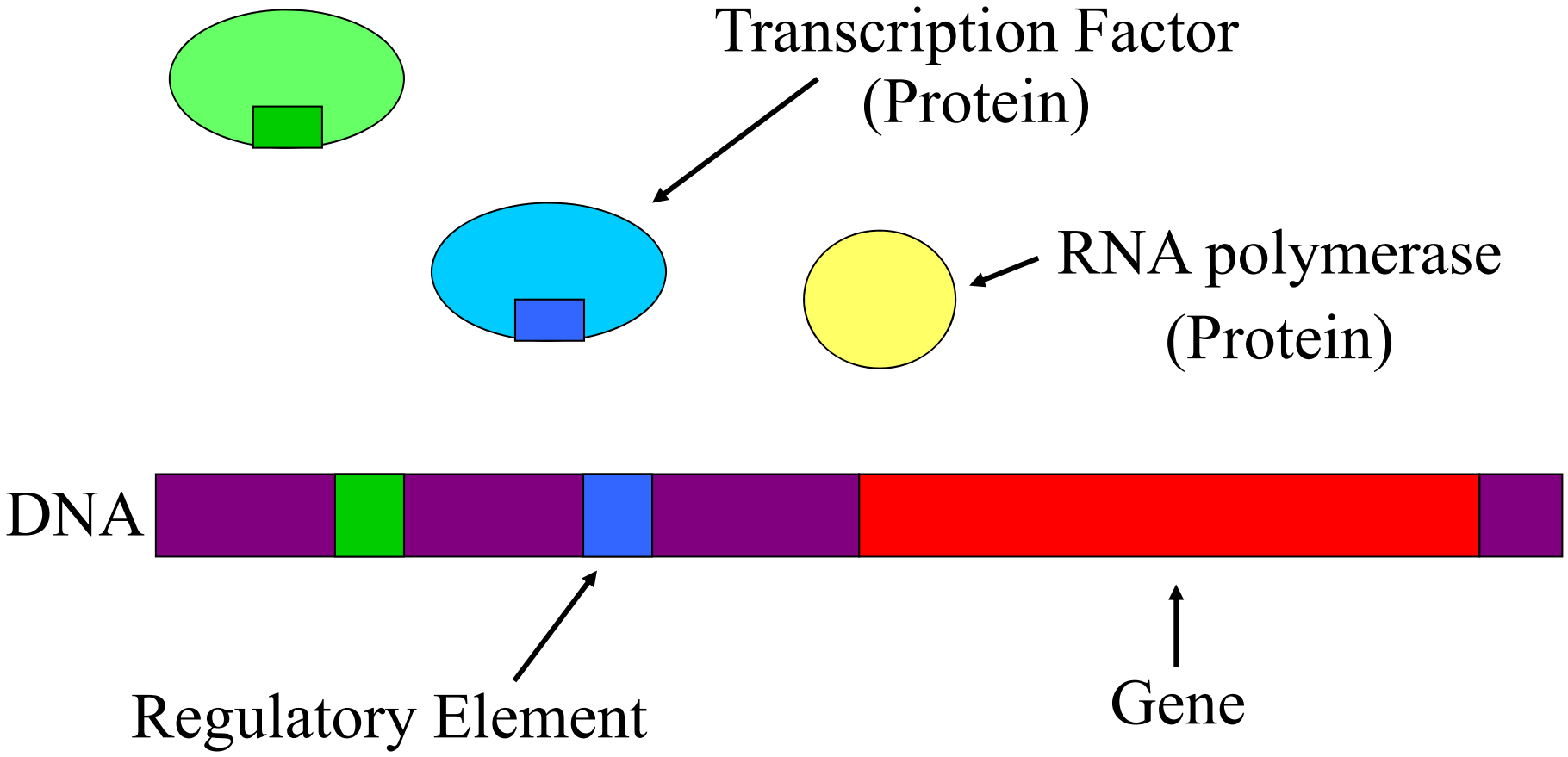
- Find promoter motifs associated with **co-regulated** or **functionally related** genes



