10-810
Advanced Algorithms and Models for Computational Biology

Ziv Bar-Joseph
zivbj@cs.cmu.edu
WeH 4107

Eric Xing
epxing@cs.cmu.edu
WeH 4127

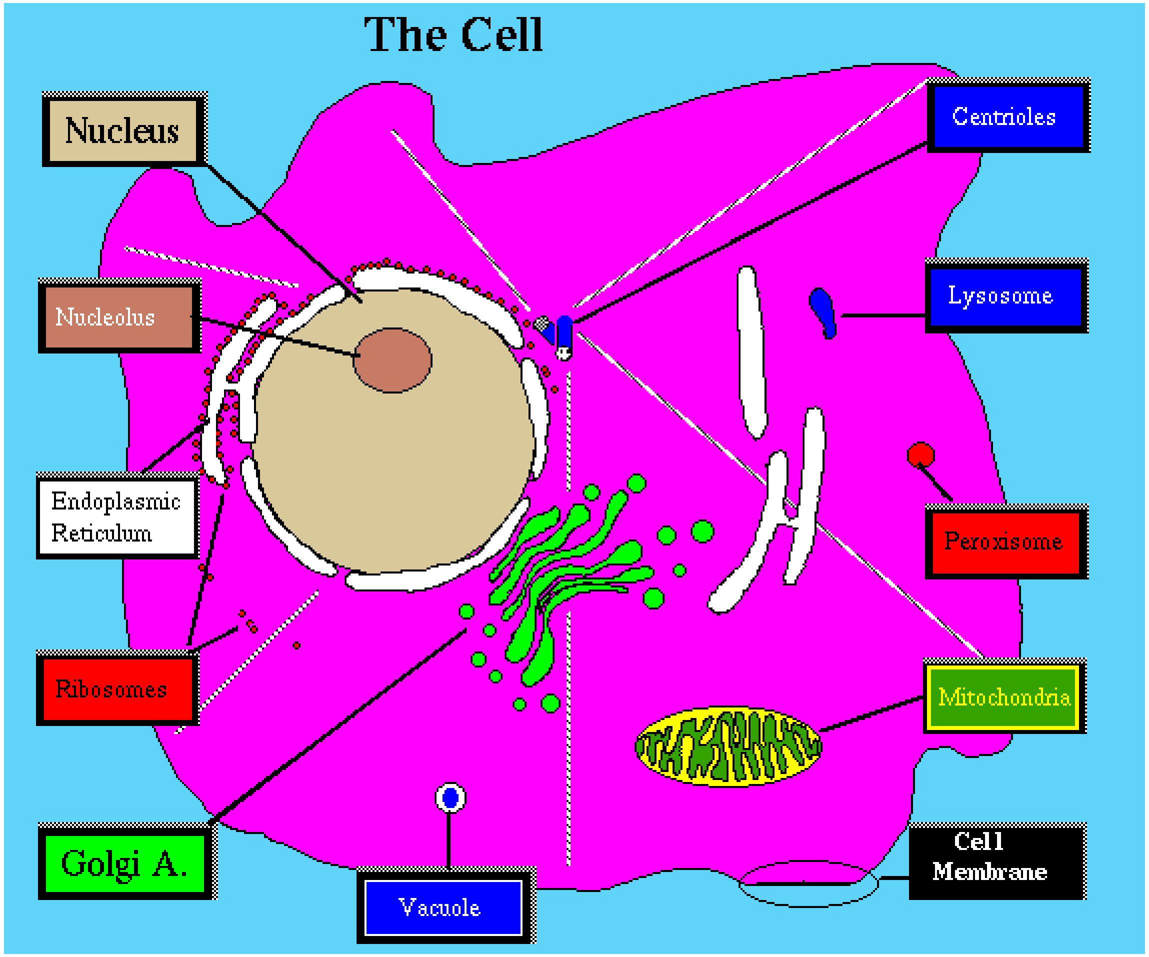
http://www.cs.cmu.edu/~epxing/Class/10810-06/

Topics

- Introduction (1 Week)
- Sequence analysis (3 weeks)
- Gene expression (4 weeks)
- Genetics (3 weeks)
- Systems biology (3 weeks)
- Projects (1 week)

