# Quantitative Biology

- Building mathematical models to understand and predict behavior of living systems
- A wide range of hierarchical models of life:
  - Self contained replicating molecules (?)
  - Cells
  - Co-operating cells
  - Populations
  - Artificial life

  Extensive use of abstraction (good computer programming)

# CS: 426 Introduction to computational (***molecular***) biology

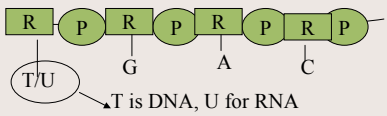
- A bottom to top approach
  - What is a "minimal" model of life?

    A self-replicating molecule that uses material and energy from the environment
  - The most basic biological molecules are DNA, RNA, proteins, and lipids. Which is a candidate for a minimal model?

    For a long time people try to prove that proteins are the best choice, but more recently RNA became the more natural choice.

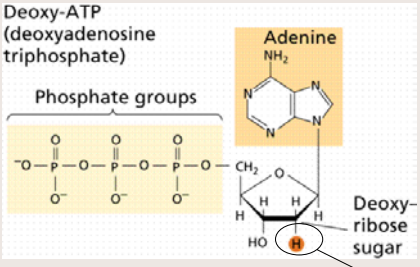
# Why RNA?

- RNA performs two functions coding information and synthesis
  - Like essentially all the important biological molecules RNA is a linear polymer that is made of a small number of monomers
  - The different monomers are attached to the same backbone (sugar rings (ribose) linked by phosphates)




# RNA/DNA (continue)

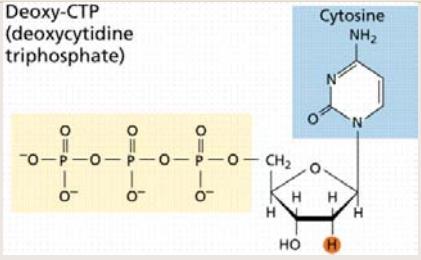
- **Closer look at basic syntax**




***Called by "friends and family" : A***

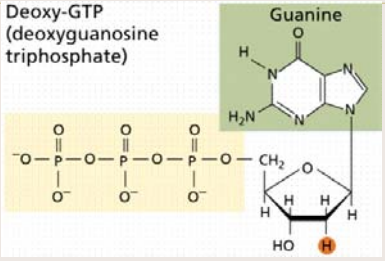
For RNA
Replace H by OH

## More RNA/DNA




***Called C***

## RNA/DNA: basic




***Called G So far we have A,C,G for DNA we also have T for RNA U. Total of 4 "bases".***

## The bases

Adenine

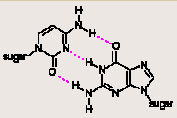
Guanine

Thymine
*DNA only*

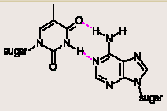
Uracil
*RNA only*

Cytosine

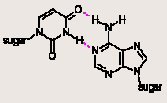
## The base pairs (hydrogen bonds)



CG

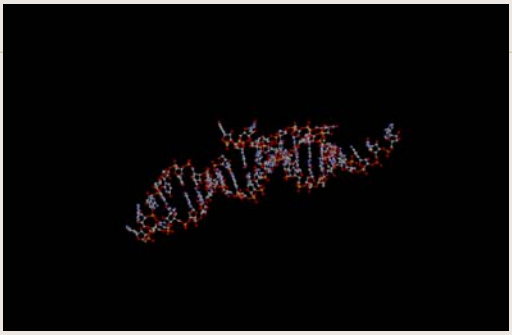


AT



AU

## RNA



Hydrogen bonding between complementing bases:
G:C or A:T/U (DNA/RNA)

## RNA (the n-th time)

- The base pairing C:G and A:T determine structure of the RNA (must match!). One dimensional sequence determines shape e.g.
- G G G A G C U C A A C U C U C C C C C C C U U U U C C G A G G G U C A U C G G A A C C A
- Determine the shape from sequence using maximum pairing condition (RNA structure prediction problem).

