COMP90016 – Computational Genomics *Genomics II*

Department of Computing and Information Systems

The University of Melbourne

Outlook

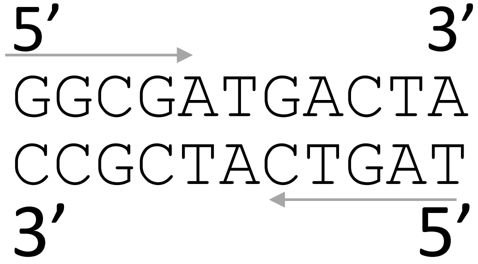
- DNA in cells
  - Organisation
  - Information flow
  - Alleles
- DNA information content

DNA Summary and Example

- From an information relevant point of view, DNA strands are sequences of nucleotides A, C, G, and T, for example GGCGATGACTA
- Each base/nucleotide is arranged opposite to its matching counterpart (A<->T, C<->G), forming a double-stranded double helix. For example

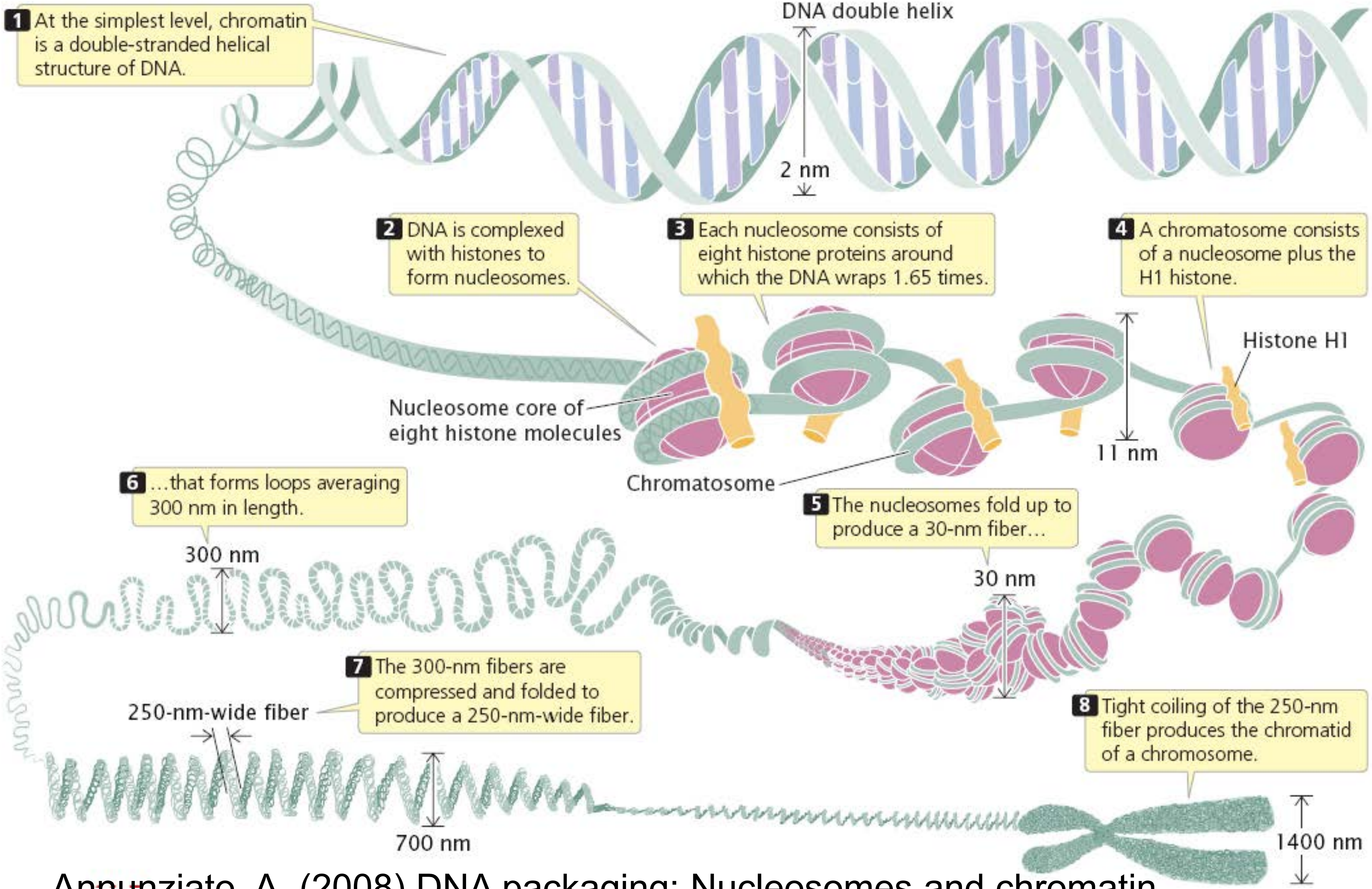
```
GGCGATGACTA
CCGCTACTGAT
```

- The strands can be read (sequencing) or transcribed (cell-internal processes) in one direction only: in the 5’ to 3’ direction. For example



```
5’          3’
GGCGATGACTA
CCGCTACTGAT
3’          5’
```

Organisation of DNA



1 At the simplest level, chromatin is a double-stranded helical structure of DNA.

2 DNA is complexed with histones to form nucleosomes.

3 Each nucleosome consists of eight histone proteins around which the DNA wraps 1.65 times.

4 A chromatosome consists of a nucleosome plus the H1 histone.

5 The nucleosomes fold up to produce a 30-nm fiber…

6 …that forms loops averaging 300 nm in length.

7 The 300-nm fibers are compressed and folded to produce a 250-nm-wide fiber.

8 Tight coiling of the 250-nm fiber produces the chromatid of a chromosome.

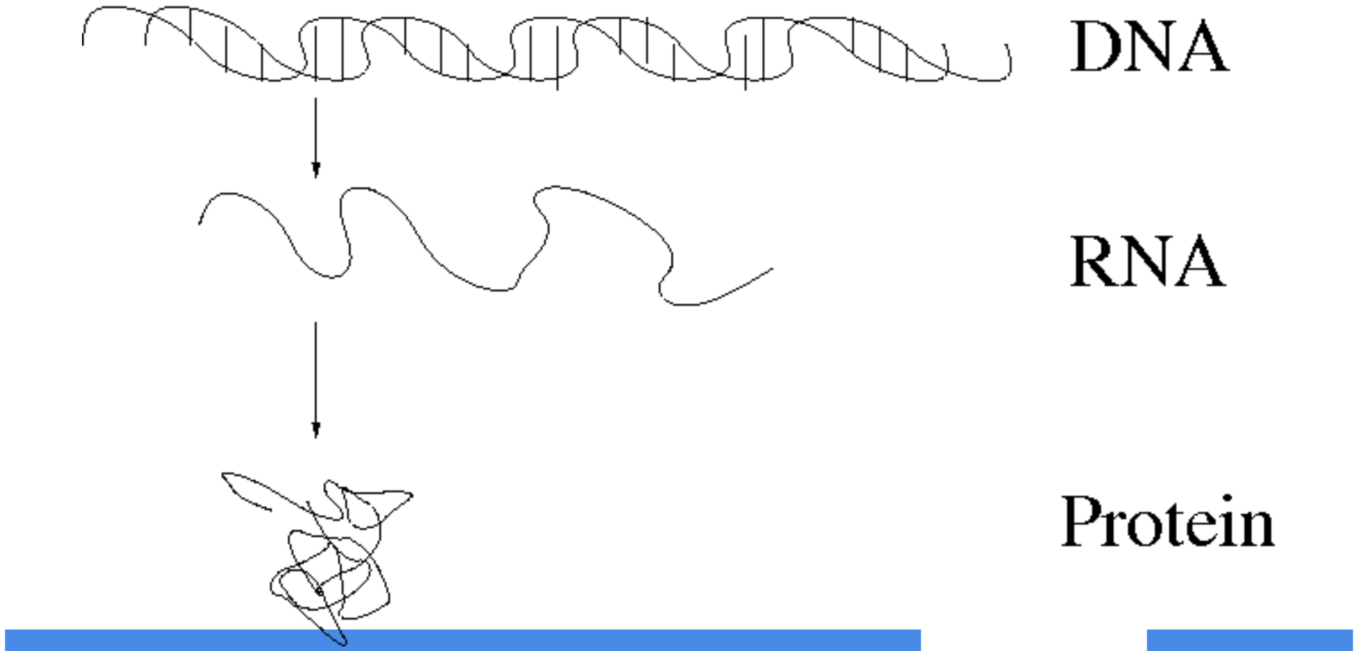
Annunziato, A. (2008) DNA packaging: Nucleosomes and chromatin. Nature Education 1(1)

