Pattern Recognition in Bioinformatics

Pattern Recognition in Bioinformatics

Lecturers

Christian Nørgaard Storm Pedersen
E-mail: cstorm@birc.au.dk
Office: 1110.325

Thomas Mailund
E-mail: mailund@birc.au.dk
Office: 1110.326

Course homepage

http://www.birc.au.dk/~cstorm/courses/PRiB_f12

Lectures

Mondays 10.15-12.00 and Wednesdays 15.15-16.00

in DI-Nygaard 184

Info about week *x* available on Thursday in week *x*-1 at

Week 1: Hidden Markov Models (HMMs)

Week 2: HMMs

Week 3: HMMs

Week 4: Stochastic Context Free Grammars (SCFGs)

Week 5: SCFGs

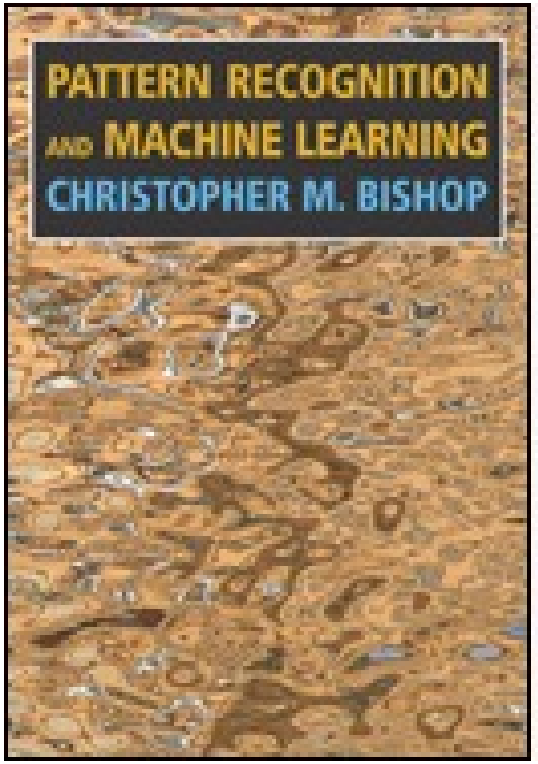
Week 6: Pattern Discovery.

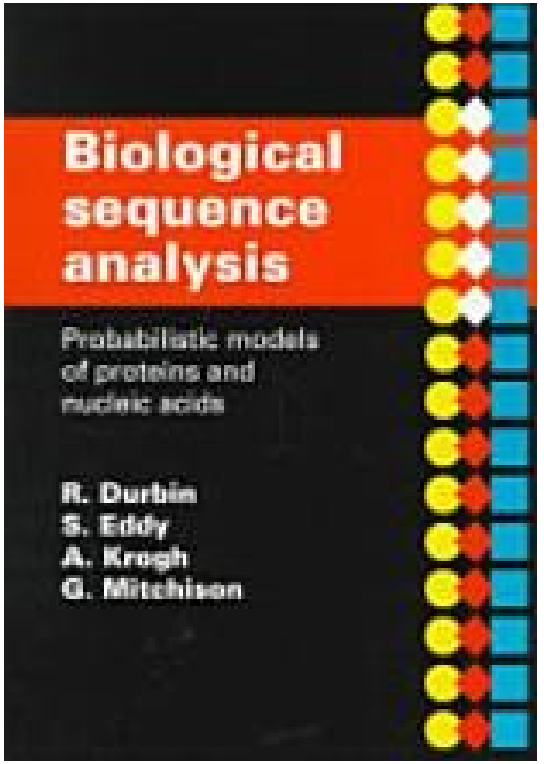
Week 7: Pattern Discovery.

http://www.birc.au.dk/~cstorm/courses/PRiB_f12/schedule.html

Literature

Book chapters and papers will be available via WWW





Mandatory Projects and Exam

There are three mandatory projects.

Basic HMM algorithms (30/1 – 8/2)

An HMM for gene finding (8/2 – 22/2)

An SCFG for RNA structure prediction (22/2 – 7/3)

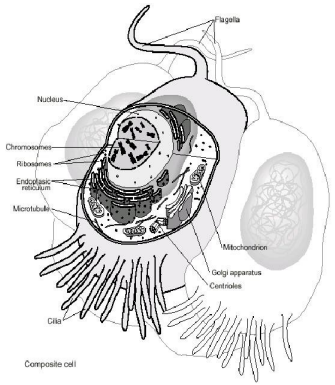
Work in groups of 2-3 students

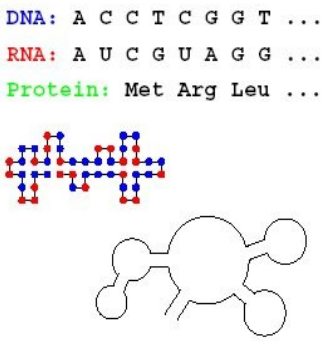
Implementation, experiments, small report, discussion

The exam is an individual 20 minute oral exam which includes a presentation of a topic covered in the class

What is bioinformatics?