Transcriptional Regulation

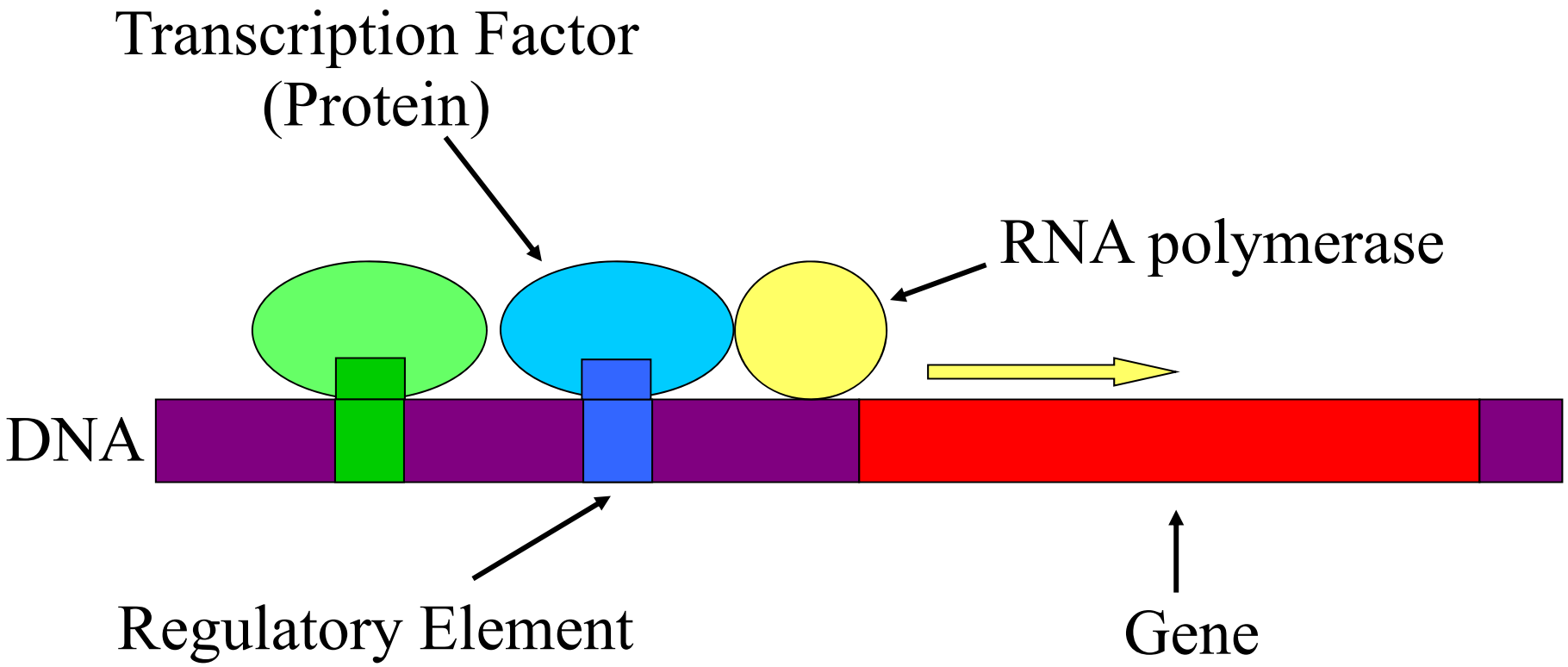
- The transcription of each gene is controlled by a regulatory region of DNA relatively near the transcription start site (TSS).
- two types of fundamental components
  - short DNA regulatory elements
  - *gene regulatory proteins* that recognize and bind to them.