Introduction to Molecular Biology

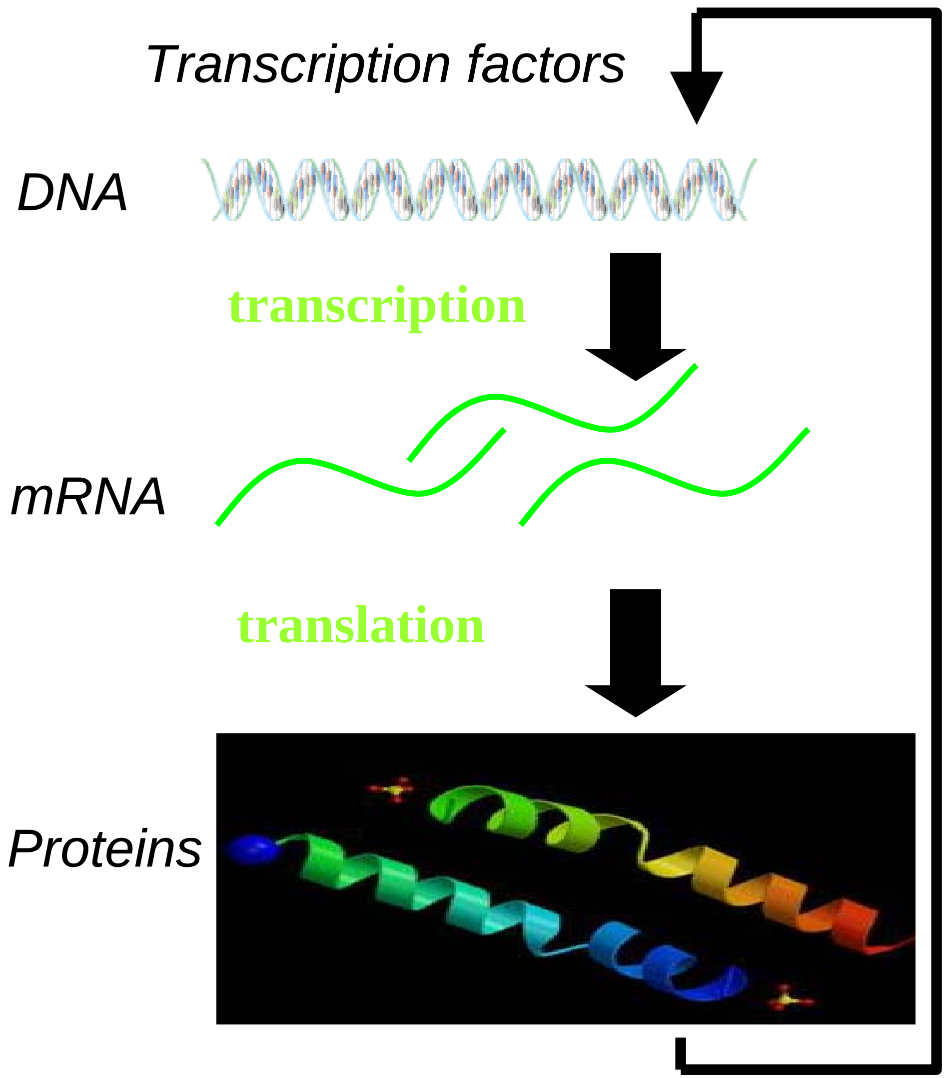
The Eukaryotic Cell



Cells Type

- Eukaryots:
  - Plants, animals, humans
  - DNA resides in the nucleus
  - Contain also other compartments
- Prokaryots:
  - Bacteria
  - Do not contain compartments

Central Dogma



The Genome

Genome

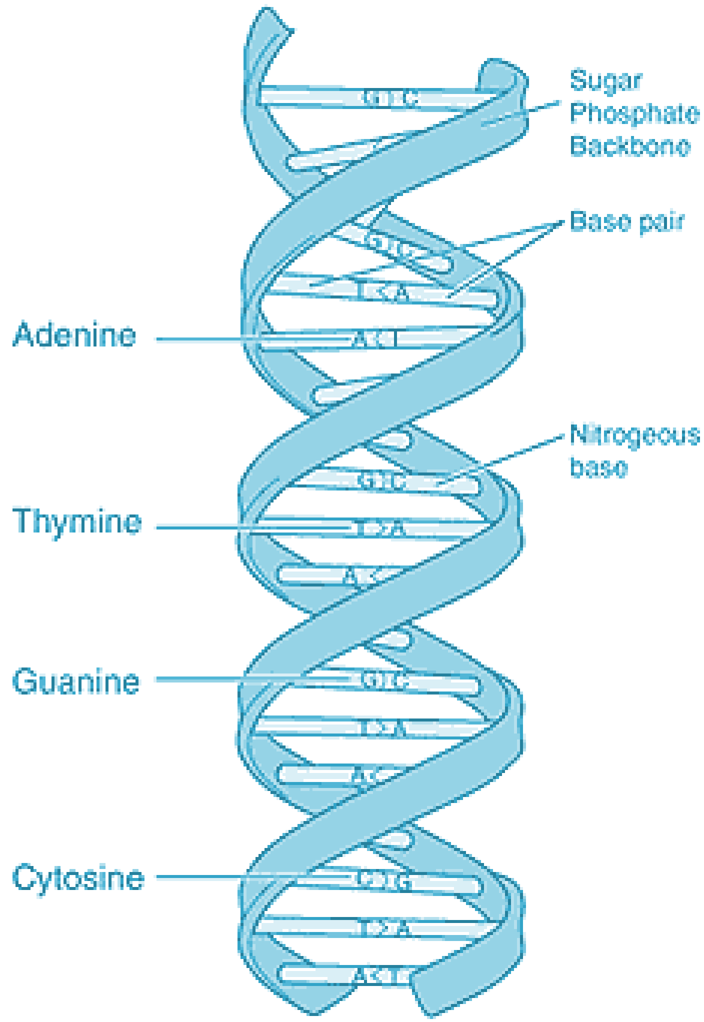
- A genome is an organism’s complete set of DNA (including its genes).
- However, in humans less than 3% of the genome actually encodes for genes.
- A part of the rest of the genome serves as a control regions (though that’s also a small part).
- The goal of the rest of the genome is unknown (a possible project …).

DNA

- Four letters alphabet: A, C, G and T.
- Complementary base pairing: A-T and G-C
- Double stranded.
- Encodes proteins in segments called genes.
- Control regions in front of genes.
- Similar in all cells in the body.

DNA

- Four letters alphabet: A, C, G and T.
- Complementary base pairing: A-T and G-C
- Double stranded.
- Encodes proteins in segments called genes.
- Control regions in front of genes.
- Similar in MOST cells in the body.


Sugar
Phosphate
Backbone
Base pair
Adenine
Nitrogeous
base
Thymine
Guanine
Cytosine

Comparison of Different Organisms

	Genome size	Num. of genes
E. coli	$.05*10^8$	4,200
Yeast	$.15*10^8$	6,000
Worm	$1*10^8$	18,400
Fly	$1.8*10^8$	13,600
Human	$30*10^8$	30,000
Plant	$1.3*10^8$	25,000

DNA is Divided Between Chromosomes

- Single molecule of DNA
- 46 chromosomes in human cells
- 22 pairs, and two sex chromosomes: X and Y
- For different diseases, many diseases related genes reside on the same chromosome

Genes

What is a gene?

Promoter Protein coding sequence Terminator

Genomic DNA

Example of a Gene: Gal4 DNA

ATGAAGCTACTGTCTTCTATCGAACAAGCATGCGATATTTGCCGACTTAAAAAGCTCAAG
TGCTCCAAAGAAAAACCGAAGTGCGCCAAGTGTCTGAAGAACAACTGGGAGTGTCGCTAC
TCTCCCAAAACCAAAAGGTCTCCGCTGACTAGGGCACATCTGACAGAAGTGGAATCAAGG
CTAGAAAGACTGGAACAGCTATTTCTACTGATTTTTCCTCGAGAAGACCTTGACATGATT
TTGAAAATGGATTCTTTACAGGATATAAAAGCATTGTTAACAGGATTATTTGTACAAGAT
AATGTGAATAAAGATGCCGTCACAGATAGATTGGCTTCAGTGGAGACTGATATGCCTCTA
ACATTGAGACAGCATAGAATAAGTGCGACATCATCATCGGAAGAGAGTAGTAACAAAGGT
CAAAGACAGTTGACTGTATCGATTGACTCGGCAGCTCATCATGATAACTCCACAATTCCG
TTGGATTTTATGCCCAGGGATGCTCTTCATGGATTTGATTGGTCTGAAGAGGATGACATG
TCGGATGGCTTGCCCTTCCTGAAAACGGACCCCAACAATAATGGGTTCTTTGGCGACGGT
TCTCTCTTATGTATTCTTCGATCTATTGGCTTTAAACCGGAAAATTACACGAACTCTAAC
GTTAACAGGCTCCCGACCATGATTACGGATAGATACACGTTGGCTTCTAGATCCACAACA
TCCCGTTTACTTCAAAGTTATCTCAATAATTTTCACCCCTACTGCCCTATCGTGCACTCA
CCGACGCTAATGATGTTGTATAATAACCAGATTGAAATCGCGTCGAAGGATCAATGGCAA
ATCCTTTTTAACTGCATATTAGCCATTGGAGCCTGGTGTATAGAGGGGGAATCTACTGAT
ATAGATGTTTTTTACTATCAAAATGCTAAATCTCATTTGACGAGCAAGGTCTTCGAGTCA