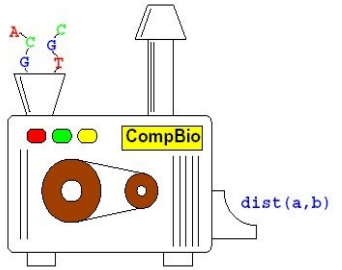
Construction and application af algorithms and programs for collecting, handling, and analysis of biological data ...






```
Input:  a[1..n],b[1..m]
Output: dist(a,b)

D[0,0..m]=D[0..n,0]=0
FOR i=1 TO n DO
  FOR j=1 TO m DO
    D[i,j]=
    min(D[i-1,j-1]+
    d(a[i],b[j]),
    D[i-1,j]+1,
    D[i,j-1]+1)
  OD
OD
RETURN D[n,m]
```



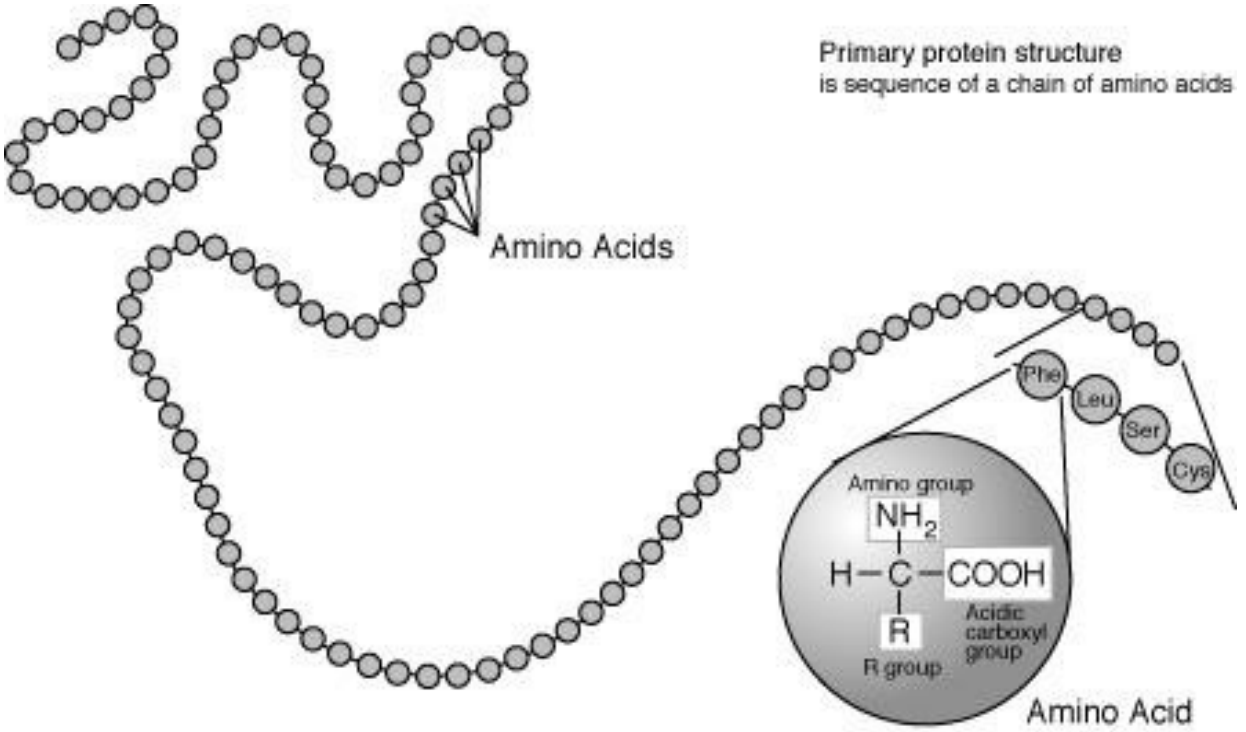
Biological questions, models of biology, formulation of computational problems, construction of effective algorithms, development of useful tools ...