Organisation of DNA 2

- When doing whole genome sequencing, the process unwinds the DNA of any organizational states.
- However, the packaging of DNA has profound implications on how it is utilized within a cell.
  - Different parts of the DNA are accessible in different cell types.
  - This is part of the different functionality of cell types (brain, lung, liver, muscle…)
- There are sequencing assays other than WGS, that study DNA accessibility:
  - DNase-seq
  - ATAC-seq
- These techniques are outside our scope, but keep them in mind as potential avenues of genomics exploration in your future projects.
  Open bioinformatics challenges:
  - Peak calling of accessible regions.
  - Footprinting of nucleosomes and transcription factors binding to DNA during experiment.

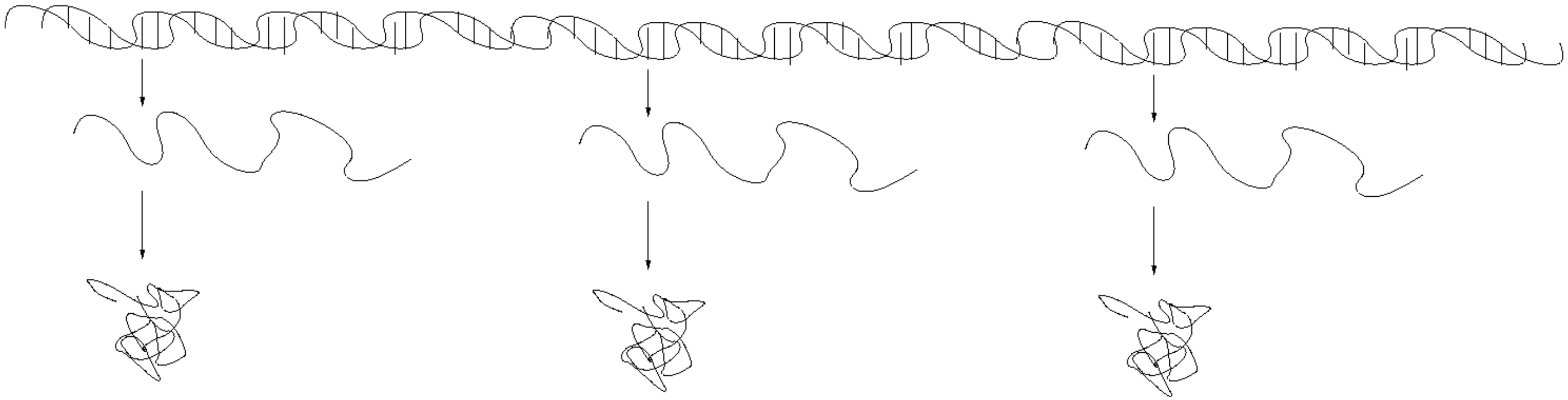
DNA information flow

- DNA is the blueprint, proteins do the work.



• DNA→ RNA: transcription	1-to-1
• RNA → protein: translation	3-to-1

DNA information flow (2)



- The genome has lots of genes (coding sequences).
- And lots of un-transcribed space (non-coding regions).

Genes and Proteins

- Gene (common definintion):
  A piece of DNA that encodes a protein.
- Proteins do most of the work in a cell.
- The link (translation):
  Three DNA bases code for one amino acid, *e.g.* ATG → methionine (met, M).

Heredity and Genes

- There are over 20 thousand genes in the human genome.
- Everyone has all of those genes (Y chromosome aside).
  - In fact, everybody has two copies of each gene.
- But the copies may be different.
  - There may be many different versions (alleles) of a single gene present in a population.
  - Each individual has two (possibly identical) copies from the available variants (one of which he or she got from their mother, one from the father).
- Genes determine how we look, walk, talk, feel, … etc in a direct or complex way.
  - Having a certain variant of a gene can directly determine our eye colour.
  - Having certain sets of variants for many different genes may increase our risk for dementia in the future.
- The traits (eye colour etc) are called the phenotype.
- The set of gene variants for an individual are called the genotype.

Heredity and Genes 2

- Genotypes can be dominant or recessive:
  - Since we have two copies of each gene, there may be two different proteins available for translation.
  - A gene variant is called dominant if a single copy is sufficient to dictate the phenotype. Example: Eye colour is dominated by the brown eye variant, so a single copy of the "brown eye gene" (it's more complex than that, and we will study this further), will cause a brown eyed phenotype.
  - The blue eyed trait is recessive. A person needs a genotype with two copies of a certain variant in order to have blue eyes.
- The different variants of genes are also called alleles.
- Further Reading: **The Gene: An Intimate History**

DNA Information Flow Regulation

- There are many stages and processes that regulate gene expression (translation) – which can either improve or disrupt this process:
    - DNA accessibility (see above).
    - Availability of transcription factors (proteins that need to be present – interact with DNA - to initiate the transcription process).
    - Transportations and denaturing of RNA.
    - Availability/presence of enhancers/silencers (proteins that interact with the DNA outside of genes to increase or decrease transcription.
    - Compatibility of RNA with ribosomes (which do the translation).
    - Splice site modifications.
    - DNA methylation.
    - And many more.