Regulation of Genes



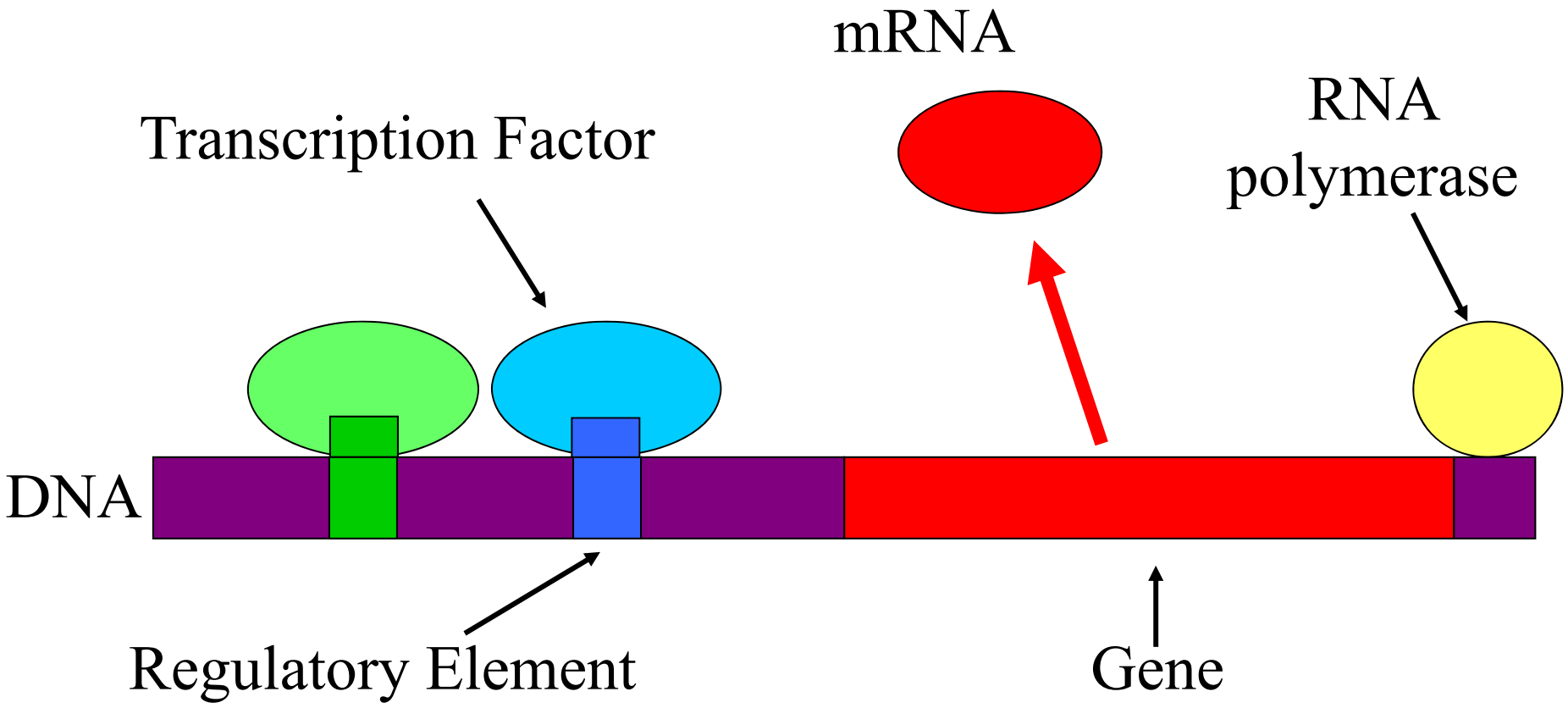
source: M. Tompa, U. of Washington

Regulation of Genes



source: M. Tompa, U. of Washington

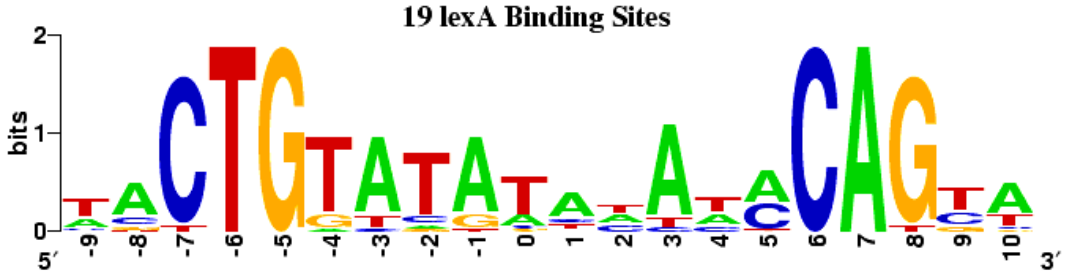
Regulation of Genes



source: M. Tompa, U. of Washington

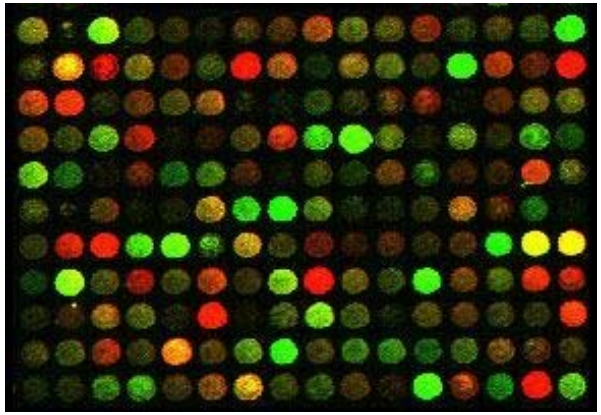