Finding patterns in strings

```
>NC_002737.1 Streptococcus pyogenes M1 GAS, complete genome.
TTGTTGATATTCTGTTTTTTCTTTTTTAGTTTTCCACATGAAAAATAGTTGAAAACAATA
GCGGTGTCCCCTTAAAATGGCTTTTCCACAGGTTGTGGAGAACCCAAATTAACAGTGTTA
ATTTATTTTCCACAGGTTGTGGAAAAACTAACTATTATCCATCGTTCTGTGGAAAACTAG
AATAGTTTATGGTAGAATAGTTCTAGAATTATCCACAAGAAGGAACCTAGTATGACTGAA
AATGAACAAATTTTTTGGAACAGGGTCTTGGAATTAGCTCAGAGTCAATTAAAACAGGCA
ACTTATGAATTTTTTGTTCATGATGCCCGTCTATTAAAGGTCGATAAGCATATTGCAACT
ATTTACTTAGATCAAATGAAAGAGCTCTTTTGGGAAAAAAATCTTAAAGATGTTATTCTT
ACTGCTGGTTTTGAAGTTTATAACGCTCAAATTTCTGTTGACTATGTTTTCGAAGAAGAC
CTAATGATTGAGCAAAATCAGACCAAAATCAACCAAAAACCTAAGCAGCAAGCCTTAAAT
TCTTTGCCTACTGTTACTTCAGATTTAAACTCGAAATATAGTTTTGAAAACTTTATTCAA
GGAGATGAAAATCGTTGGGCTGTTGCTGCTTCAATAGCAGTAGCTAATACTCCTGGAACT
ACCTATAATCCTTTGTTTATTTGGGGTGGCCCTGGGCTTGGAAAAACCCATTTATTAAAT
GCTATTGGTAATTCTGTACTATTAGAAAATCCAAATGCTCGAATTAAATATATCACAGCT
GAAAACTTTATTAATGAGTTTGTTATCCATATTCGCCTTGATACCATGGATGAATTGAAA
GAAAAATTTCGTAATTTAGATTTACTCCTTATTGATGATATCCAATCTTTAGCTAAAAAA
ACGCTCTCTGGAACACAAGAAGAGTTCTTTAATACTTTTAATGCACTTCATAATAATAAC
AAACAAATTGTCCTAACAAGCGACCGTACACCAGATCATCTCAATGATTTAGAAGATCGA
TTAGTTACTCGTTTTAAATGGGGATTAACAGTCAATATCACACCTCCTGATTTTGAAACA
CGAGTGGCTATTTTGACAAATAAAATTCAAGAATATAACTTTATTTTTCCTCAAGATACC
ATTGAGTATTTGGCTGGTCAATTTGATTCTAATGTCAGAGATTTAGAAGGTGCCTTAAAA
GATATTAGTCTGGTTGCTAATTTCAAACAAATTGACACGATTACTGTTGACATTGCTGCC
GAAGCTATTCGCGCCAGAAAGCAAGATGGACCTAAAATGACAGTTATTCCCATCGAAGAA
ATTCAAGCGCAAGTTGGAAAATTTTACGGTGTTACCGTCAAAGAAATTAAAGCTACTAAA
CGAACACAAAATATTGTTTTAGCAAGACAAGTAGCTATGTTTTTAGCACGTGAAATGACA
GATAACAGTCTTCCTAAAATTGGAAAAGAATTTGGTGGCAGAGACCATTCAACAGTACTC
CATGCCTATAATAAAATCAAAAACATGATCAGCCAGGACGAAAGCCTTAGGATCGAAATT
GAAACCATAAAAAACAAAATTAAATAACATGTGGAAAAGAATATCTTTTATGAAATAGTT
ATCCACAAGTTGTGAACATCCATTTAGTCTTGGATTCTCTCGTTTATTTAGAGTTATCCA
CTATATACACAAGACCTACTACTACTACTTATTATTATACTTATTAAATAAAGGAGTTCT
CATGATTCAATTTTCAATTAATCGCACATTATTTATTCATGCTTTAAATACAACTAAACG
TGCTATTAGCACTAAAAATGCCATTCCTATTCTTT ...
```

Proteins

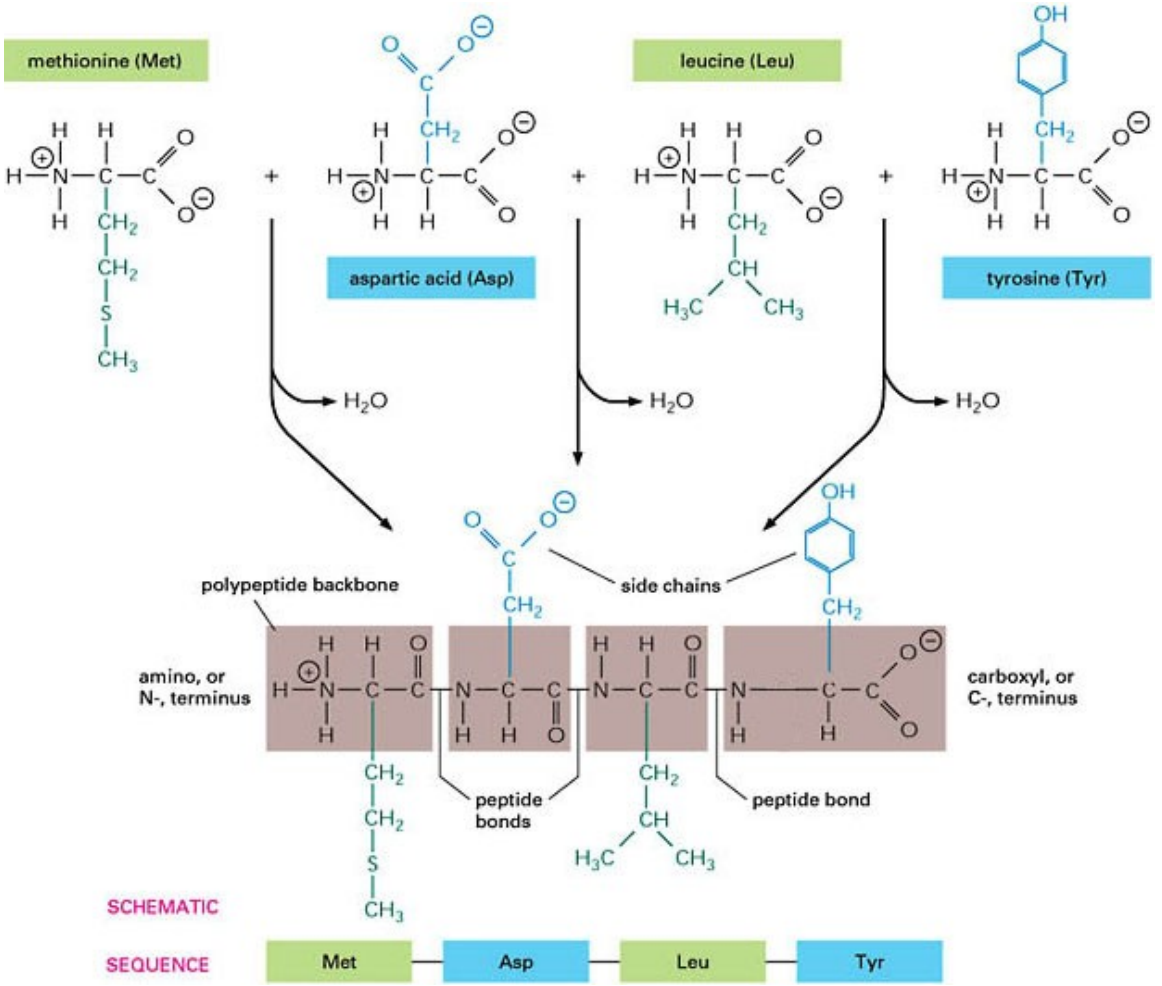
The building blocks of all living organisms