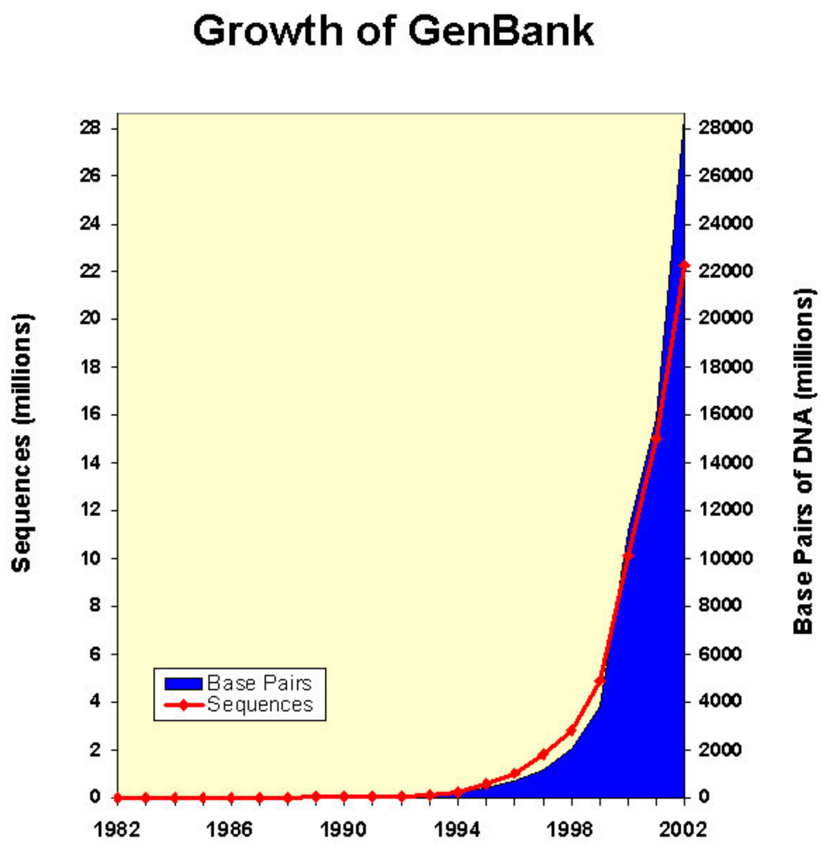
Genes Encode for Proteins

1st letter	Second Letter: U		C		A		G		3rd letter
U	UUU UUC UUA UUG	Phe Leu	UCU UCC UCA UCG	Ser	UAU UAC UAA UAG	Tyr Stop Stop	UGU UGC UGA UGG	Cys Stop Trp	U C A G
C	CUU CUC CUA CUG	Leu	CCU CCC CCA CCG	Pro	CAU CAC CAA CAG	His Gln	CGU CGC CGA CGG	Arg	U C A G
A	AUU AUC AUA AUG	Ile Met	ACU ACC ACA ACG	Thr	AAU AAC AAA AAG	Asn Lys	AGU AGC AGA AGG	Ser Arg	U C A G
G	GUU GUC GUA GUG	Val	GCU GCC GCA GCG	Ala	GAU GAC GAA GAG	Asp Glu	GGU GGC GGA GGG	Gly	U C A G

Example of a Gene: Gal4 AA

MKLLSSIEQACDICRLKKLKCSKEKPKCAKCLKNNWECRYSPKTKRSPLTRAHLTEVESR
LERLEQLFLLIFPREDLDMILKMDSLQDIKALLTGLFVQDNVNKDAVTDRLASVETDMPL
TLRQHRISATSSSEESSNKGQRQLTVSIDSAAHHDNSTIPLDFMPRDALHGFDWSEEDDM
SDGLPFLKTDPNNNGFFGDGSLLCILRSIGFKPENYTNSNVNRLPTMITDRYTLASRSTT
SRLLQSYLNNFHPYCPIVHSPTLMMLYNNQIEIASKDQWQILFNCILAIGAWCIEGESTD
IDVFYYQNAKSHLTSKVFESGSIILVTALHLLSRYTQWRQKTNTSYNFHSFSIRMAISLG
LNRDLPSSFSDSSILEQRRRIWWSVYSWEIQLSLLYGRSIQLSQNTISFPSSVDDVQRTT
TGPTIYHGIIETARLLQVFTKIYELDKTVTAEKSPICAKKCLMICNEIEEVSRQAPKFLQ
MDISTTALTNLLKEHPWLSFTRFELKWKQLSLIIYVLRDFFTNFTQKKSQLEQDQNDHQS
YEVKRCSIMLSDAAQRTVMSVSSYMDNHNVTPYFAWNCSYYLFNAVLVPIKTLLSNSKSN
AENNETAQLLQQINTVLMLLKKLATFKIQTCEKYIQVLEEVCAPFLLSQCAIPLPHISYN
NSNGSAIKNIVGSATIAQYPTLPEENVNNISVKYVSPGSVGPSPVPLKSGASFSDLVKLL
SNRPPSRNSPVTIPRSTPSHRSVTPFLGQQQQLQSLVPLTPSALFGGANFNQSGNIADSS