Motif Finding Problem

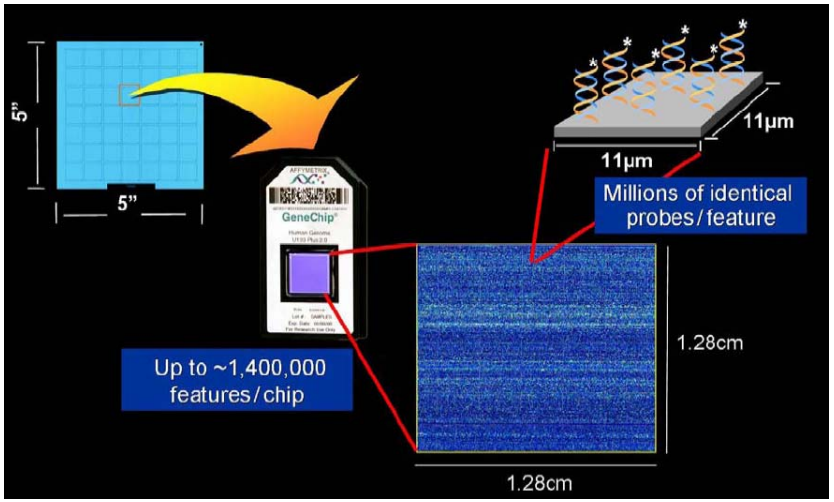
- Characterizing the motif: Positional weight matrix



- Finding the motif
  - Gibbs Sampler (AlignACE)
  - EM algorithm (MEME)

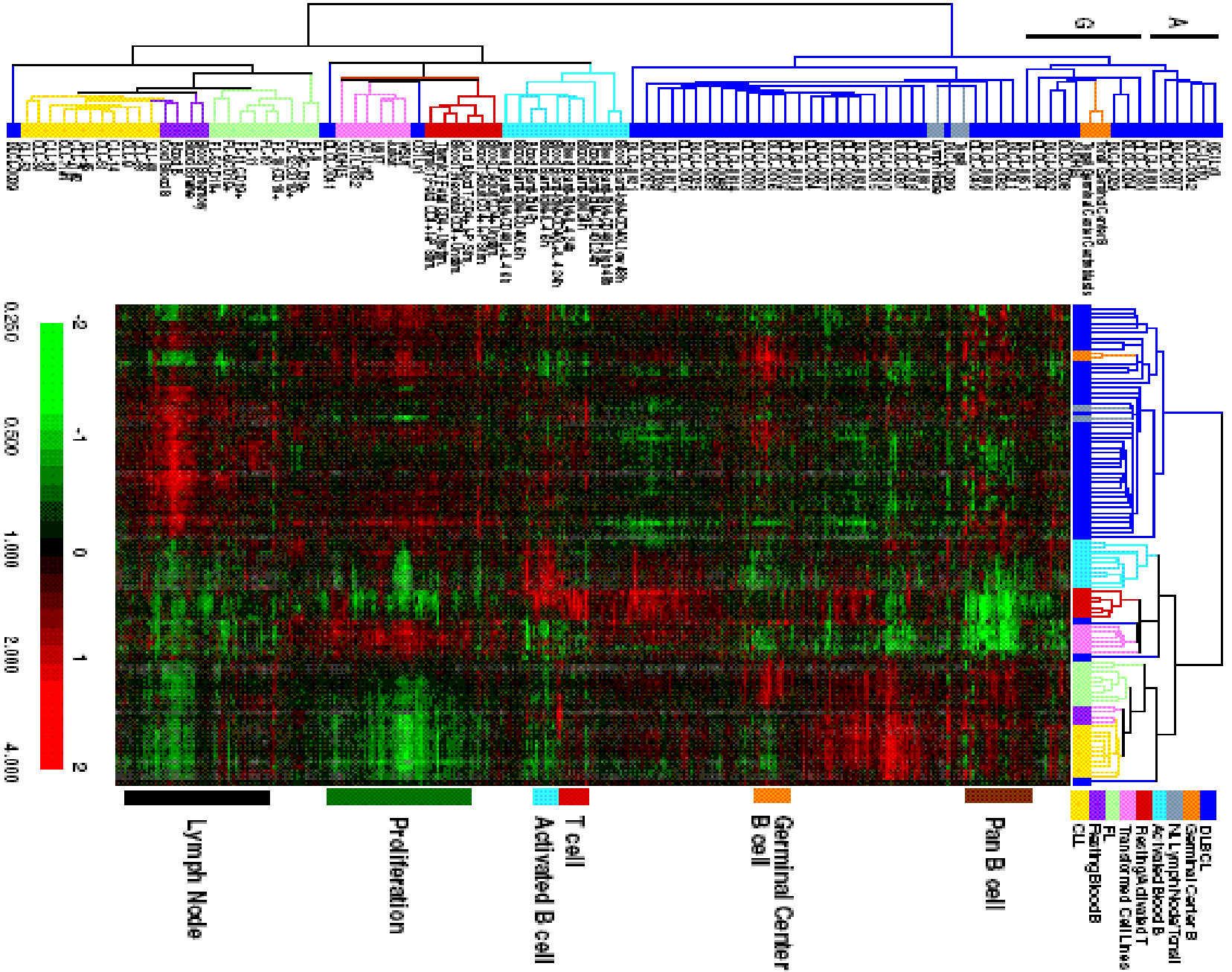
Topic V: Gene Expression Data Clustering and Biomarker Discovery