Structural proteins: tissue binding block

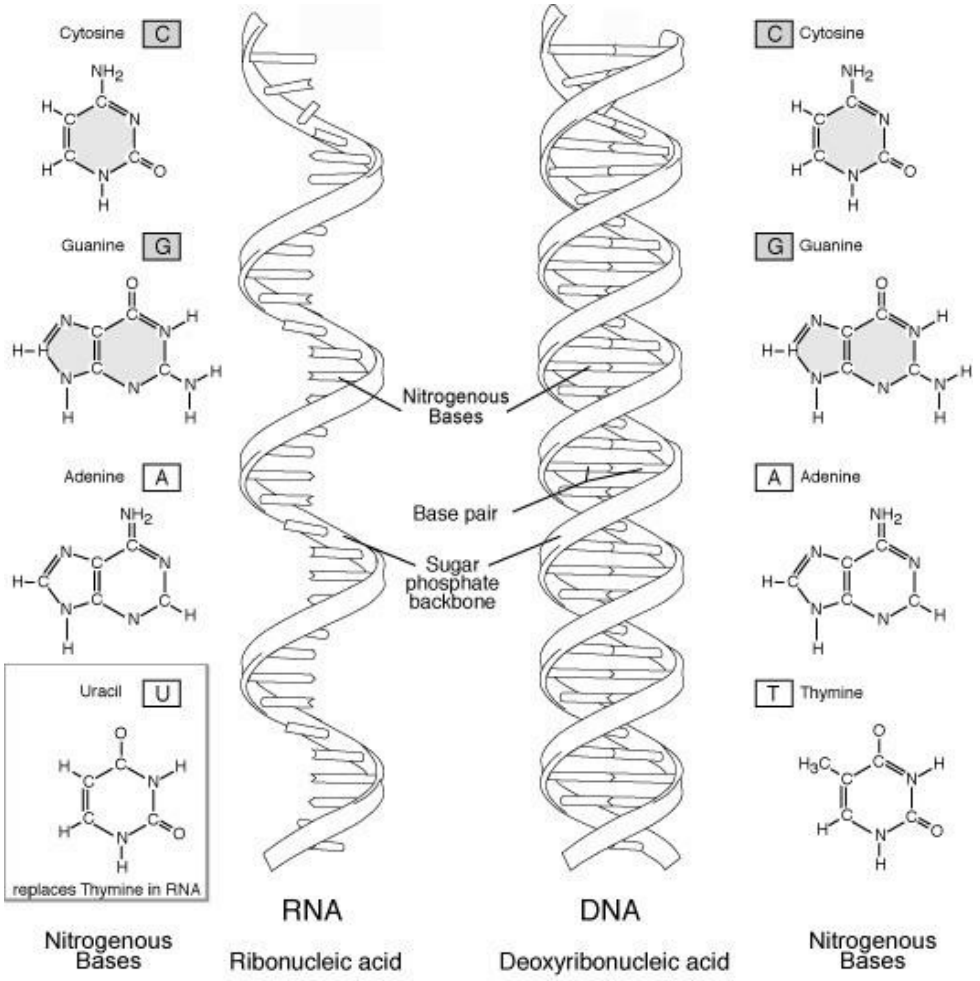
Enzymes: catalysts of chemical reactions