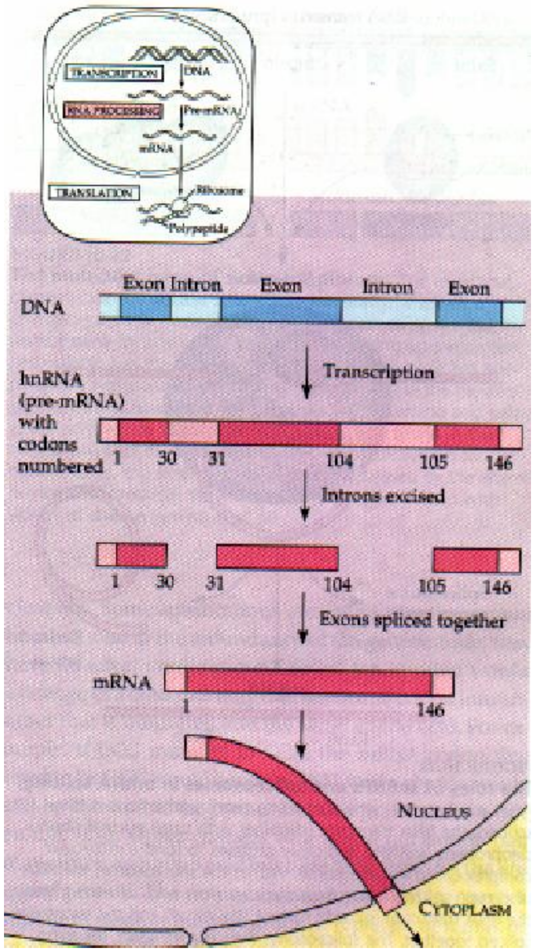
Number of Genes in Public Databases



Structure of Genes in Mammalian Cells

- Within coding DNA genes there can be “junk” regions (Introns)
- Exons are segments of DNA that contain the gene’s information coding for a protein
- Need to cut Introns out of RNA and splice together Exons before protein can be made

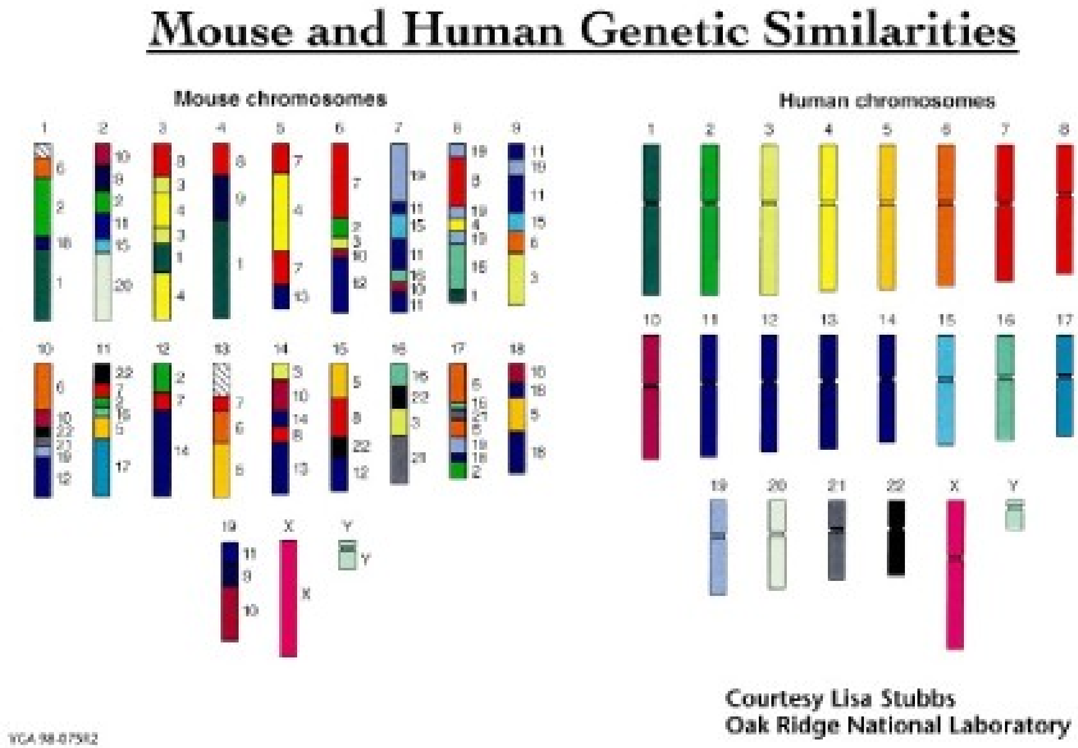
- Alternative splicing increases the potential number of different proteins, allowing the generation of millions of proteins from a small number of genes.


TRANSCRIPTION
DNA
RNA PROCESSING
Pre-mRNA
mRNA
TRANSLATION
Ribosome
Polypeptide
Exon
Intron
Exon
Intron
Exon
DNA
Transcription
hnRNA (pre-mRNA) with codons numbered
1
30
31
104
105
146
Introns excised
1
30
31
104
105
146
Exons spliced together
mRNA
1
146
NUCLEUS
CYTOPLASM

Identifying Genes in Sequence Data

- Predicting the start and end of genes as well as the introns and exons in each gene is one of the basic problems in computational biology.
- Gene prediction methods look for *ORFs* (Open Reading Frame).
- These are (relatively long) DNA segments that start with the start codon, end with one of the end codons, and do not contain any other end codon in between.
- Splice site prediction has received a lot of attention in the literature.

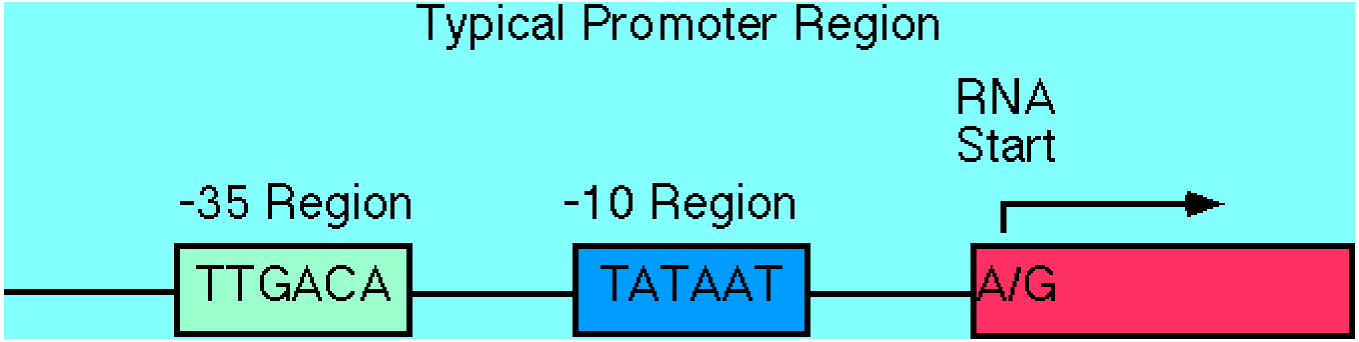
Comparative genomics



Regulatory Regions

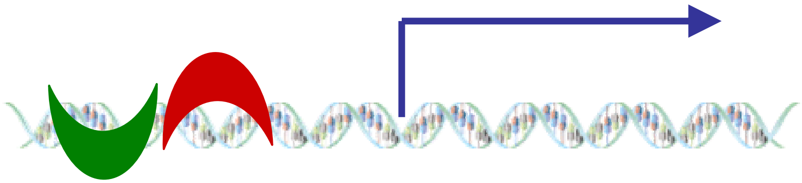
Promoter

The promoter is the place where RNA polymerase binds to start transcription. This is what determines which strand is the coding strand.