## RNA (n+1) time

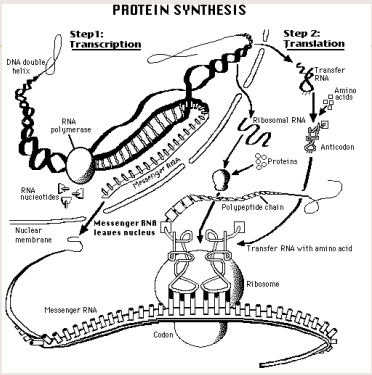
## RNA (n+2)

- RNA has globular shape
- Double molecular function:
  - Linear sequence stores information (information in biology is stored linearly)
  - Three dimensional shape determines enzymatic (machine-like) activity
  - **Can be dangerous for the data to process itself!**

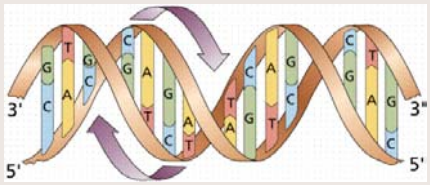
## Life cannot be made too simple

- As a cautionary step nature separated information from processing and cell function. Genetic information (in particular) must be protected exceptionally well.
- Instead of one molecule that does it all we now have two: DNA (information) Proteins (operations). RNA is kept in the middle probably for historical reasons and additional safety nets.

## From storage to machines




## DNA is a "boring" molecule that "only" stores linear information




In contrast to RNA no complex three dimensional fold or enzymatic function (only hard disk no CPU). A **double** helix (A/B/Z) of complementing base pairs, very useful in information processing and corrections

## DNA in the most common B form

Genetic Table From DNA To proteins

From 4 bases To 20 amino acids

Many errors Allowed.

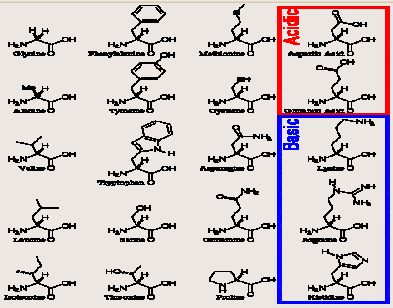
Not all DNA codes proteins

| | T | C | A | G |
|---|---|---|---|---|
| **T** | TTT Phe (F)<br>TTC "<br>TTA Leu (L)<br>TTG " | TCT Ser (S)<br>TCC "<br>TCA "<br>TCG " | TAT Tyr (Y)<br>TAC<br>TAA **Ter**<br>TAG **Ter** | TGT Cys (C)<br>TGC<br>TGA **Ter**<br>TGG Trp (W) |
| **C** | CTT Leu (L)<br>CTC "<br>CTA "<br>CTG " | CCT Pro (P)<br>CCC "<br>CCA "<br>CCG " | CAT His (H)<br>CAC "<br>CAA Gln (Q)<br>CAG " | CGT Arg (R)<br>CGC "<br>CGA "<br>CGG " |
| **A** | ATT Ile (I)<br>ATC "<br>ATA "<br>**ATG** Met (M) | ACT Thr (T)<br>ACC "<br>ACA "<br>ACG " | AAT Asn (N)<br>AAC "<br>AAA Lys (K)<br>AAG " | AGT Ser (S)<br>AGC "<br>AGA Arg (R)<br>AGG " |
| **G** | GTT Val (V)<br>GTC "<br>GTA "<br>GTG " | GCT Ala (A)<br>GCC "<br>GCA "<br>GCG " | GAT Asp (D)<br>GAC "<br>GAA Glu (E)<br>GAG " | GGT Gly (G)<br>GGC "<br>GGA "<br>GGG " |

# More on DNA

- In higher organisms a lot of "junk" non coding DNA (up to 90 percent)
- Junk DNA may play a role in regulation controlling gene transcription/introns + exons
- Reading frames: Can shift a base to read different amino acids, e.g. (codon – 3 basepairs)
  - (TAA)(TCG)(AAT)(GGG)C == XSNG
  - T(AAT)(CGA)(ATG)(GGC) == NRMG
  - TA(ATC)(GAA)(TGG)GC == IEW
  - There are actually six ways of reading the DNA…
  - Open reading frame