Chain of amino acids



$S \in \{A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y\}^*$

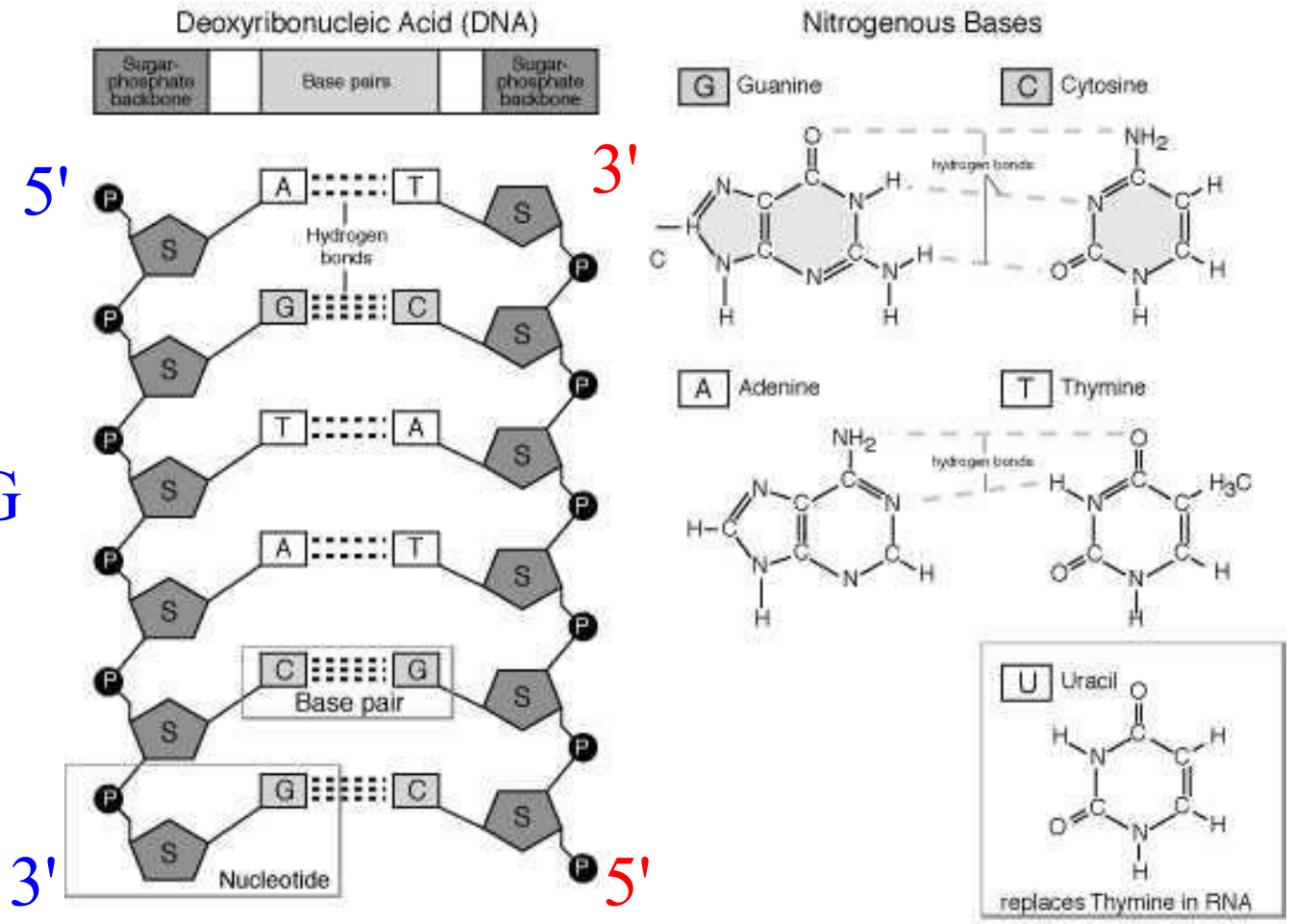
Nucleic Acids - RNA and DNA



The carrier of genetic information - The blueprints of proteins

DNA - deoxyribonucleic acid

AGTACG

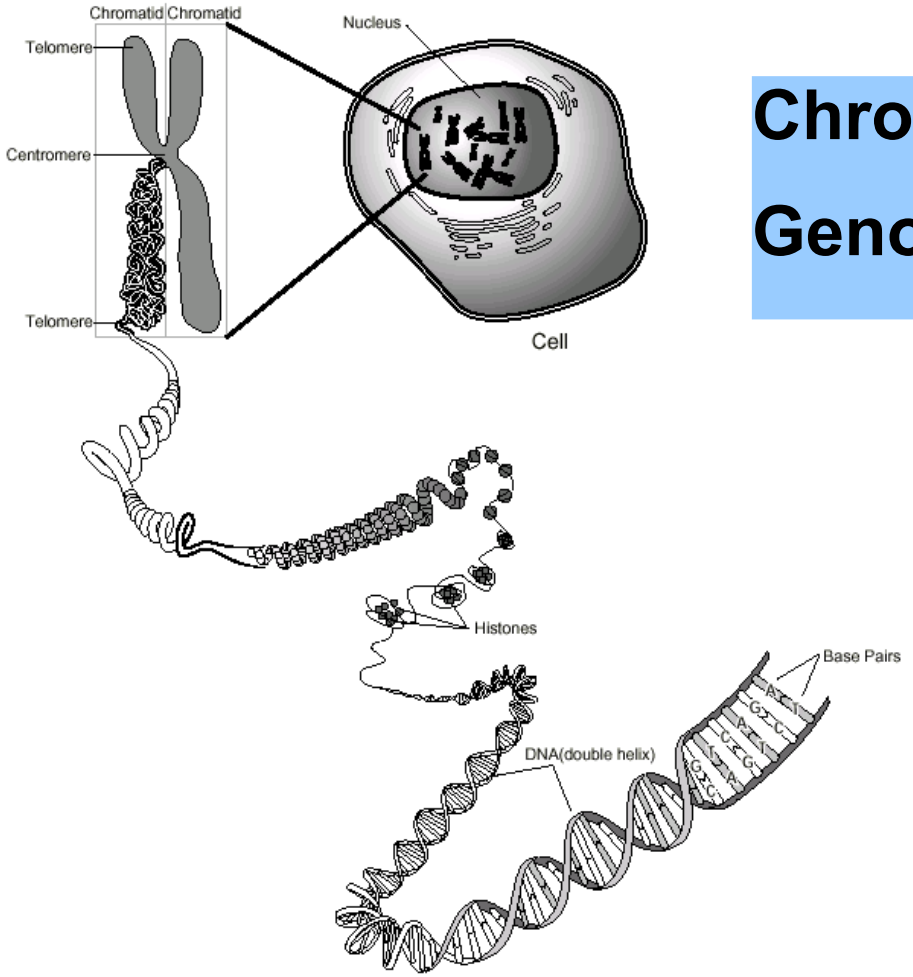


A DNA molecule is two complementary strands of nucleotides

