Course Homepage

www.cs.fiu.edu/~giri/teach/BioinfS06.html

- Lecture notes, required reading material, homework, announcements, etc.
- Class: 2:00-4:45 PM, ECS 235.

Statistics Background

Basic Statistics

- Probability & Conditional Probability
- Probability space
- Random Variables
- Mean
- Standard Deviation
- Discrete vs Continuous
- Important results
  - Bayes theorem, Chebychev inequality, Markov's inequality, sum of expectations, Central Limit Theorem, …

Discrete Distributions

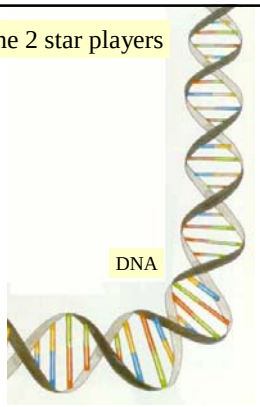
- Bernoulli
  - Distribution of heads/tails
- Binomial
  - Distr of # of heads in n coin tosses
- Geometric
  - Distr of # of tails before a head appears
- Negative binomial
  - Distr of # of tosses before r heads appear
- Poisson
  - Distr of # of heads in a large # of tosses

Continuous Distributions

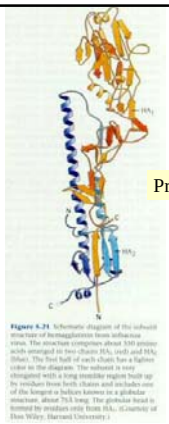
- Uniform
- Normal, Gaussian
- Exponential
- Gamma
- Beta
- Extreme-value

Molecular Biology Background

The 2 star players



DNA



Protein

The Players

DNA

String with alphabet {A, C, G, T}

Nucleotides/Bases

RNA

String with alphabet {A, C, G, U} **Bases**

Protein

String with 20-letter alphabet

Amino acids/Residues

Typical DNA Sequence

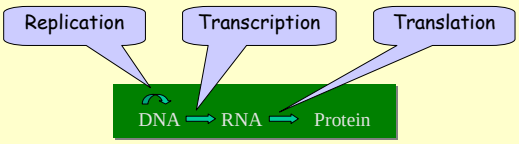
```
   1 gggagaacac ccggagaagg aggaggaggc gaagaaaagc aacagaagcc cagttgctgc
  61 tccaggtccc tcggacagag ctttttccat gtggagactc tctcaatgga cgtgcccctt
 121 agtgcttctt agacggactg cggtctccta aaggtcgacc atggtggccg ggacccgctg
 181 tcttctagtg ttgctgcttc cccaggtcct cctgggcggc gcggccggcc tcattccaga
 241 gctgggccgc aagaagttcg ccgcggcatc cagccgaccc ttgtcccggc cttcggaaga
 301 cgtcctcagc gaatttgagt tgaggctgct cagcatgttt ggcctgaagc agagacccac
 361 ccccagcaag gacgtcgtgg tgccccccta tatgctagat ctgtaccgca ggcactcagg
 421 ccagccagga gcgcccgccc cagaccaccg gctggagagg gcagccagcc gcgccaacac
 481 cgtgcgcagc ttccatcacg aagaagccgt ggaggaactt ccagagatga gtgggaaaac
 541 ggccccggcgc ttcttcttca atttaagttc tgtccccagt gacgagtttc tcacatctgc
 601 agaactccag atcttccggg aacagataca ggaagctttg ggaaacagta gtttccagca
 661 ccgaattaat atttatgaaa ttataaagcc tgcagcagcc aacttgaaat ttcctgtgac
 721 cagactattg gacaccaggt tagtgaatca gaacacaagt cagtgggaga gcttcgacgt
 781 caccccagct gtgatgcggt ggaccacaca gggacacacc aaccatgggt ttgtggtgga
 841 agtggcccat ttagaggaga acccaggtgt ctccaagaga catgtgagga ttagcaggtc
 901 tttgcaccaa gatgaacaca gctggtcaca gataaggcca ttgctagtga cttttggaca
 961 tgatggaaaa ggacatccgc tccacaaacg agaaaagcgt caagccaaac acaaacagcg
```

