CSE 397-497:
Computational Issues in Molecular Biology

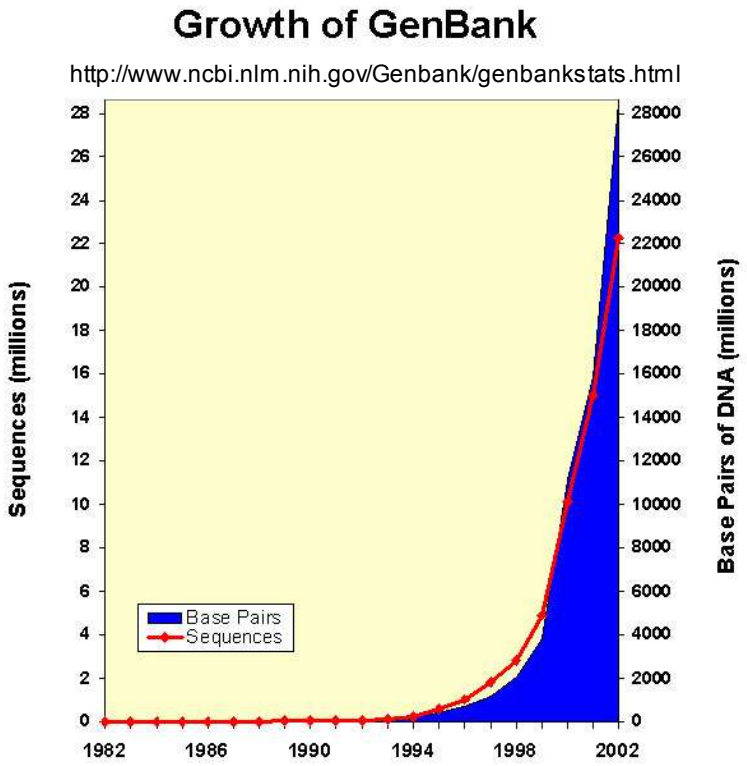
Lecture 1

Spring 2004

Motivation

"Biology easily has 500 years of exciting problems to work on."

Donald E. Knuth



Fortunately, no prior study of biology is needed for this course. We'll learn what we need.

This won't make you a biologist – you'll be a computer scientist who better understands this important new application area.

Synergies with: algorithms, databases, graphics, pattern recognition, machine learning, graph theory, robotics, etc.

Plan for today

1. Discussion of course structure and logistics:
   - requirements,
   - grading,
   - resources.

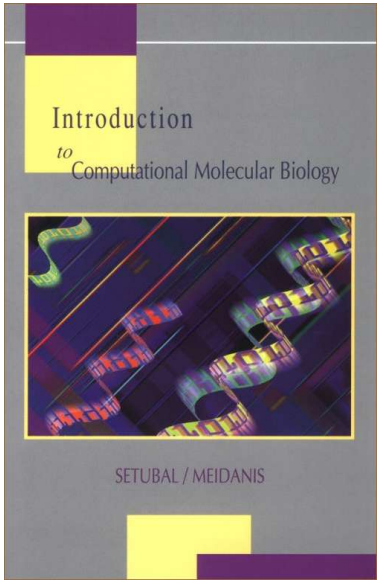
2. Brief introduction to molecular biology.

3. Assigned readings for next time.

Note – I will be away this coming Thursday. Our next class will be Tuesday, January 27.

Instructor

Professor Daniel Lopresti
PL 404B ~ x85782 ~ dal9@lehigh.edu
Office hours: 2:00 – 4:00 W (or by appt)



Text: *Introduction to Computational Molecular Biology* by João Setubal and João Meidanis.

While your text covers most topics rather well, we will supplement it extensively.

Basic ground-rules

Collaboration is prohibited unless I explicitly state otherwise.

Cell phones, wireless email, etc. must be off during class.

University Policy on Disabilities:

"If you have a disability for which you are or may be requesting accommodations, please contact your professor and the Office of Academic Services, Room 212, University Center or call (610-758-4152) as early as possible in the semester. University policy states that you must notify your professor seven (7) days prior to the exam."

Course requirements

CSE 397-497 will be run as a quasi-seminar course.

This means:

- Some of the material will be drawn from research papers.
- Class participation is vital and will weigh heavily in grading.
- You will be required to prepare and present one lecture during the semester (I will work closely with you on this).
- Final project (or paper) due at the end of the semester.
- No homework assignments. However, you are expected to come to class prepared (i.e., having done the readings).

In addition, students taking the course as CSE 497 must also serve as "scribe" for one student lecture.

Course requirements

To reiterate:

- Do all of the assigned readings.
- Attend every class (barring extreme circumstances).
- Be prepared to discuss the material.