DNA: $s \in \{A,C,G,T\}^*$ RNA: $s \in \{A,C,G,U\}^*$

Cells and Chromosomes

Any organisme consists of one to billions of cells



Chromosomes: Large DNA molecules

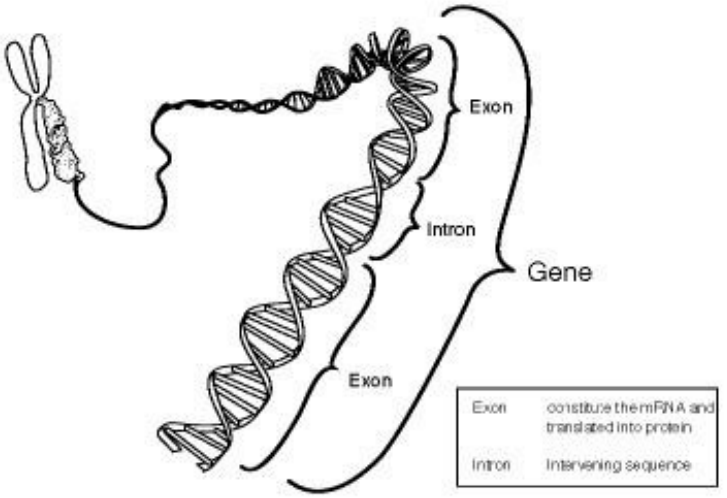
Genome: The collection of chromosomes

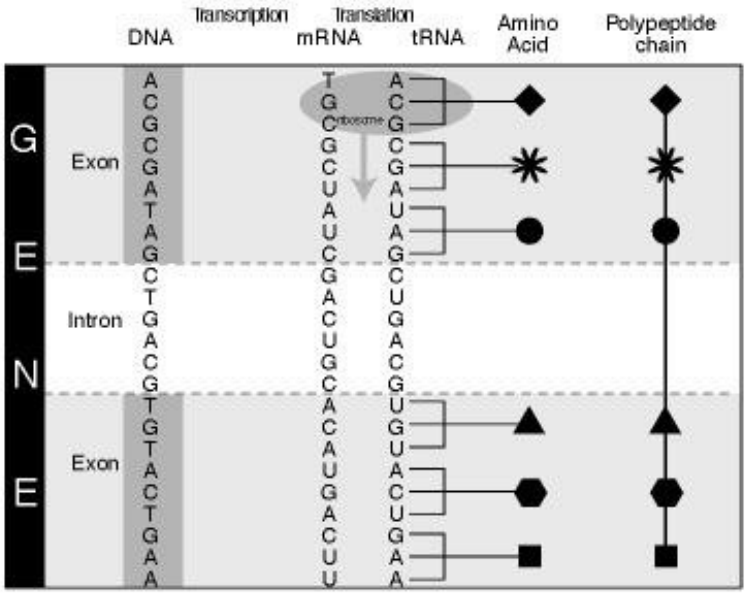
Human: 23 pairs, 3.100.000.000 bp

Fruit fly: 4 pairs, 200.000.000 bp