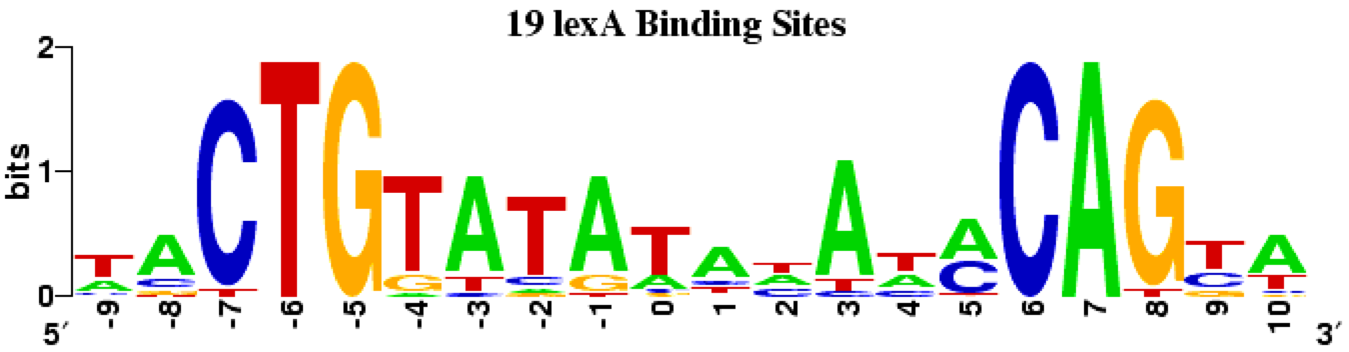


DNA Binding Motifs

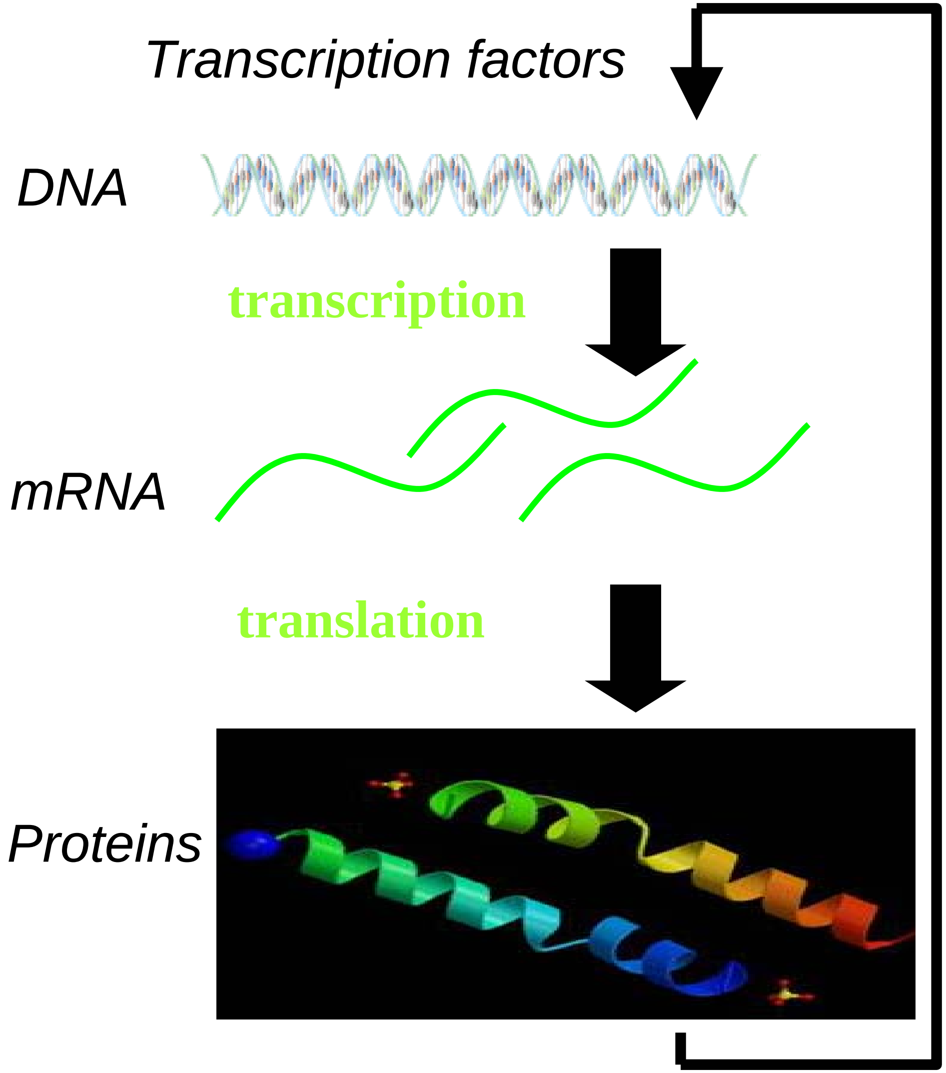
- In order to recruit the transcriptional machinery, a transcription factor (TF) needs to bind the DNA in front of the gene.
- TFs bind in to short segments which are known as DNA binding motifs.
- Usually consists 6 – 8 letters, and in many cases these letters generate palindromes.



Example of Motifs



Central Dogma



RNA

Three major types (plus a recently discovered regulatory RNA).