Transcription and translation in more detail: Prokaryotic process

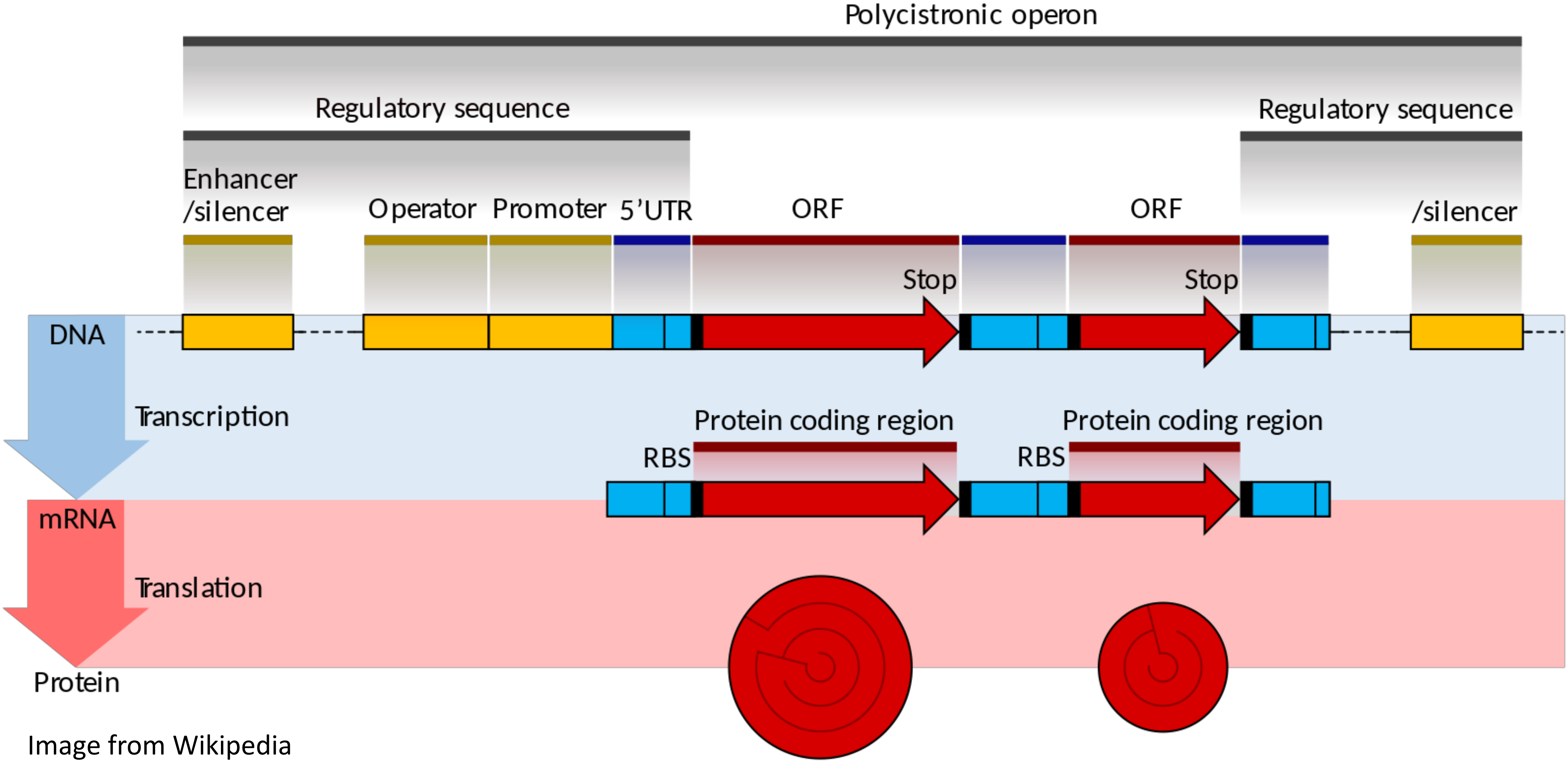


Image from Wikipedia

Transcription and translation in more detail: Eukaryotic process

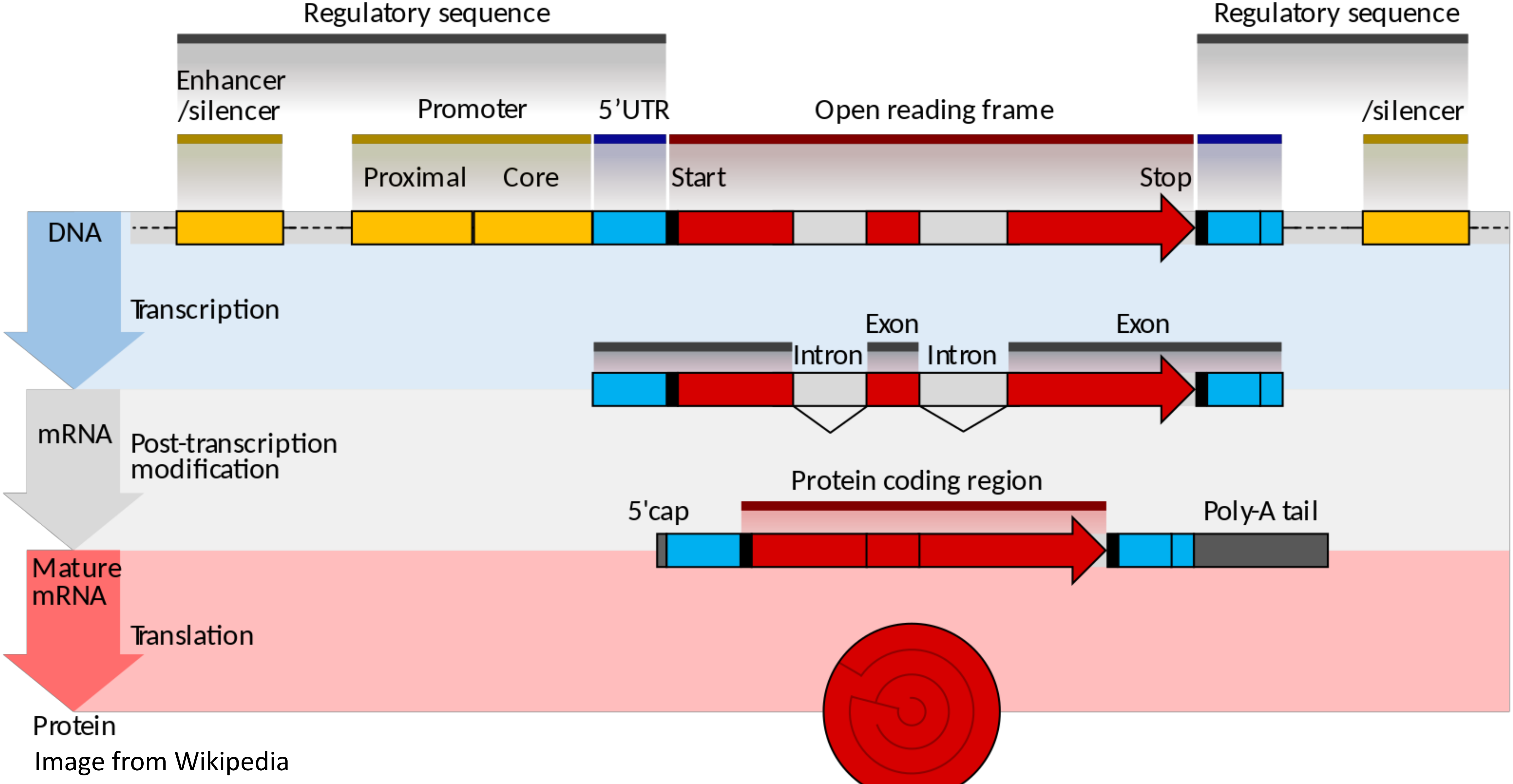
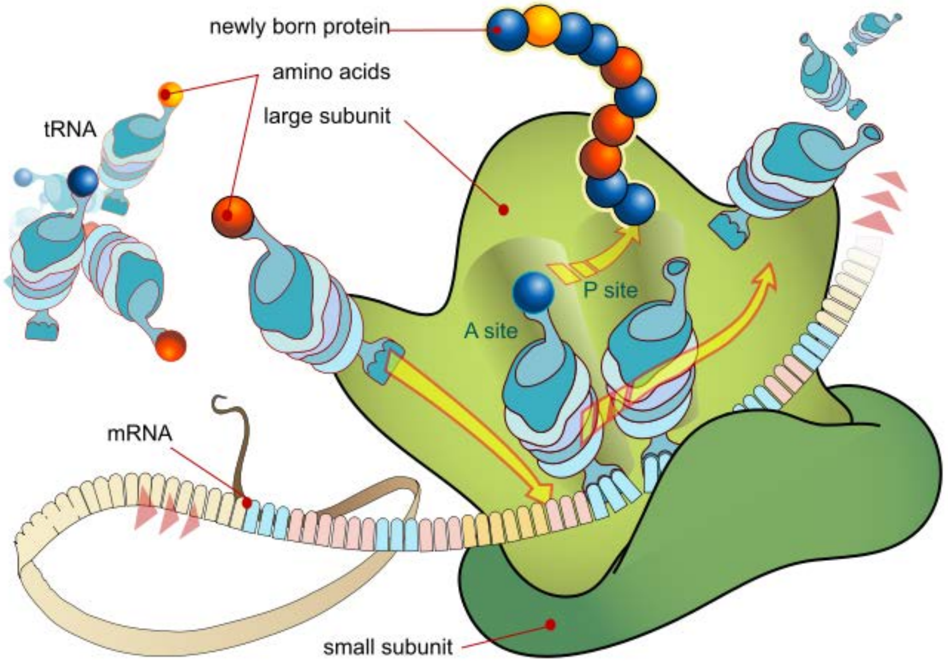