Yeast: 16 pairs, 10.000.000 bp

Genes - Blue prints of proteins





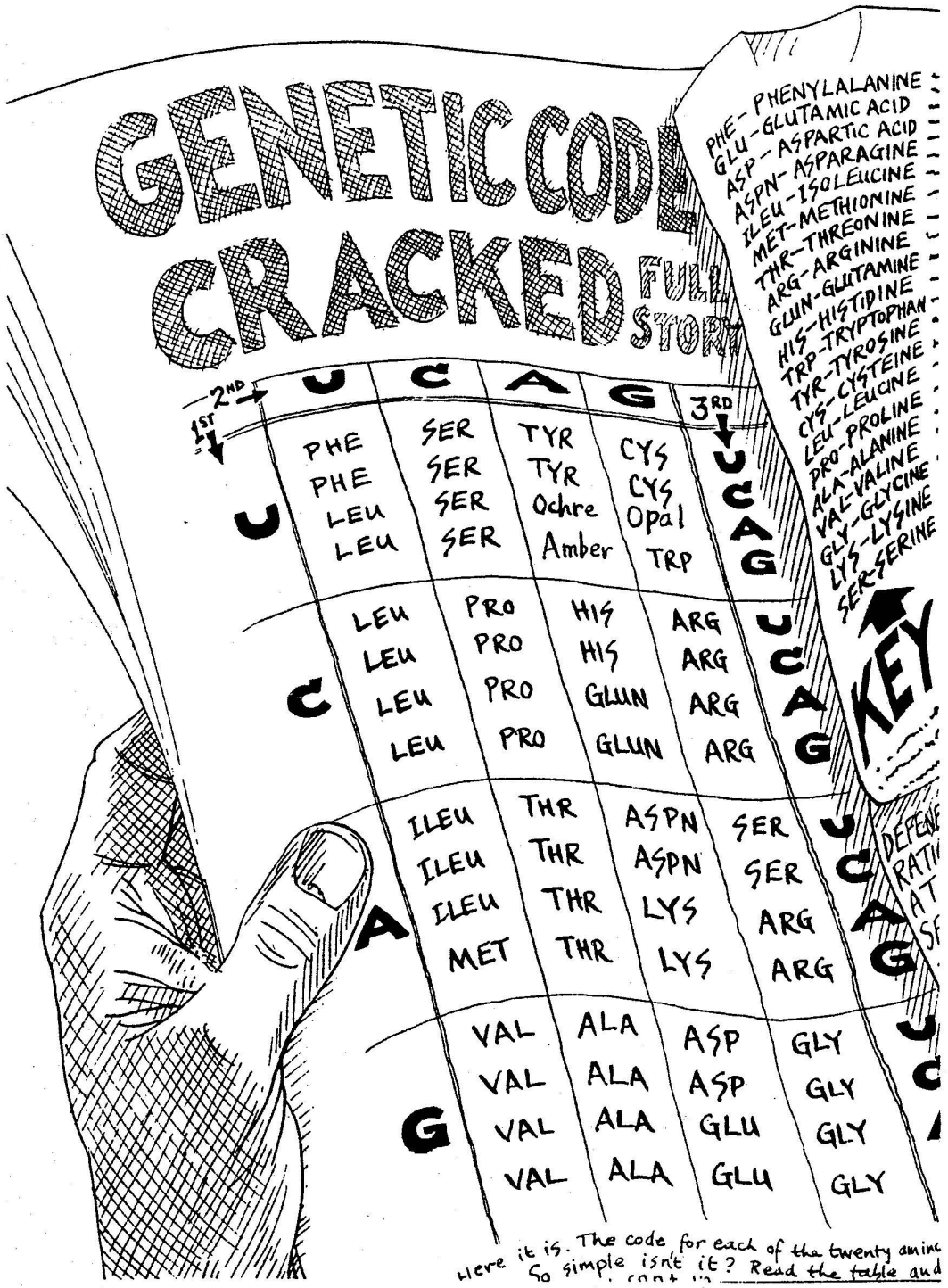
Each protein is encoded in a stretch of DNA. A gene ...

Which is expressed when the protein is needed ...

Important problem

Locating genes on the genome and determining how they get expressed ...

Recognizing the patterns that indicates a gene ...