- mRNA – messenger RNA
- tRNA – Transfer RNA
- rRNA – ribosomal RNA
- RNAi – RNA interference

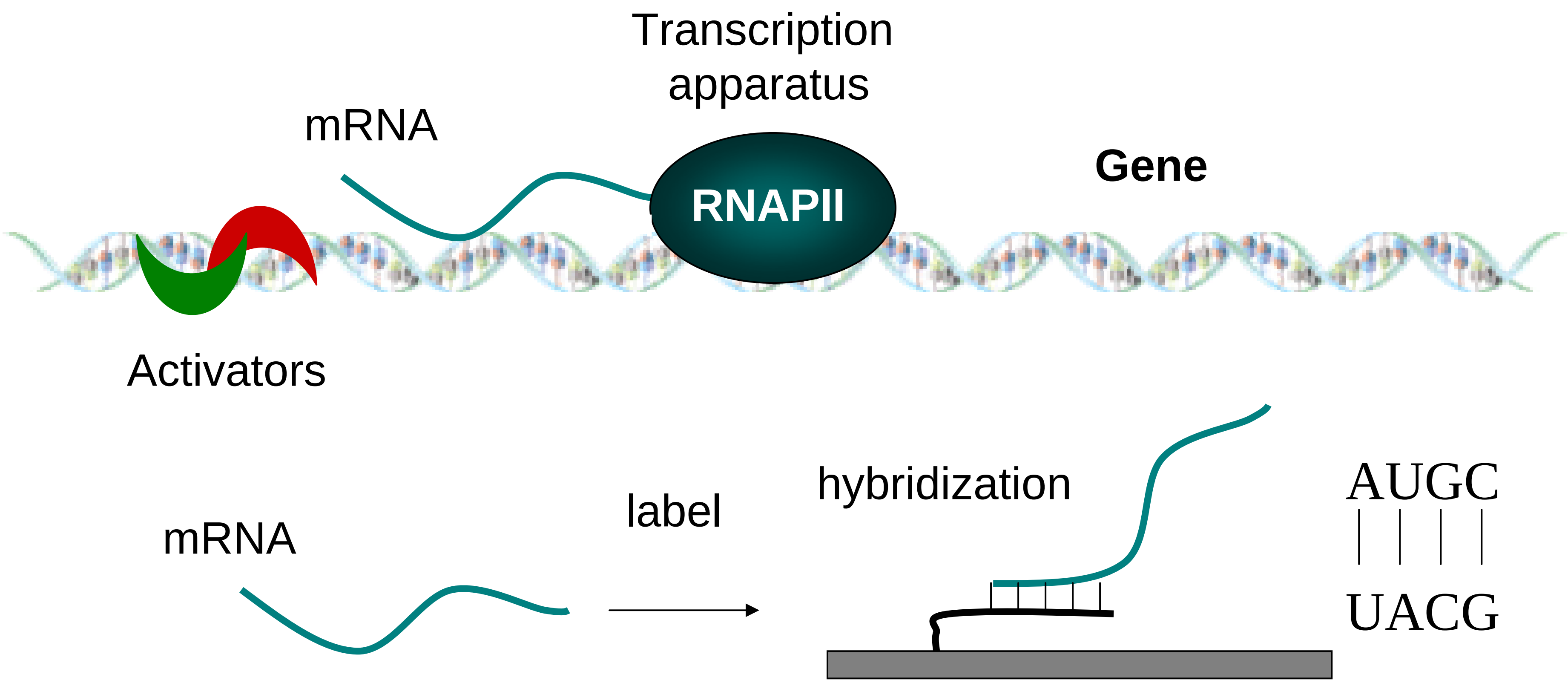
Messenger RNA

- Basically, an intermediate product
- Transcribed from the genome and translated into protein
- Number of copies correlates well with number of proteins for the gene.
- Unlike DNA, the amount of messenger RNA (as well as the number of proteins) differs between different cell types and under different conditions.

Complementary base-pairing

• mRNA is transcribed from the DNA

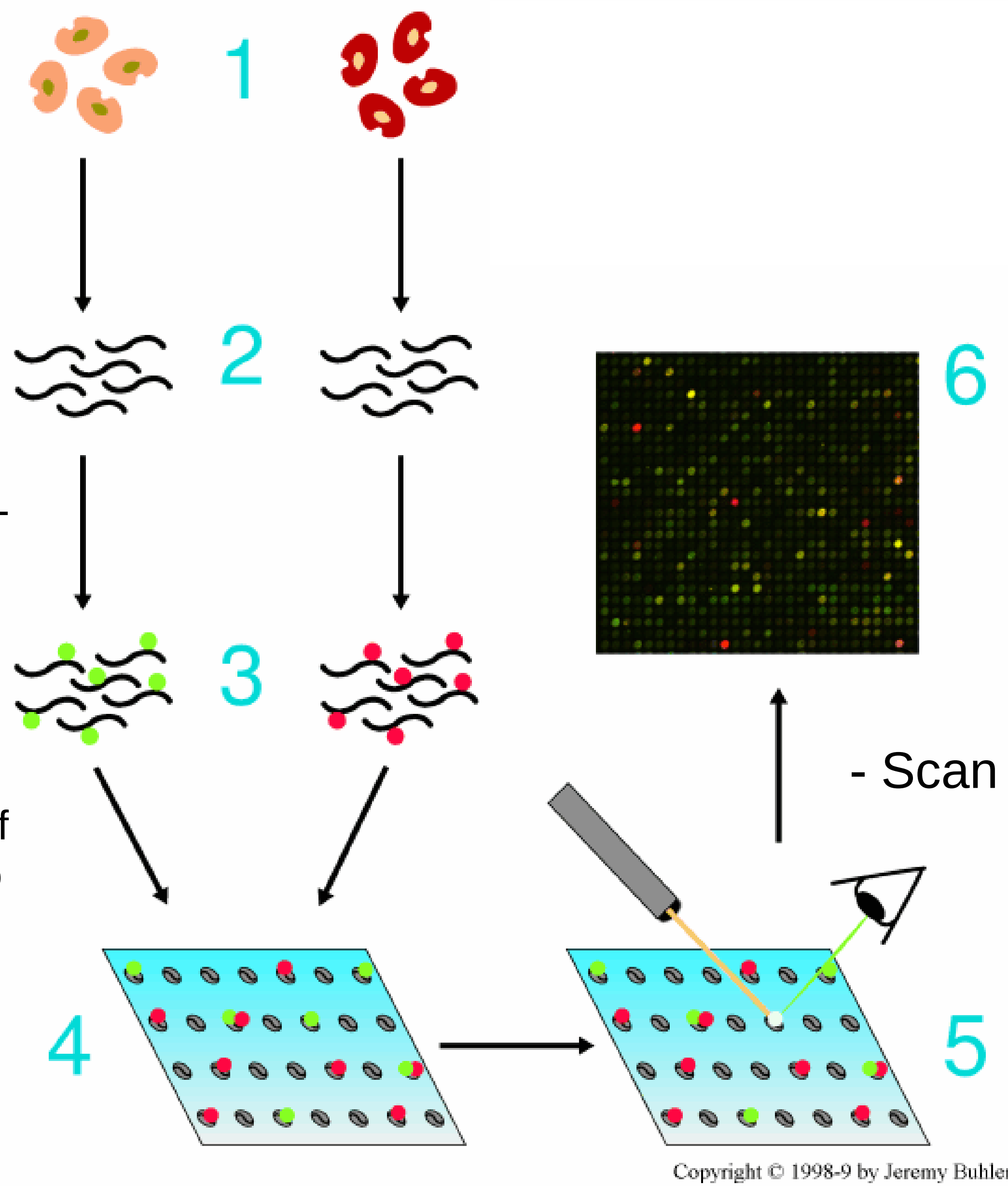
• mRNA (like DNA, but unlike proteins) binds to its complement