Microarrays

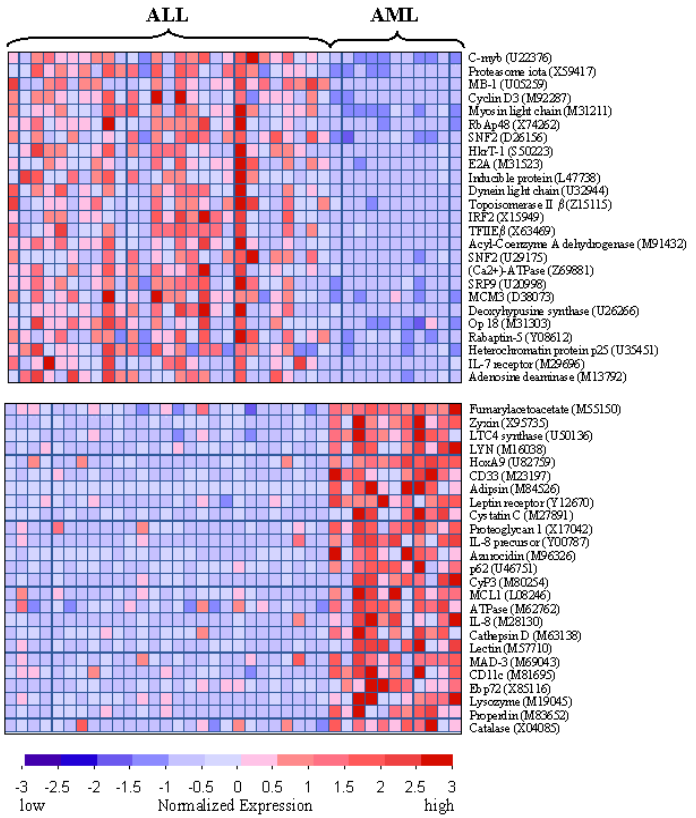
- ***DNA microarray*** technology rely on the hybridization properties of nucleic acids to monitor DNA or RNA abundance on a genomic scale in different types of cells.