Based on how the course goes – whether or not it really runs as a seminar with active student participation – I reserve the right either to hold a final exam or to call it off.

poor student participation in class $\Rightarrow$ final exam

good student participation in class $\Rightarrow$ no final exam

I will make this determination no later than one month before the scheduled date of the final.

Course requirements

Class attendence / participation means:

- Showing up having done the readings, asking questions, making comments, and contributing new insights.

Presenting a lecture means:

- Meeting with me several times in advance and then presenting a class lecture and leading the discussion.

Final project or paper means:

- Implementing and testing one of the algorithms we study, or writing a 15-page paper on a topic relating to the course.

Scribing a student lecture (for CSE 497 students) means:

- Taking notes during the lecture in question and later editing them so that they can serve as a useful reference.

Course grading

Class attendence / participation = 100 points.

Lecture = 25 points (preparation) + 100 points (delivery).

Final project or paper = 100 points.

Scribe* = 25 points.

Final exam (if we need it) = 100 points.

** Note that CSE 397 and CSE 497 point totals will be different and each will be curved separately.*

Student lectures

General procedure and schedule:

- You choose general topic area (next Thursday, Jan. 29).
- I assign date you will lecture (next Friday, Jan. 30).
- First meeting with me (two weeks before your lecture date): discuss material you will present and describe your plan.
- Second meeting with me (one week before your lecture date): show me your near-final lecture and discussion topics.
- Third meeting with me (one day before your lecture): final run-through.

Use of PowerPoint for slides is encouraged, but not required.

Course topics

Topic areas we will hopefully have time to cover:

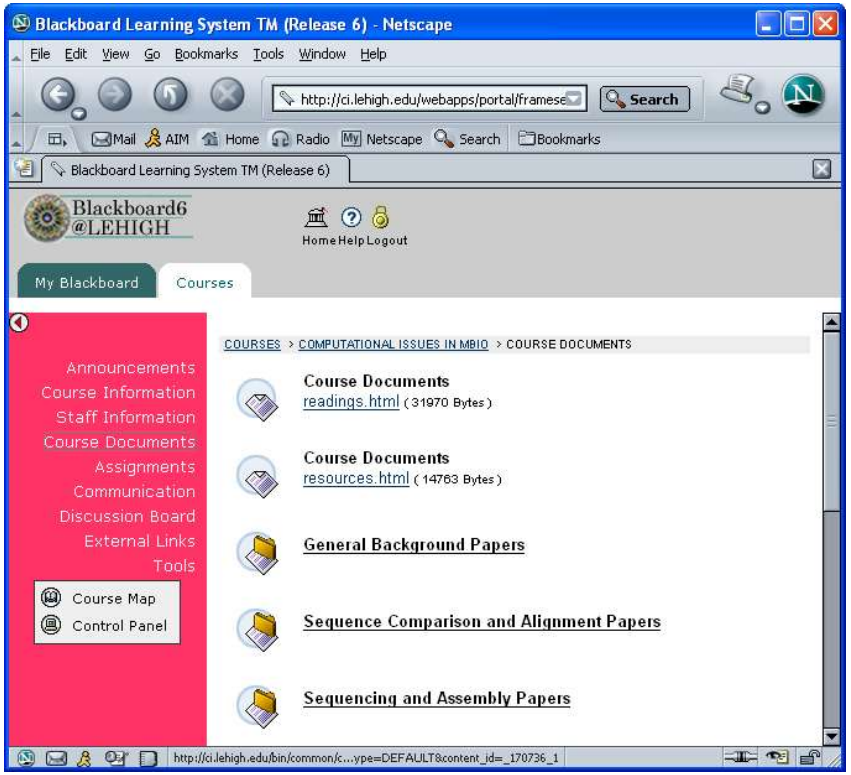
- introduction to molecular biology for computer scientists,
- pairwise sequence comparison & alignment,
- multiple sequence alignment,
- sequencing and sequence assembly,
- physical mapping of DNA,
- advanced topics: DNA microarrays, genome rearrangements, RNA and protein structure prediction, etc.

We will mostly follow your textbook, but supplement it by:

- returning to original versions of seminal papers,
- studying current papers and topics too recent for text,
- experimenting with tools that use the algorithms in question.

Lehigh Blackboard system

CSE 397-497 is registered on Blackboard: http://ci.lehigh.edu



- Much supplementary material will be posted.
- E.g., pointers to resources, papers you are expected to read, etc.
- Lecture slides will be posted (a little in advance, I hope).
- Remember to check Blackboard frequently!