Typical protein sequence

```
/translation="MVAGTRCLLVLLLPQVLLGGAAGLIPELGRKKFAAASSRPLSRP
SEDVLSEFELRLLSMFGLKQRPTPSKDVVVPPYMLDLYRRHSGQPGAPAPDHRLERAA
SRANTVRSFHHEEAVEELPEMSGKTARRFFFNLSSVPSDEFLTSAELQIFREQIQEAL
GNSSFQHRINIYEIIKPAAANLKFPVTRLLDTRLVNQNTSQWESFDVTPAVMRWTTQG
HTNHGFVVEVAHLEENPGVSKRHVRISRSLHQDEHSWSQIRPLLVTFGHDGKGHPLHK
REKRQAKHKQRKRLKSSCKRHPLYVDFSDVGWNDWIVAPPGYHAFYCHGECPFPLADH
LNSTNHAIVQTLVNSVNSKIPKACCVPTELSAISMLYLDENEKVVLKNYQDMVVEGCG
CR"
```


Central Dogma

- DNA acts as a template to replicate itself.
- DNA is transcribed into RNA.
- RNA is translated into **Protein**.



Basic Genetic Processes

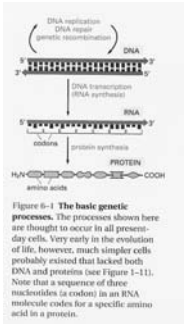
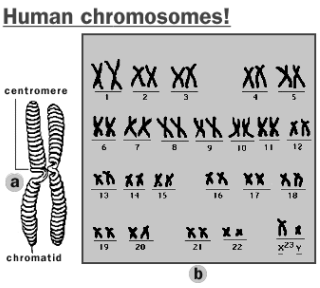


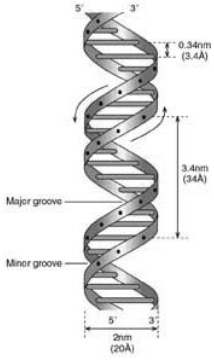
Figure 6–1 **The basic genetic processes.** The processes shown here are thought to occur in all present-day cells. Very early in the evolution of life, however, much simpler cells probably existed that lacked both DNA and proteins (see Figure 1–11). Note that a sequence of three nucleotides (a codon) in an RNA molecule codes for a specific amino acid in a protein.

Chromosomes

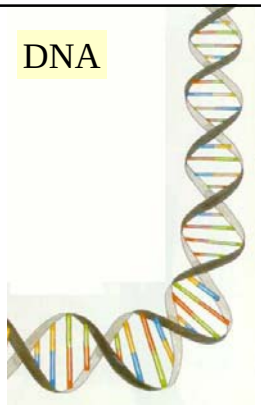


Chromosomes

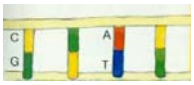
DNA Molecule



DNA



Complementary Bases

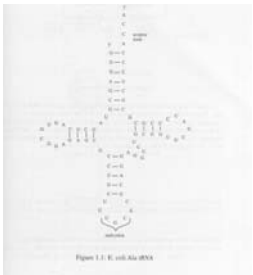


Proteins – Amino acids

amino acid	3 letter code	1 letter code
alanine	Ala	A
arginine	Arg	R
aspartic acid	Asp	D
asparginine	Asn	N
cysteine	Cys	C
glutamic acid	Glu	E
glutamine	Gln	Q
glycine	Gly	G
histine	His	H
isoleucine	Ile	I
leucine	Leu	L
lysine	Lys	K
methionine	Met	M
phenylalanine	Phe	F
proline	Pro	P
serine	Ser	S
threonine	Thr	T
tryptophan	Trp	W
tyrosine	Tyr	Y
valine	Val	V

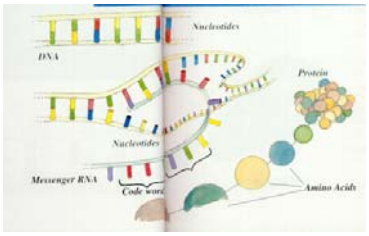
Table 1.1: *Amino acid abbreviations*

RNA

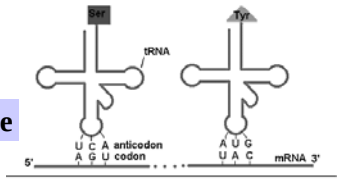


Genes

DNA ⟶ RNA ⟶ Protein



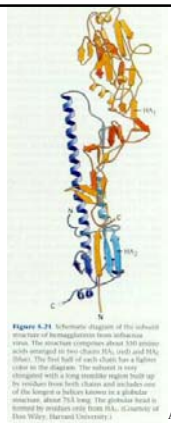
The Genetic Code

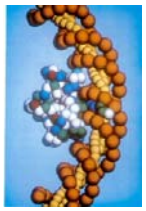


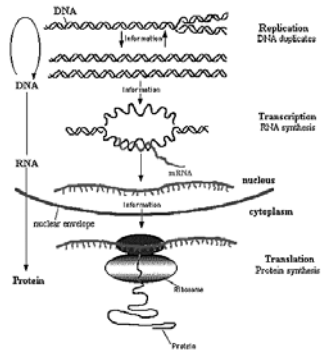
2nd base in codon

1st base in codon	U	C	A	G	3rd base in codon
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STOP	STOP	A
	Leu	Ser	STOP	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