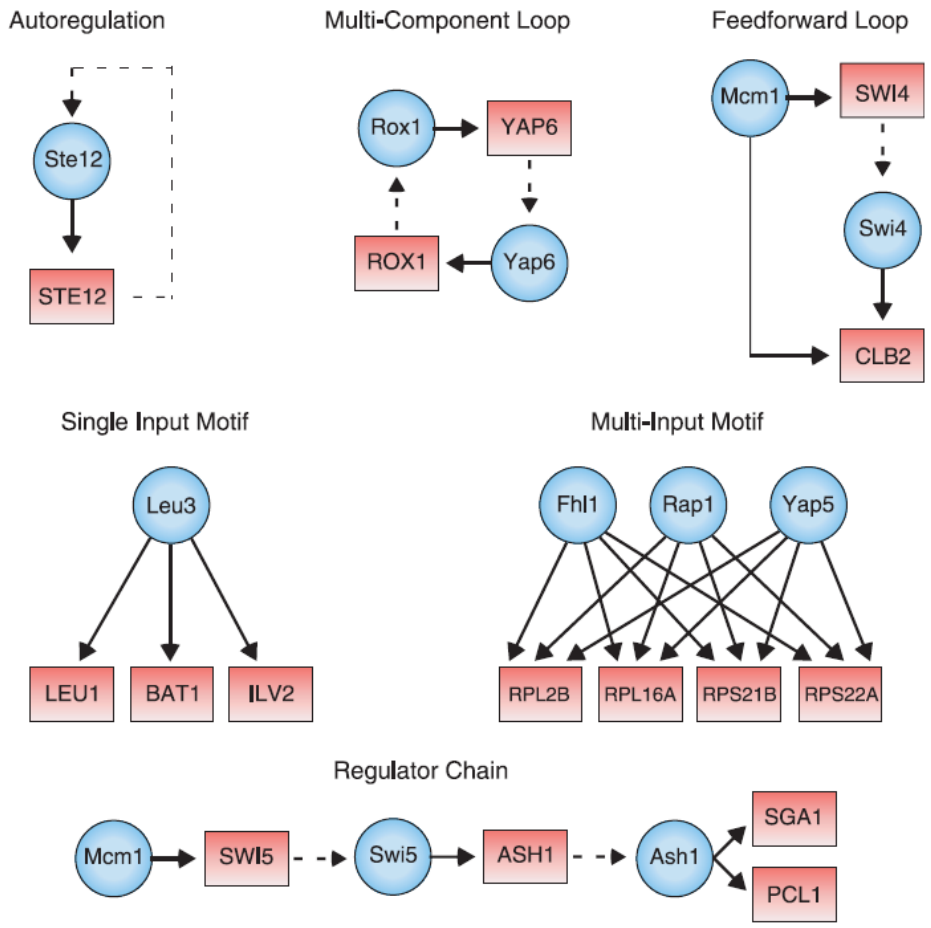
Alizadeh et al., Nature 403:503-11, 2000

Classification

- Given: the case, control gene expression data
- Find: a set of genes (biomarker) can discriminate two classes.
- Method: variable selection



Topic VI: Regulatory Network Inference from Gene Expression Data



Lee et al. Science 298: 799(2002)

Network Inference: Reverse Engineering

- Given: a large set of gene expression observations
- Find:
  - Wiring diagram
  - Transition rules

  To fit the observation data
- Methods
  - Bayesian Network
  - Gaussian graphical model