Hybridization and Scanning—Glass slide arrays

- Prepare Cy3, Cy5-labeled ss cDNA

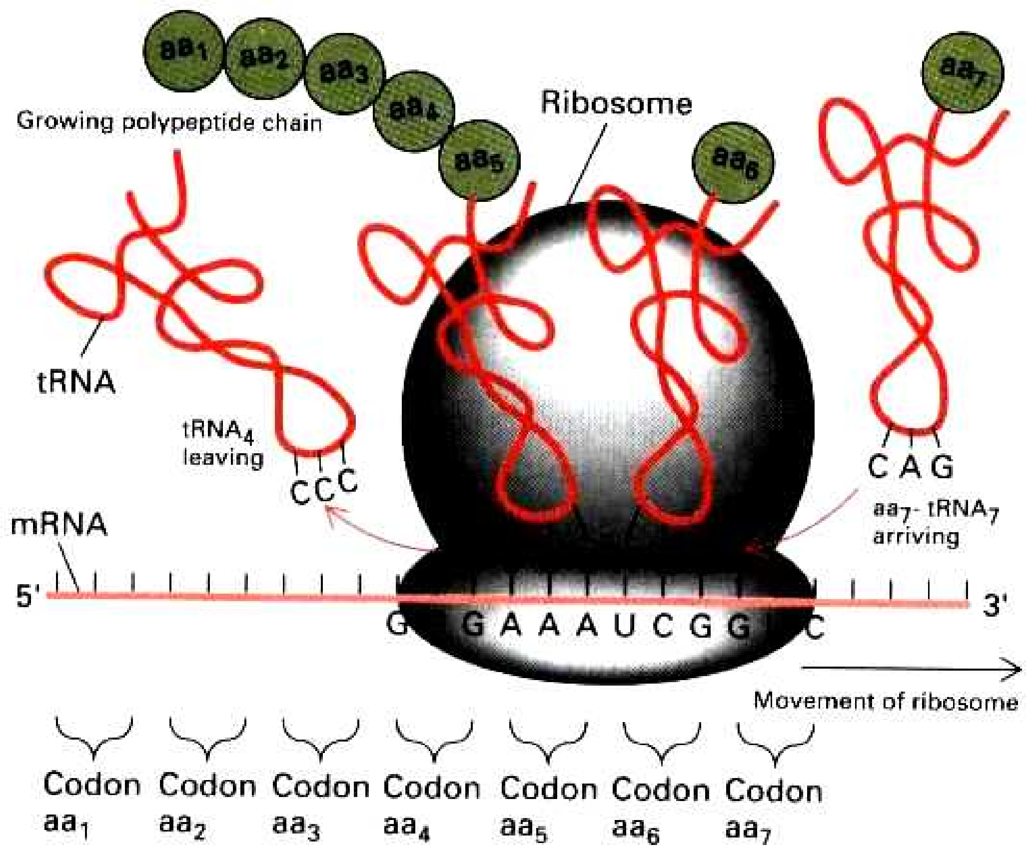
- Hybridize 600 ng of labeled ss cDNA to glass slide array



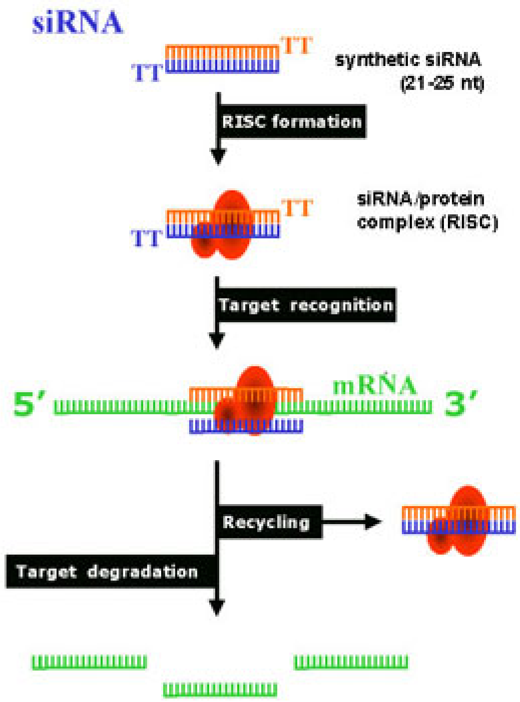
The Ribosome

- Decoding machine.
- Input: mRNA, output: protein
- Built from a large number of proteins and a number of RNAs.
- Several ribosomes can work on one mRNA