Image from Wikipedia

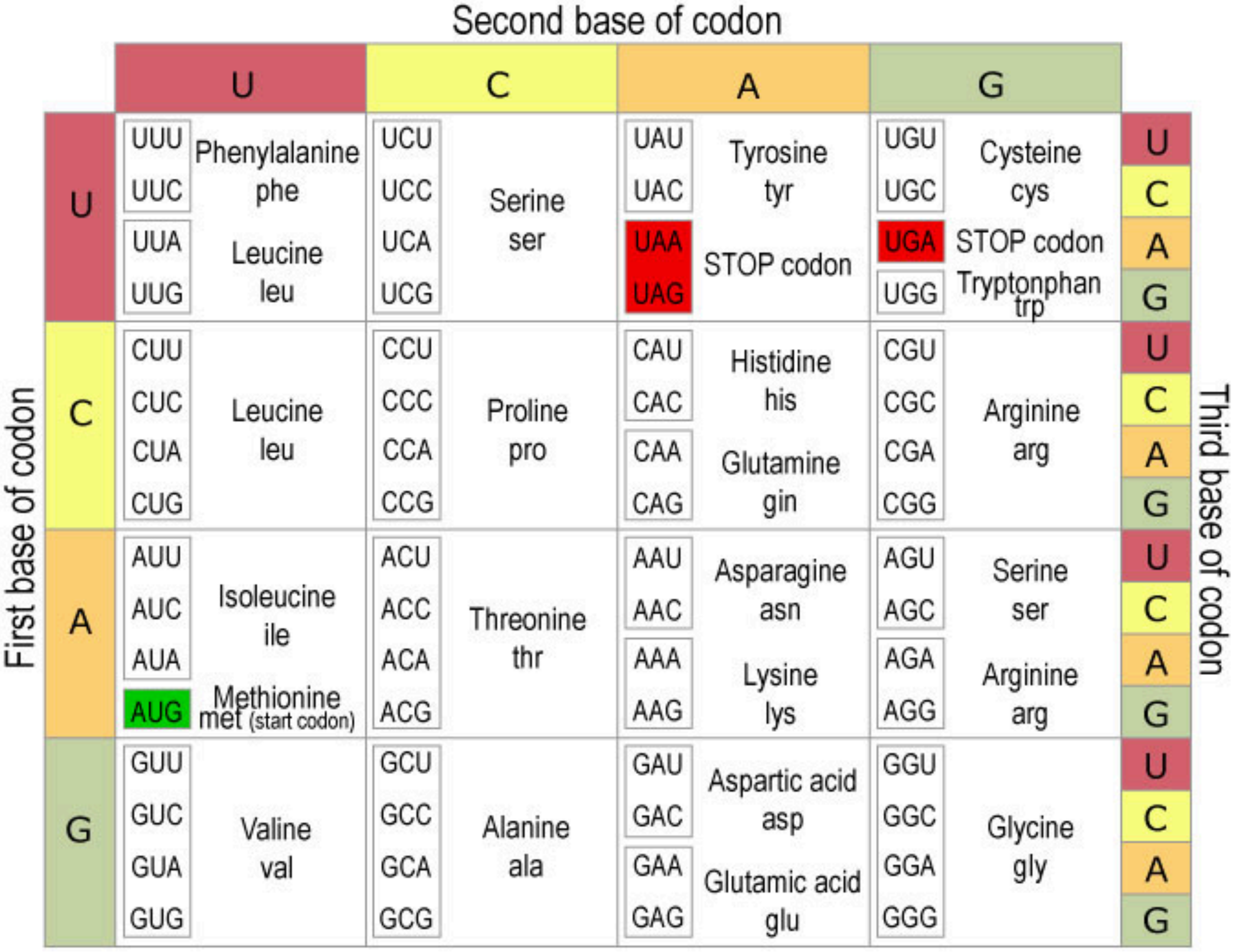
Translation

- Translation is transferring the information in the RNA into a new molecule by ribosomes.
  - The protein is made out of amino acids.
  - Any three nucleotides (codon) specify exactly one amino acid.
  - Similar to DNA synthesis (turning single-stranded DNA into double-stranded), the ribosome complements every three nucleotides by a corresponding amino acid.
- The translated amino acid chain will form a complex 3D structure giving the protein its properties.



Redundancy in Translation

- DNA alphabet = 4 bases
- Given that 3 bases in DNA specify one amino acid in protein, how many amino acids could a DNA alphabet of size 4 (theoretically) encode?
- How many amino acids are there?



First base of codon	Second base of codon: U		Second base of codon: C		Second base of codon: A		Second base of codon: G		Third base of codon
U	UUU	Phenylalanine phe	UCU	Serine ser	UAU	Tyrosine tyr	UGU	Cysteine cys	U
	UUC		UCC		UAC		UGC		C
	UUA	Leucine leu	UCA		UAA	STOP codon	UGA	STOP codon	A
	UUG		UCG		UAG		UGG	Tryptonphan trp	G
C	CUU	Leucine leu	CCU	Proline pro	CAU	Histidine his	CGU	Arginine arg	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	Glutamine gin	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	Isoleucine ile	ACU	Threonine thr	AAU	Asparagine asn	AGU	Serine ser	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	Lysine lys	AGA	Arginine arg	A
	AUG	Methionine met (start codon)	ACG		AAG		AGG		G
G	GUU	Valine val	GCU	Alanine ala	GAU	Aspartic acid asp	GGU	Glycine gly	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	Glutamic acid glu	GGA		A
	GUG		GCG		GAG		GGG		G

© Clinical Tools, Inc.

Reading Frame

- Since three bases encode for one amino acid, there are three distinct positions in each gene that translation could start from.
- The three positions are referred to as reading frames.
  ACGATGACTA
  ACGATGACTAA = thr, met, thr
  ACGATGACTAA = arg, stop, leu
  ACGATGACTAA = …
- Generally, only one reading frame will result in a functional protein.
  - This reading frame will start with the start codon (met) and end with a stop codon.

Discussion Questions

- What possible advantages might there be to having redundancy in the genetic code?
- What properties would a code need in order to realize this/these advantage(s)?
- What if only 2 bases encoded 1 amino acid? Assuming we still have an alphabet of 4 bases, what would the maximum number of amino acids be?
- If 4 bases encode 1 amino acid?

Extended alphabet

- Sometimes the exact identity of the base is not known:
    - Uncertainties in sequencing (wet lab).
    - Inexact patterns (biological motifs).
- In these cases, an extended alphabet is useful:
    - Single nucleotides are still: {T} {C} {A} {G}
    - Any one nucleotide, identity unknown:
      N = {T,C,A,G}

Extended nucleotide alphabet

- The 4-symbol alphabet can specify only a unique and completely specified sequence.
- However, motifs may be common to many regions, and may contain variants.
- For example:

  AAGNNNTTC, where NNN means “any three nucleotides”.
- Variations between genes are often single nucleotide differences: one of two bases may be present in any individual.
- An extended alphabet is helpful to describe these situations.

Full extended nucleotide alphabet

- Single nucleotides: {T} {C} {A} {G}
- Anything: N={A,C,G,T}
- Pyrimidines, purines: Y={C,T}, R={A,G}
- Weak/strong bonding: W={A,T}, S={C,G}
- Amino/keto: M={A,C}, K={G,T}
- V={A,C,G}; H={A,C,T}; D={A,G,T}; B={C,G,T}

- For sequences built from only single nucleotides TCAG, what is the size of the extended alphabet?