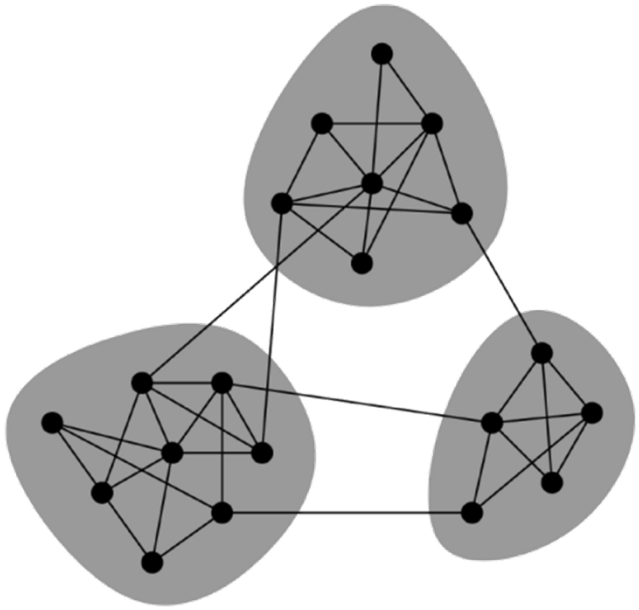
Dream Project

- DREAM: Dialogue for Reverse Engineering Assessments and Methods

- http://www.the-dream-project.org/

Topic VII: Network Analysis

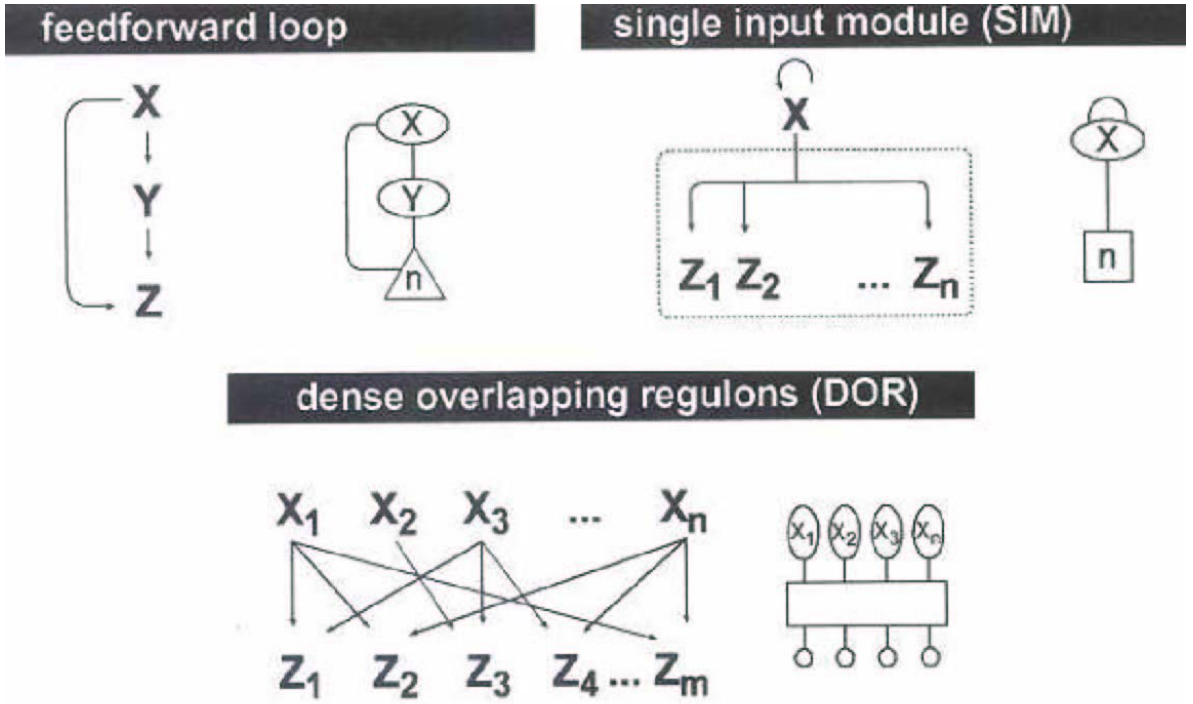
- Network modular (network clustering)



Network Motif

- Definition: Patterns of interconnections occurring in complex networks at numbers that are significantly higher than those in randomized networks (Milo, R., et. al. *Science* **298,** 824–827)

Network Motifs



Milo et al. *Science,* **298**, 2002

Topic VIII: Dimension Reduction

- Curse of dimensionality
- Visualization in low dimension

Curse of Dimensionality

- A major problem is *the curse of dimensionality.*
- If the data x lies in high dimensional space, then an enormous amount of data is required to learn distributions or decision rules.
- Example: 50 dimensions. Each dimension has 20 levels. This gives a total of 20^{50} cells. But the no. of data samples will be far less. There will not be enough data samples to learn.

Curse of Dimensionality

- One way to deal with dimensionality is to assume that we know the form of the probability distribution.
- For example, a Gaussian model in N dimensions has N + N(N-1)/2 parameters to estimate.
- Requires $O(N^2)$ data to learn reliably. This may be practical.

Dimension Reduction

- One way to avoid the curse of dimensionality is by projecting the data onto a lower-dimensional space.
- Techniques for dimension reduction:
  - Principal Component Analysis (PCA)
  - Singular value decomposition (SVD)
  - Multi-dimensional Scaling (MDS).

References

- James D. Watson, Tania A. Baker, Stephen P. Bell. Molecular Biology of the Gene. Benjamin-Cummings Publishing Company. 2008.
- Bruce Alberts. Molecular Biology of the Cell. Garland Publishing Inc. 2007.
- Jocelyn E. Krebs, Stephen T. Kilpatrick, Elliott S. Goldstein. Lewin‘s Genes XI. Jones and Bartlett Publishers, Inc. 2012.