# The monomers in a protein chain are the 20 amino acids




Name Abbr. Linear structure formula AMINO ACIDS

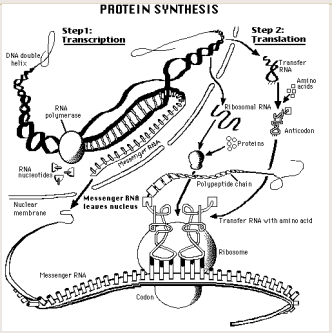
Alanine **ala** a CH3-CH(NH2)-COOH
Arginine **arg** r HN=C(NH2)-NH-(CH2)3-CH(NH2)-COOH
Asparagine **asn** n H2N-CO-CH2-CH(NH2)-COOH
Aspartic acid **asp** d HOOC-CH2-CH(NH2)-COOH
Cysteine **cys** c HS-CH2-CH(NH2)-COOH
Glutamine **gln** q H2N-CO-(CH2)2-CH(NH2)-COOH
Glutamic acid **glu** e HOOC-(CH2)2-CH(NH2)-COOH
Glycine **gly** g NH2-CH2-COOH
Histidine **his** h NH-CH=N-CH=C-CH2-CH(NH2)-COOH
Isoleucine **ile** i CH3-CH2-CH(CH3)-CH(NH2)-COOH
Leucine **leu** l (CH3)2-CH-CH2-CH(NH2)-COOH
Lysine **lys** k H2N-(CH2)4-CH(NH2)-COOH
Methionine **met** m CH3-S-(CH2)2-CH(NH2)-COOH
Phenylalanine **phe** f Ph-CH2-CH(NH2)-COOH
Proline **pro** p NH-(CH2)3-CH-COOH
Serine **ser** s HO-CH2-CH(NH2)-COOH
Threonine **thr** t CH3-CH(OH)-CH(NH2)-COOH
Tryptophan **trp** w Ph-NH-CH=C-CH2-CH(NH2)-COOH
Tyrosine **tyr** y HO-p-Ph-CH2-CH(NH2)-COOH
Valine **val** v (CH3)2-CH-CH(NH2)-COOH

## The machines (proteins)

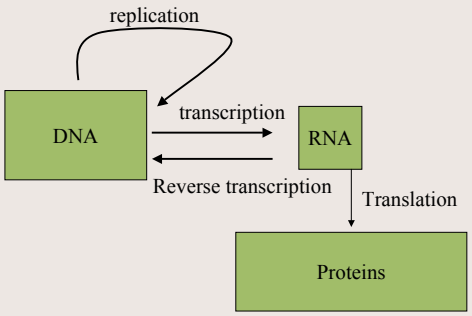
A signal RAS protein

Cytochrome C
Electron transport

## Another look on separation of information and function




## And in brief



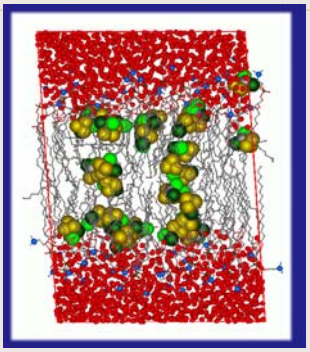

## More terminology

- We shall call a DNA segment that codes a protein – a gene
- The complete set of genetic material is called a genome
- Genomes are divided into chromosomes. Simple bacteria (Escherichia Coli) has 1 (5M basepairs). Humans have 46 chromosomes (3B basepairs)

## Experimental techniques of studying genomes (sequencing)

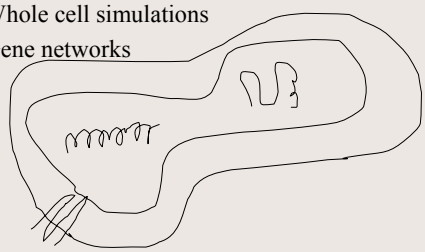
- Genetic linkage maps 1-10M bp
- Physical maps 0.1-1M bp
- Sequencing 1K-10K bp

## Lipids and cell walls



## What is life (again)?

- Divide inside from outside
  - Whole cell simulations
  - Gene networks



## What we do to understand biology?

- Use chemical physics principles
- Look for problems we already solved (homology)

# Computational tasks of molecular biology

- Assembly of DNA fragments
- Identifying Genes (DNA segments that code)
- Building evolutionary trees
- Gene networks & micro array analysis
- Whole cell simulations
- Identify gene function
  - Chemical physics
  - Homology