This encoding is called the IUPAC code, and can be reviewed at http://www.bioinformatics.org/sms/iupac.html

Number of possibilities and the power set

- Power set is the set of all subsets.
- *e.g.* for T,C,A,G the power set includes all components of the extended alphabet (super set).
- Size of the power set:
  - 2^n, n the size of the set.

DNA information

- DNA is the blueprint.
- How big is human DNA?
  - Approx 3 x 10^9 bases.
  - Approx 1 m unwound
    (contrast most eukaryotic cells 10-30 μm diameter)
  - Much (>90%) DNA is non-coding.
  - Much (>30%) DNA is repetitive.
  - Composition varies across organisms, across genome.

DNA information (discussion):

- How much computer space (in bytes) is needed to store a sequence the size of the human genome (3 x 10^9 bases)?
- Does this amount of space vary, depending on the actual sequence?
- Given that one person’s DNA varies ~1% from another’s, on average, how much space is needed to store the genomes of 10 people?

Compressibility as a Measure of Information

- Redundant data compresses well, *e.g*.
  - AAAAAAAAAAAAAAAAAA... → 100A
  - AAAAAAATAAAAAAATAAAAAAAT→7AT7AT7AT
- Repeat patterns compress*, e.g.*
  - AAAAAAATAAAAAAATAAAAAAAT→(AAAAAAAT)3, or → (7AT)3
- Unique information does not compress well

Compression

- Compression is an attempt to encode the information as succinctly as possible.
- For DNA use only *lossless* compression.
- Compression involves:
  - A model (probability of each symbol).
  - A method for encoding the model, *e.g.* use more bits for low frequency symbols.
- We will use the model to assess information content, without the coding.

Entropy (informally)

- Entropy is the theoretically least number of bits necessary to encode a sequence.
  - *e.g*. sequence AAAAAAAAAA… needs 0 bits
  - *e.g.* for alphabet A,T, need 1 bit per symbol
  - *e.g.* for lots of As and Ts, a very few Cs, on average will need n bits/symbol, n > 1

Models

- DNA models:
  - *E. coli*:
    - p(T)=p(C)=p(A)=p(G) = 0.25
    - G+C=50%
    - In a sequence, the next base is equally likely to be T,A,G, or C – so the next base carries much information.
  - *P. falciparum*:
    - p(A)=p(T)=0.4
    - p(C)=p(G)=0.1
    - G+C=20%
    - Skewed base composition, expect A or T more often; A or T give us less information

- Information theory allows us to measure the amount of disorder/uncertainty/predictability.

Entropy: a way of measuring information

- Entropy in a sequence:
  - a measure of how much information

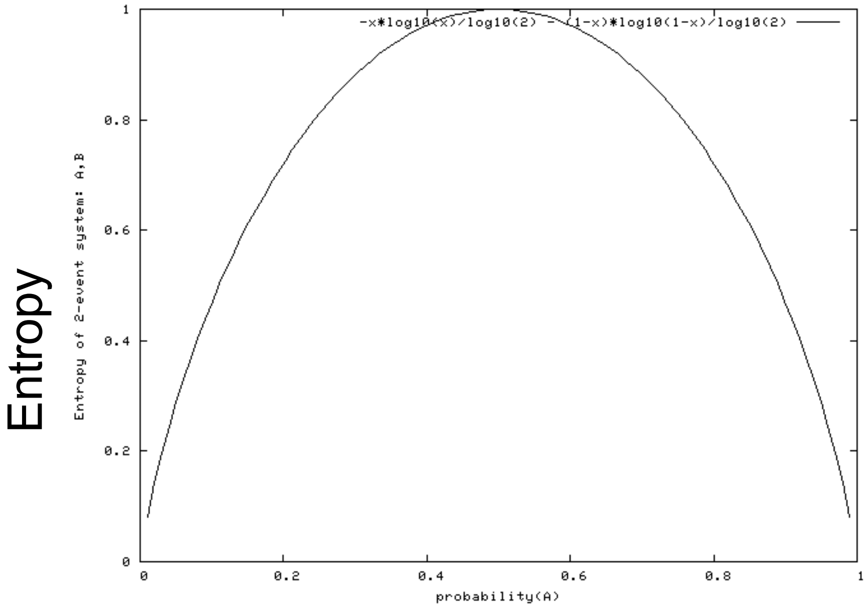
    (how much redundancy)
  - Theoretical minimum number of bits needed for maximum compression.

- Effectively, gives us an idea of the maximum compression -- *without producing the compressed output.*

Entropy: a way of measuring information

- Entropy H = -Σ_i (p_i x $\log_2 p_i$)
- Shown: entropy for 2-state system (x & y).
- H is maximal when probabilities are equal.
- Equal probabilities:
  don't know what to
  expect next!

What is the maximum
entropy for this
2-state system?



Entropy: a way of measuring information

- Entropy $H = -\Sigma_i (p_i \times \log_2 p_i)$
- What is the maximum entropy for a 4-state system, such as DNA?
- What kind of sequences have maximum entropy?
- What is the minimum entropy for a 4-state system, such as DNA?
- What kind of sequences have minimum entropy?

Models and entropy

- DNA models (4-state system):
  - *E. coli*:
    - p(T)=p(C)=p(A)=p(G) = 0.25
  - *P. falciparum*:
    - p(A)=p(T)=0.4;
    - p(C)=p(G)=0.1
- Entropy H = a measure of disorder/uncertainty
  - $H = -\Sigma_i (p_i \times \log_2 p_i)$
  - H(*E.coli* DNA)= 2.0
  - H(*P.falciparum* DNA)=1.7

Models and entropy

- $H = -\Sigma_i (p_i \times \log_2 p_i)$
  - H(*E.coli* DNA)= 2.0
  - H(*P.falciparum* DNA)=1.7
- What is entropy for DNA when:
  - p(A) = p(T) = 0.5; p(G) = p(C) = 0
  - P(A) = p(T) = 0.496; p(G) = p(C) = 0.004
  - P(A) = p(T) = 0.14; p(G) = p(C) = 0.36 (*Streptomyces coelicolor*)

Sliding window compression: detect repeats

- Example sequence:

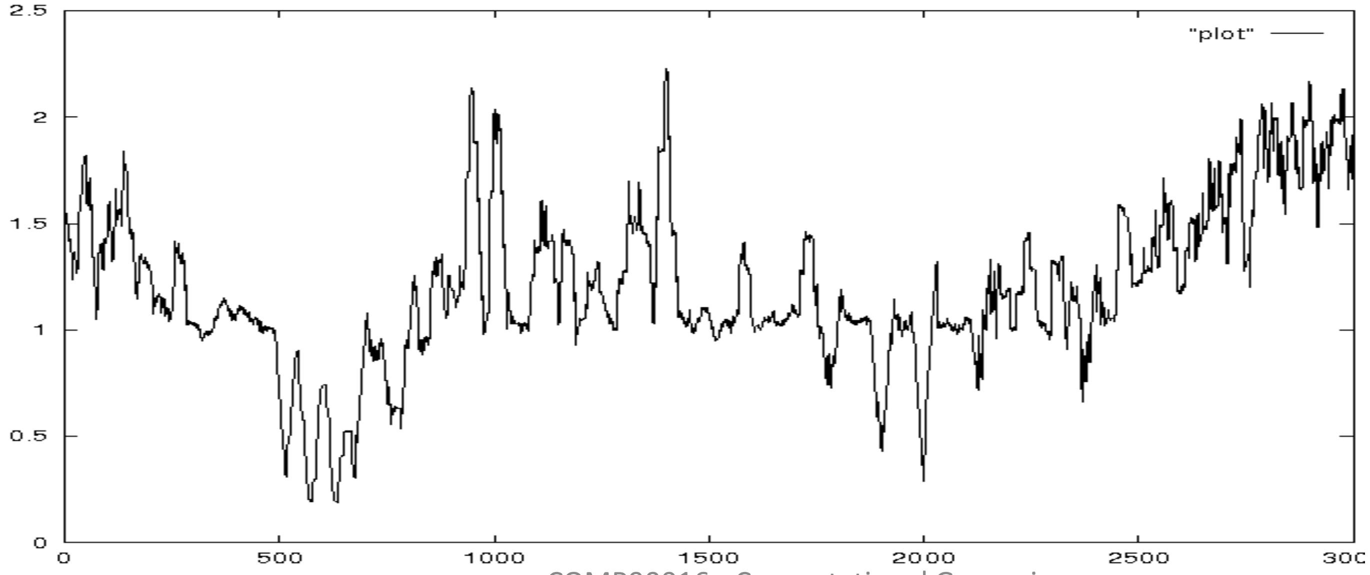
  GGAAATTGCCGCGTTGCCCAAATTGCCGCGCGTTCACA