The Ribosome

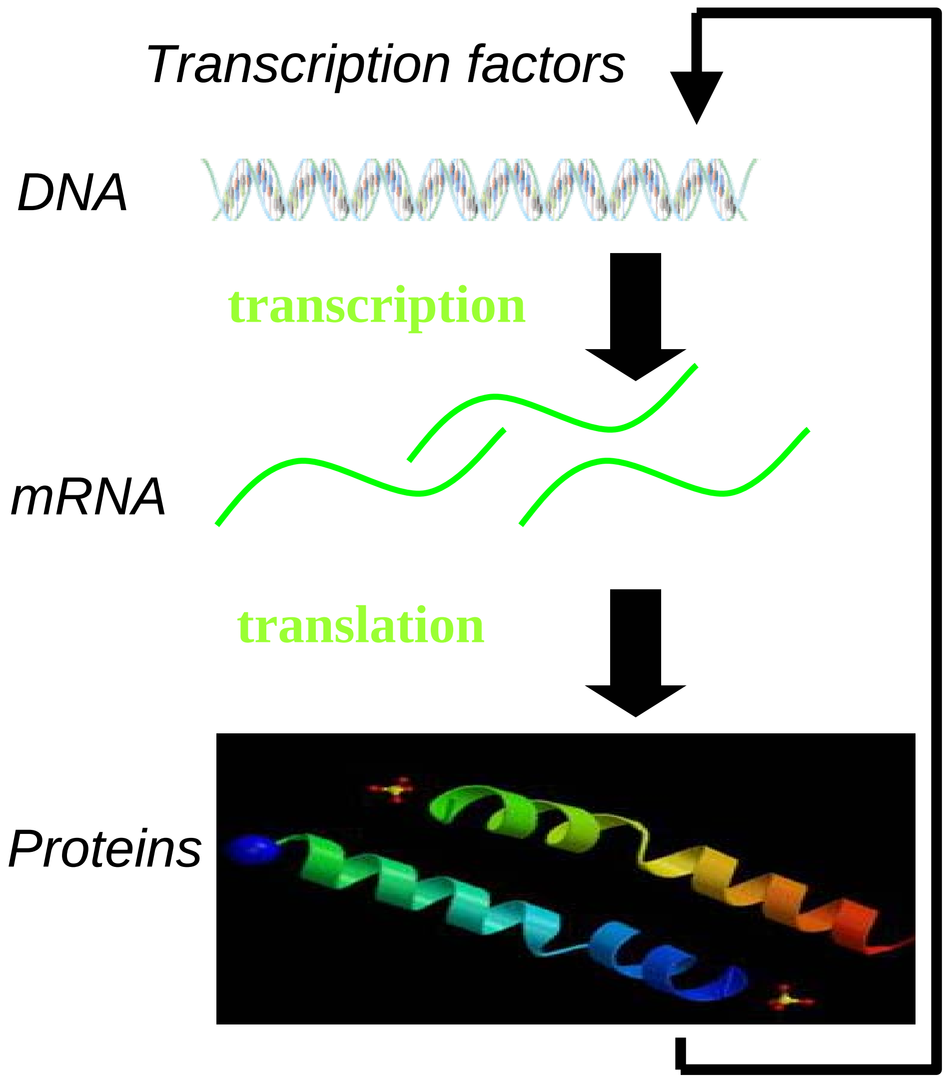


RNAi



siRNA
TT
TT
synthetic siRNA
(21-25 nt)
RISC formation
TT
TT
siRNA/protein
complex (RISC)
Target recognition
5′
mRNA
3′
Recycling
Target degradation

Central Dogma



Proteins

- Proteins are polypeptide chains of amino acids.
- Four levels of structure:

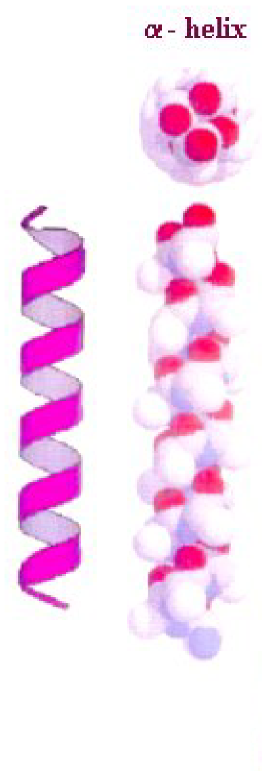
  - Primary Structure: The sequence of the protein

  - Secondary structure: Local structure in regions of the chain

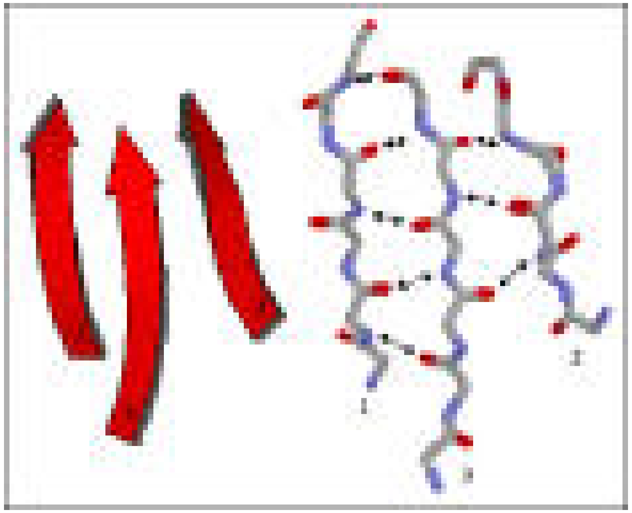
  - Tertiary Structure: Three dimensional structure

  - Quaternary Structure: multiple subunits

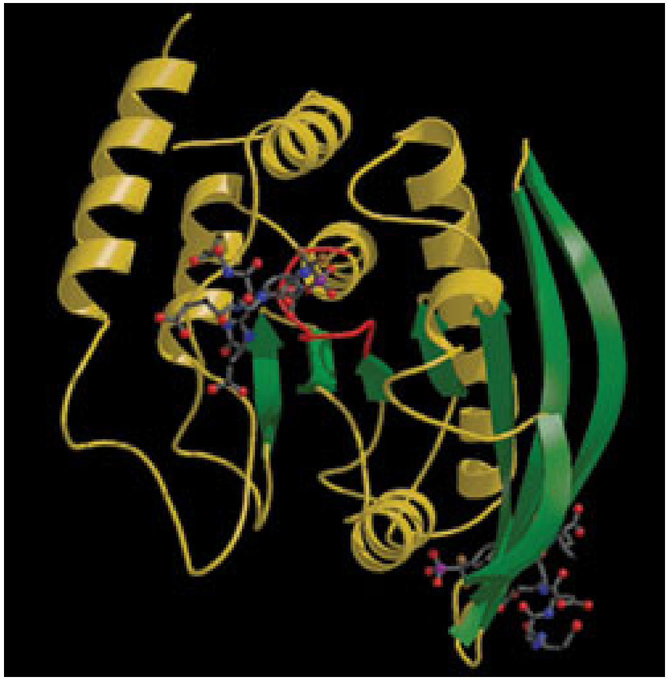
Secondary Structure: Alpha Helix



Secondary Structure: Beta Sheet



Protein Structure



Domains of a Protein

- While predicting the structure from the sequence is still an open problem, we can identify several domains within the protein.
- Domains are compactly folded structures.
- In many cases these domains are associated with specific biological function.

Assigning Function to Proteins

- While almost 30000 genes have been identified in the human genome, relatively few have known functional annotation.
- Determining the function of the protein can be done in several ways.

  - Sequence similarity to other (known) proteins
  - Using domain information
  - Using three dimensional structure
  - Based on high throughput experiments (when does it functions and who it interacts with)

Protein Interaction

In order to fulfill their function, proteins interact with other proteins in a number of ways including:

- Regulation
- Pathways, for example A -> B -> C
- Post translational modifications
- Forming protein complexes

Determining how are these protein interact, and what are the resulting pathways and networks is one of the major focuses of this class.

What you should remember

- Higher level organization of the cell
- The central dogma
- Composition of the DNA: Genes, control regions
- Properties of mRNA
- Structure of proteins