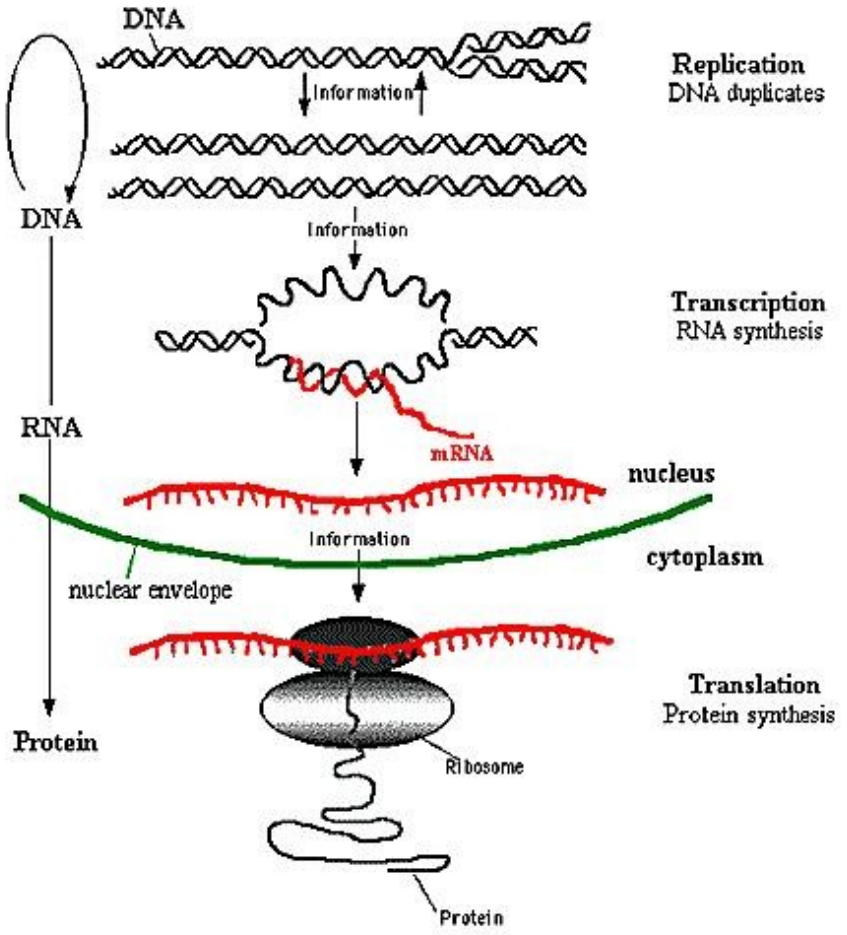

GENETIC CODE
CRACKED FULL STORY
1ST
2ND
3RD
U C A G
U
PHE SER TYR CYS U
PHE SER TYR CYS C
LEU SER Ochre Opal A
LEU SER Amber TRP G
C
LEU PRO HIS ARG U
LEU PRO HIS ARG C
LEU PRO GLUN ARG A
LEU PRO GLUN ARG G
A
ILEU THR ASPN SER U
ILEU THR ASPN SER C
ILEU THR LYS ARG A
MET THR LYS ARG G
G
VAL ALA ASP GLY U
VAL ALA ASP GLY C
VAL ALA GLU GLY A
VAL ALA GLU GLY G
PHE - PHENYLALANINE
GLU - GLUTAMIC ACID
ASP - ASPARTIC ACID
ASPN - ASPARAGINE
ILEU - ISOLEUCINE
MET - METHIONINE
THR - THREONINE
ARG - ARGININE
GLUN - GLUTAMINE
HIS - HISTIDINE
TRP - TRYPTOPHAN
TYR - TYROSINE
CYS - CYSTEINE
LEU - LEUCINE
PRO - PROLINE
ALA - ALANINE
VAL - VALINE
GLY - GLYCINE
LYS - LYSINE
SER - SERINE
KEY
Here it is. The code for each of the twenty amino
So simple isn't it? Read the table and

The Central Dogma

$s \in \{A,C,G,T\}^*$

$s \in \{A,C,G,U\}^*$

$s \in \{A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y\}^*$



The Central Dogma of Molecular Biology

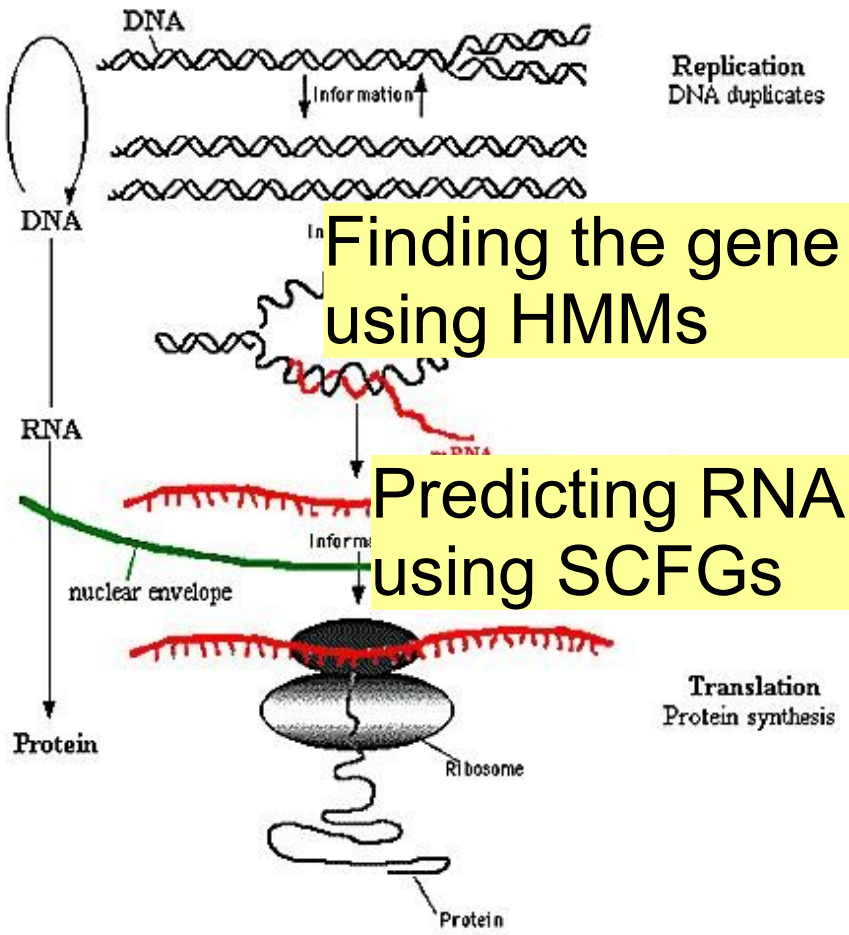
Everything are strings! A good model of important biology

Bioinformatics applications

$s \in \{A,C,G,T\}^*$

$s \in \{A,C,G,U\}^*$

$s \in \{A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y\}^*$



The Central Dogma of Molecular Biology

Finding the gene structure using HMMs

Predicting RNA structure using SCFGs

Everything are strings! A good model of important biology