  - Length = 37; repeat = 13

  - LZ compression encodes the repeat as:
    - signal-for-repeat
    - position of earlier copy
    - length of repeat
  - GGAAATTGCCGCGTTGCCC**0213**CACA

Sliding window compression usefulness

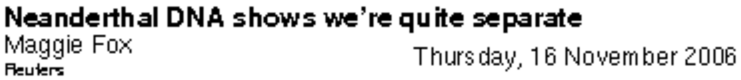
- Average bits/base will go down at the region of the repeat -- because there is less information.



Data from P. falciparum, chromosome 3, centromere region, Stern and Allison, 2001

How is information theory used in genomics?

- DNA is encoded as a string of symbols.
- Entropy measures the *information* in a string or substring.
- Use to:
  - Locate repeated/similar sequences motifs, related genes, homologs, pseudogenes.
  - Filter out low information regions before comparing sequences.
  - Separate DNA from different organisms.
  - Find different regions in DNA.



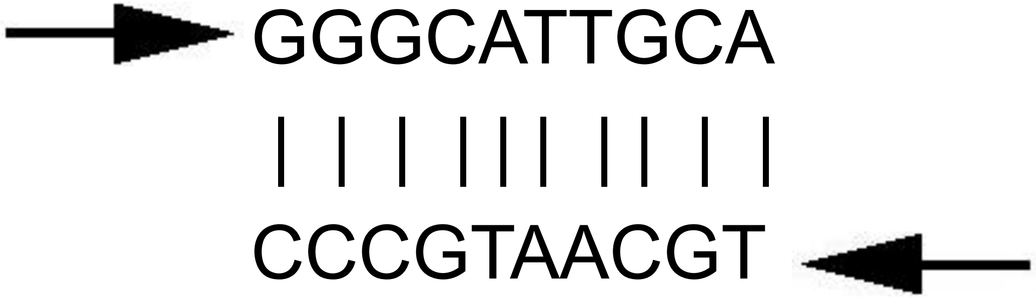
Neanderthal DNA shows we're quite separate
Maggie Fox
Reuters
Thursday, 16 November 2006

Sequence, structure, and function

- All macromolecules have:
  - Sequence
  - Structure
  - Function
- Bioinformatics connects sequence, higher order structure, and function.

DNA structure

- Primary structure (sequence):



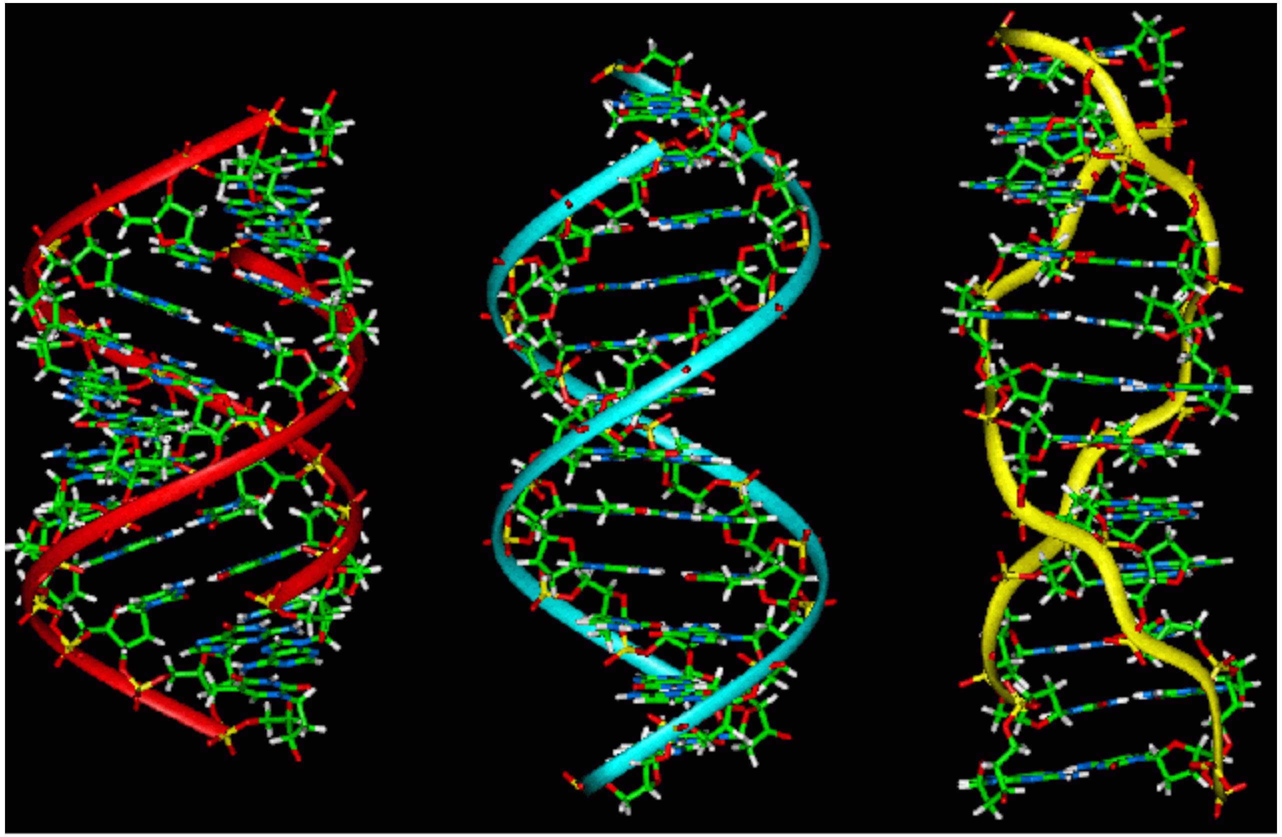
GGGCATTGCA

| | | ||| || | |

CCCGTAACGT

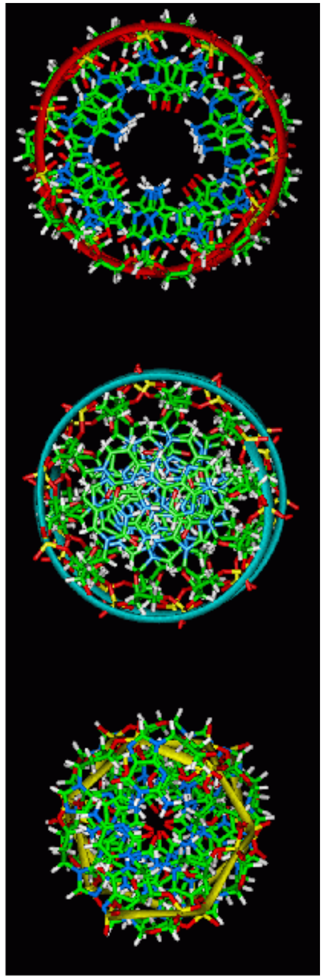
- Secondary structure
  - Watson-Crick helix, RNA folding
- Higher-order structure
  - 3-dimensional refinements of helix shape