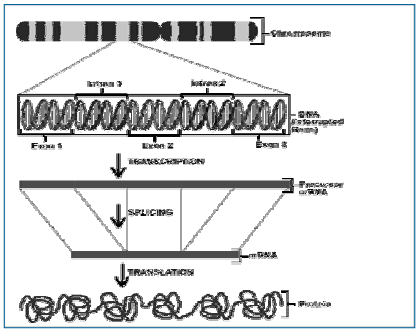
The Genetic Code



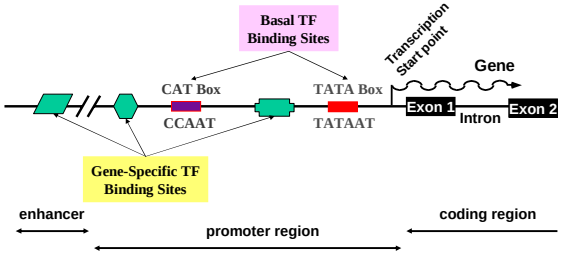




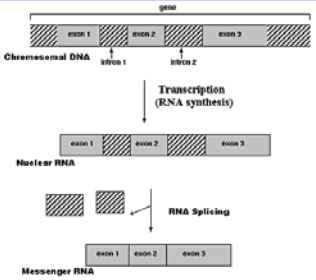
The Central Dogma of Molecular Biology



Transcription Regulation

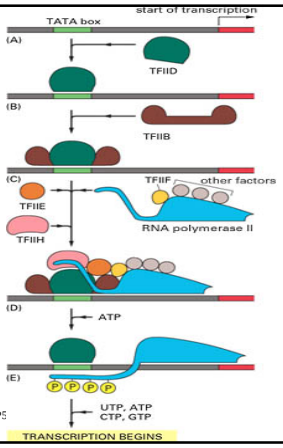


DNA Transcription

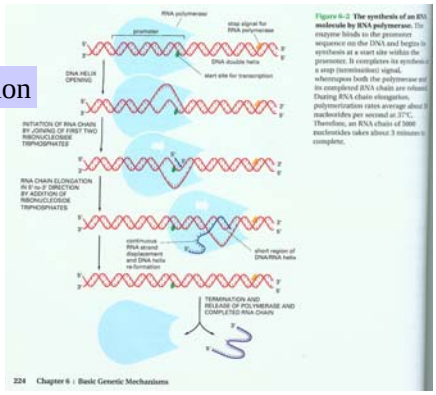


RNA synthesis and processing

Transcription Initiation



Transcription



Transcription Steps

RNA polymerase needs many transcription factors (TFIIA,TFIIB, etc.)

(A) The promoter sequence (TATA box) is located 25 nucleotides away from transcription initiation site.

(B) The TATA box is recognized and bound by transcription factor TFIID, which then enables the adjacent binding of TFIIB. DNA is somewhat distorted in the process.

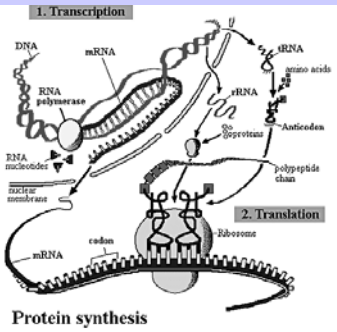
(D) The rest of the general transcription factors as well as the RNA polymerase itself assemble at the promoter. What order?

(E) TFIIH then uses ATP to phosphorylate RNA polymerase II, changing its conformation so that the polymerase is released from the complex and is able to start transcribing. As shown, the site of phosphorylation is a long polypeptide tail that extends from the polymerase molecule.

Transcription Factors

- The general transcription factors have been highly conserved in evolution; some of those from human cells can be replaced in biochemical experiments by the corresponding factors from simple yeasts.

Protein Synthesis



Protein synthesis

Protein Synthesis:

Incorporation of amino acid into protein