Some examples of resources listed on Blackboard

1. Background material

National Human Genome Research Institute Talking Glossary of Genetic Terms

http://www.genome.gov/page.cfm?pageID=10002096

2. Databases

GenBank

http://www.ncbi.nlm.nih.gov/Genbank/index.html

3. Research groups

University of Pennsylvania Center for Bioinformatics

http://www.pcbi.upenn.edu/

4. Tools

MolBiol.Net Directory

http://www.molbiol.net

5. Similar courses at other universities

Carnegie-Mellon: Computational Molecular Biology and Genomics (15-856)

http://www-2.cs.cmu.edu/~durand/03-711/

Introduction to molecular biology

Life as we know it is largely determined by two classes of molecules, *proteins* and *nucleic acids*.

Proteins play a key role in almost all biological activities in our bodies. *Structural proteins* form the building blocks for our tissues. *Enzymatic proteins* act as *catalysts* for chemical reactions that otherwise would occur far too slowly to be useful.

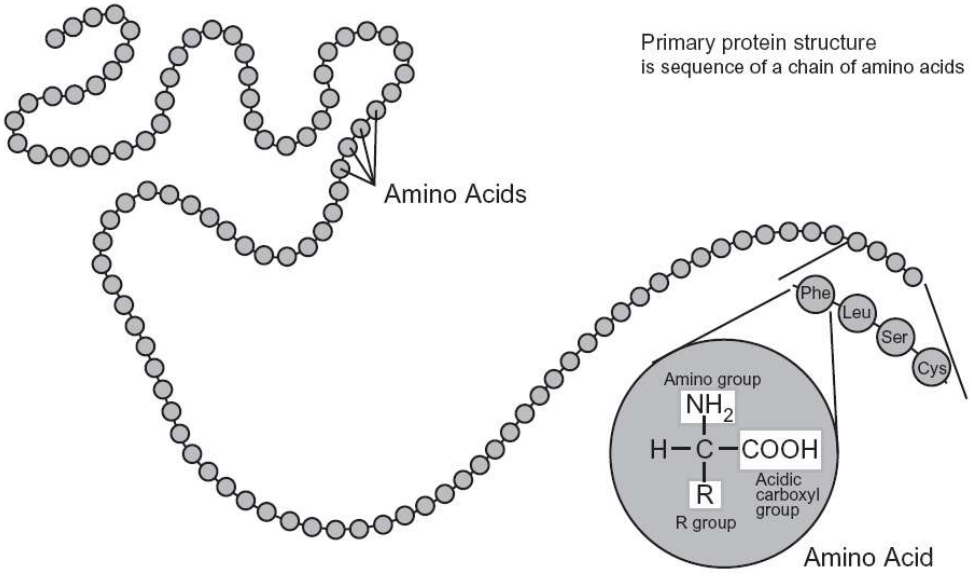
There are two basic kinds of nucleic acids: *ribonucleic acid* (*RNA*) and *deoxyribonucleic acid* (*DNA*).

DNA is sometimes called the "blueprint of life" because it encodes the information necessary to manufacture proteins.

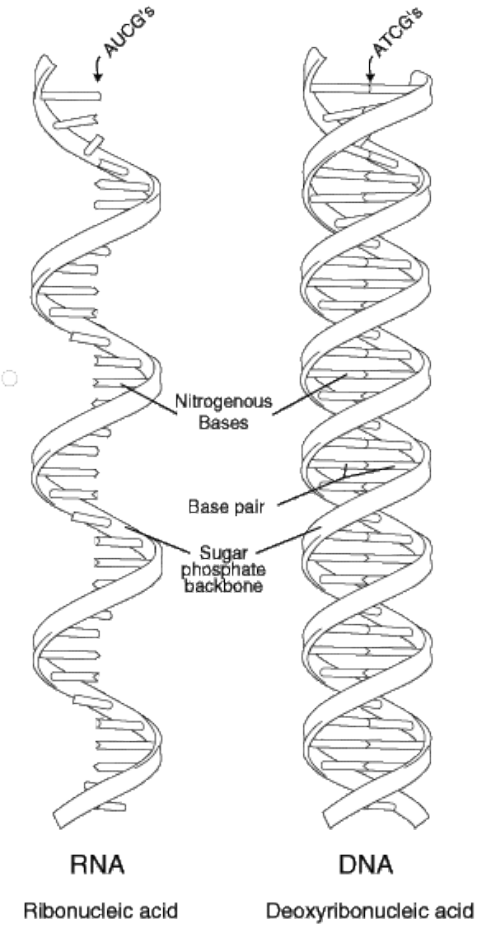
RNA performs several different functions connected to taking this information and making use of it.