DNA Structure



B-DNA A-DNA Z-DNA

image from http://www.answers.com



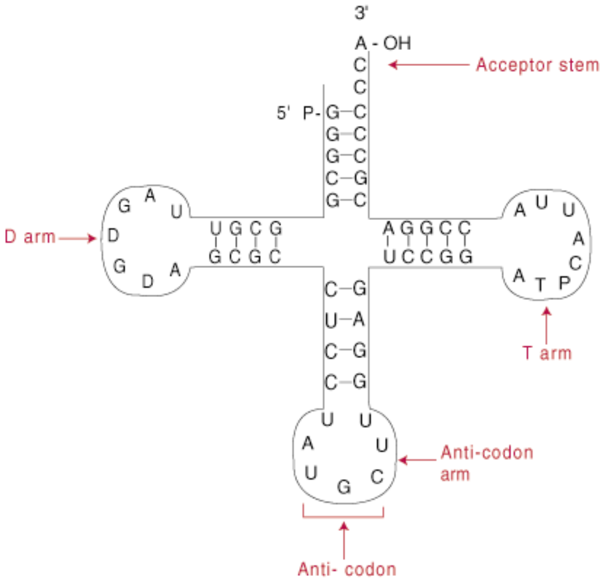
Top view (B=top)

RNA Structure

- Primary structure (sequence):

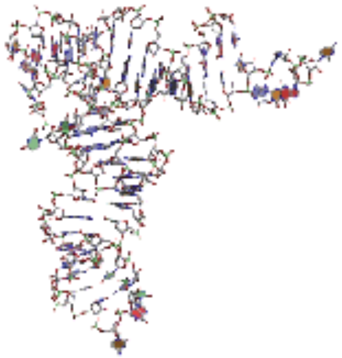
  GGGCGGCGUUA…

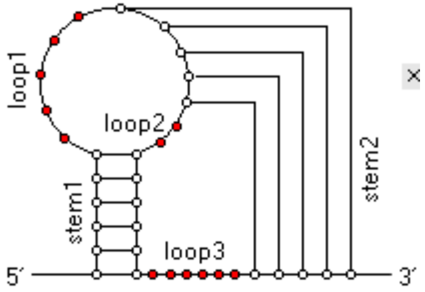
- Secondary structure (2-dimensional):

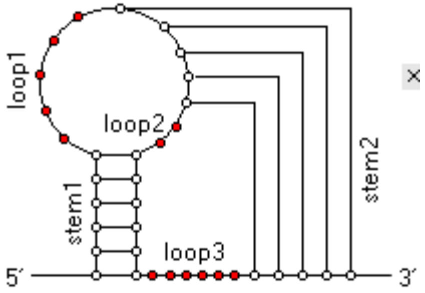


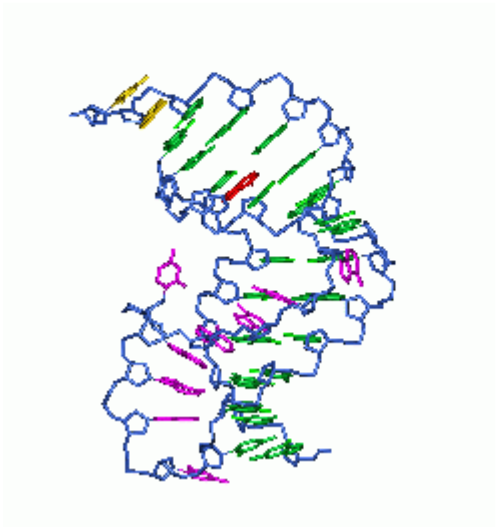
RNA Structure

- Tertiary structure (3-dimensional):









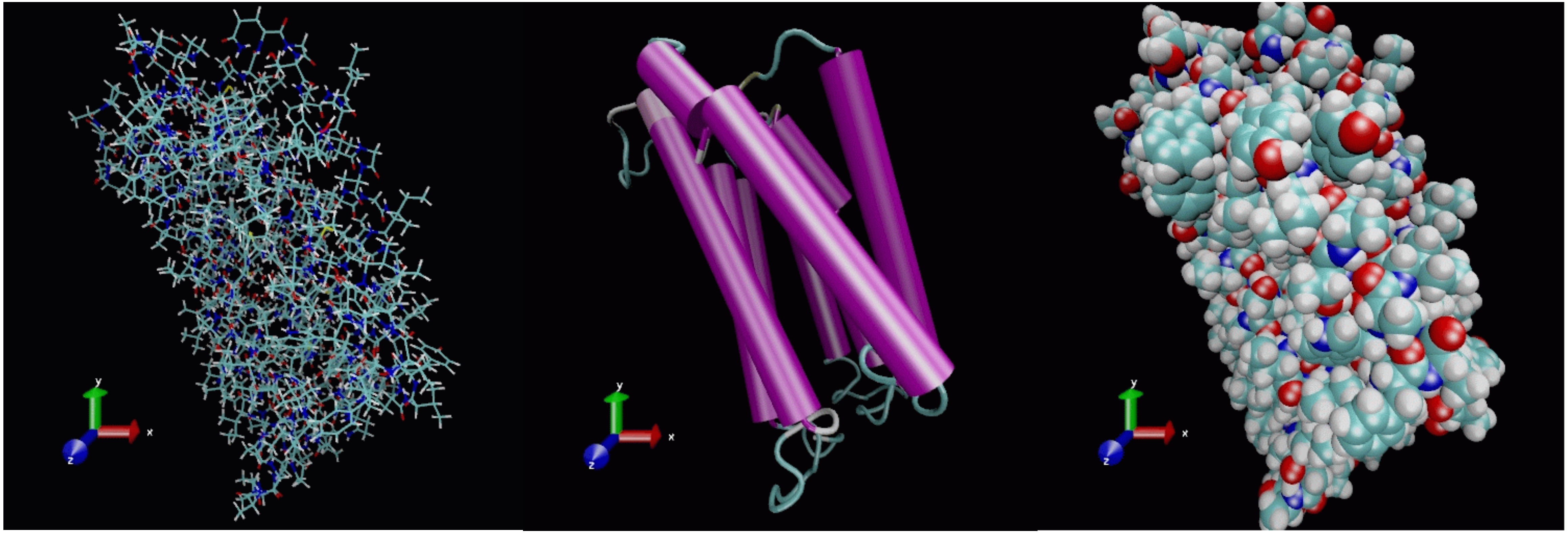
  - Extra-planar hydrogen-bonds (pseudoknots)
  - Weaker forces: van der Waals,

Protein Structure

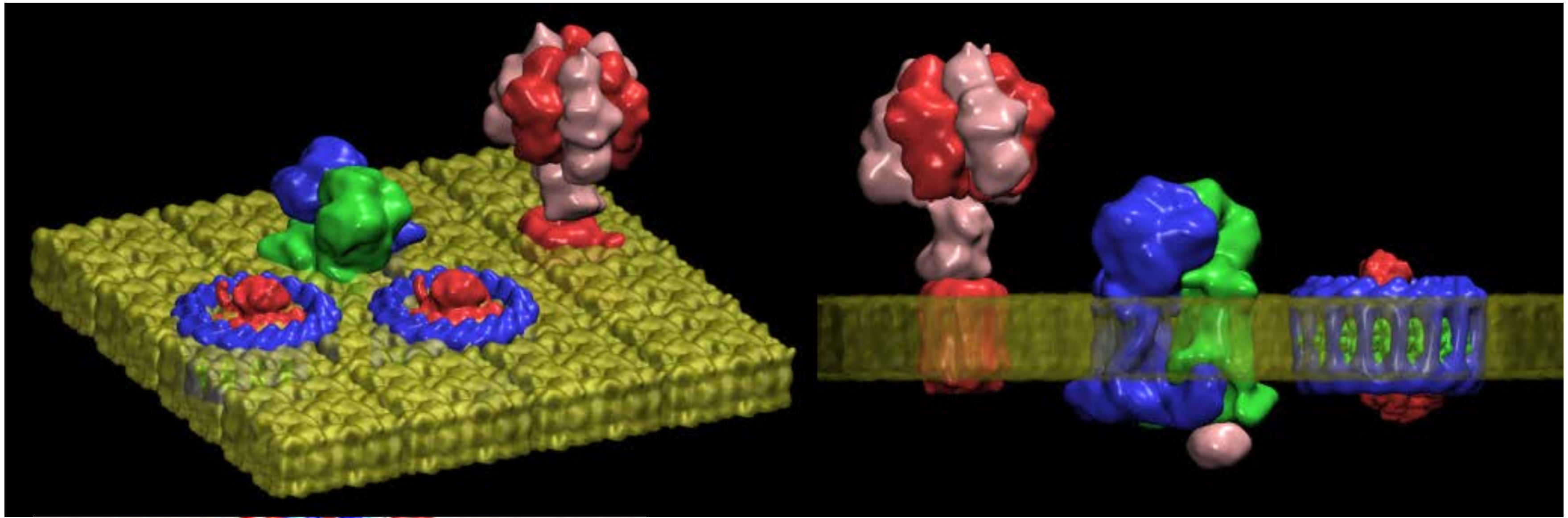
- Primary structure (sequence):

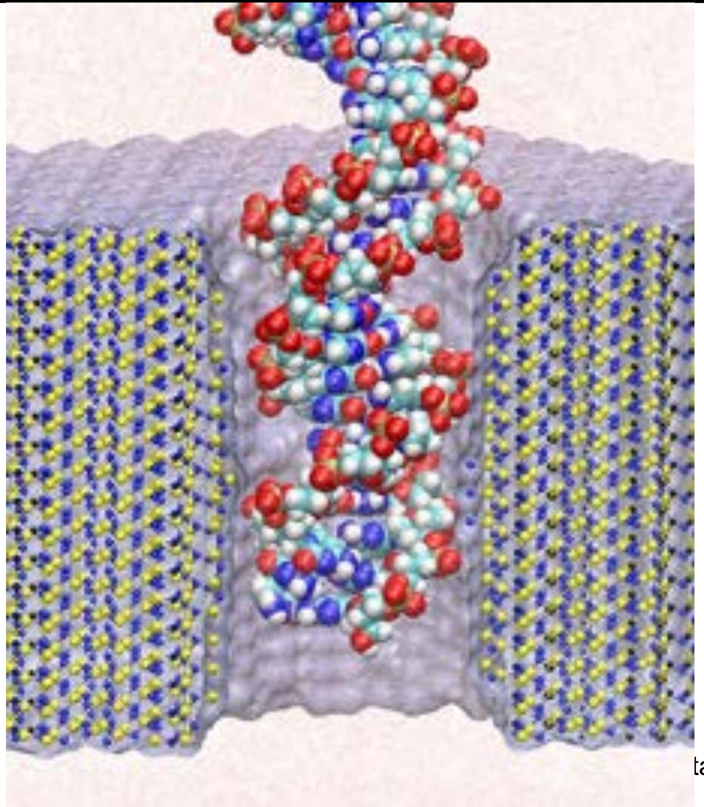
  MKVFLTYVKI… alphabet size 20
- Secondary structure (major features):
  - Helices (coils)
  - Sheets
  - Loops
- Tertiary structure (3-dimensional)
- Quaternary structure (subunits)

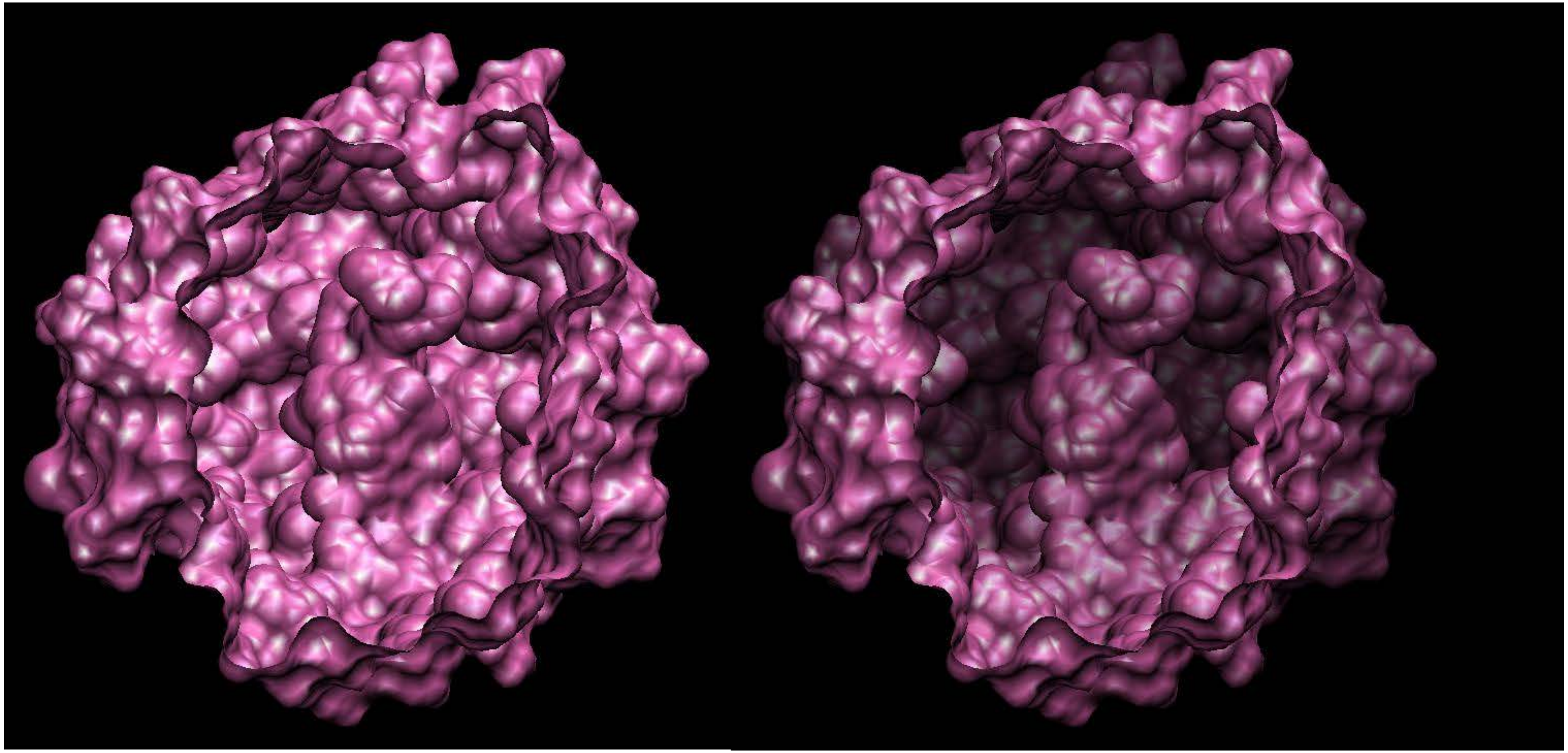
Protein Structure: Visualizations



University of Illinois, Urbana-Champagne, Computational Biophysics Group
http://www.ks.uiuc.edu/Research/vmd/allversions/repimages







Sequence, Structure, and Function

- Sequence determines 3-dimensional structure (mostly).
- Structure determines function (mostly).

- Study of sequences:
  - Analysis of components.
  - Comparison.
  - Structure prediction.
  - Genetic engineering.