The sequence nature of biology

Both proteins and nucleic acids are *macromolecules*, long chains of much simpler molecules.



In the case of proteins, these basic building blocks are *amino acids*.



In DNA and RNA, they are *nucleotides*.

http://www.accessexcellence.org/AB/GG/aminoAcid.html
http://www.accessexcellence.org/AB/GG/rna.html

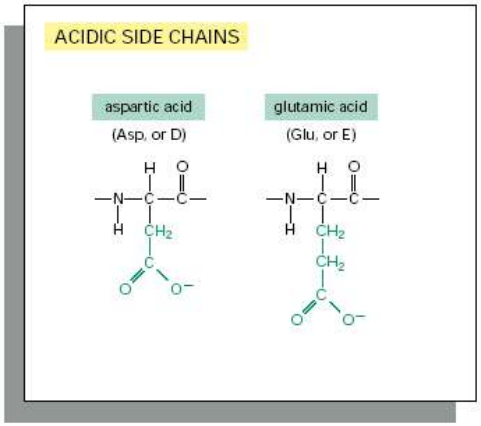
Proteins and amino acids

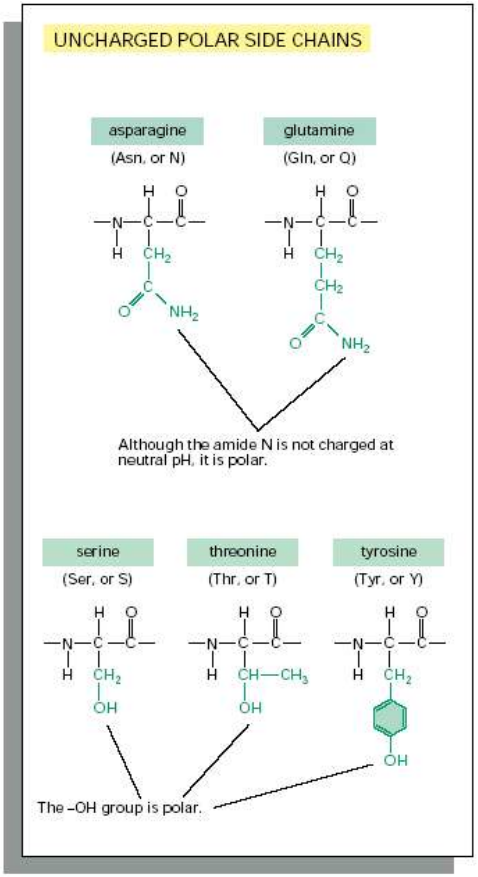
There are 20 different kinds of amino acids.

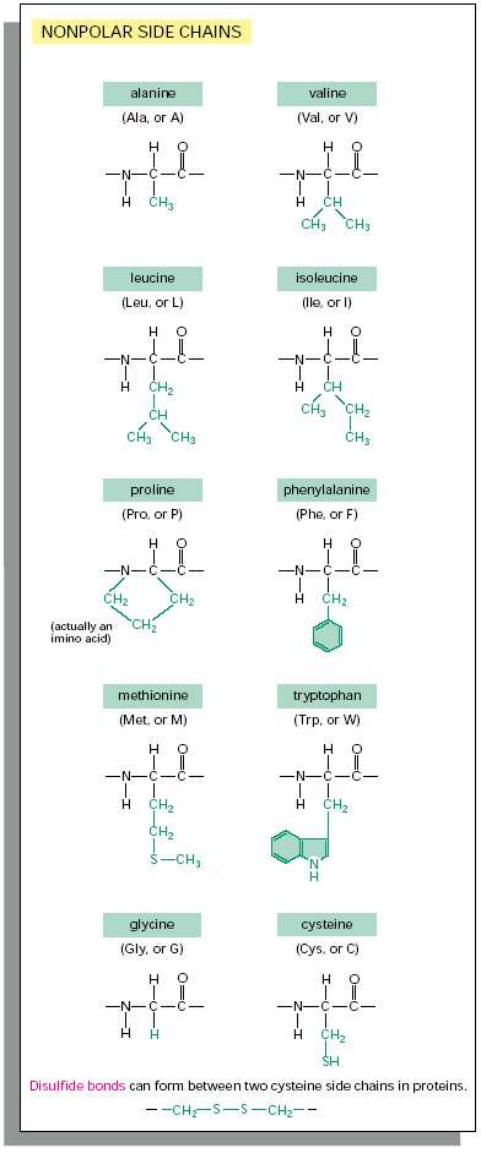
Each has identical:

- central carbon atom (C),
- single hydrogen (H),
- amino group (NH_2),
- carboxy group (COOH).

Side chain is difference.

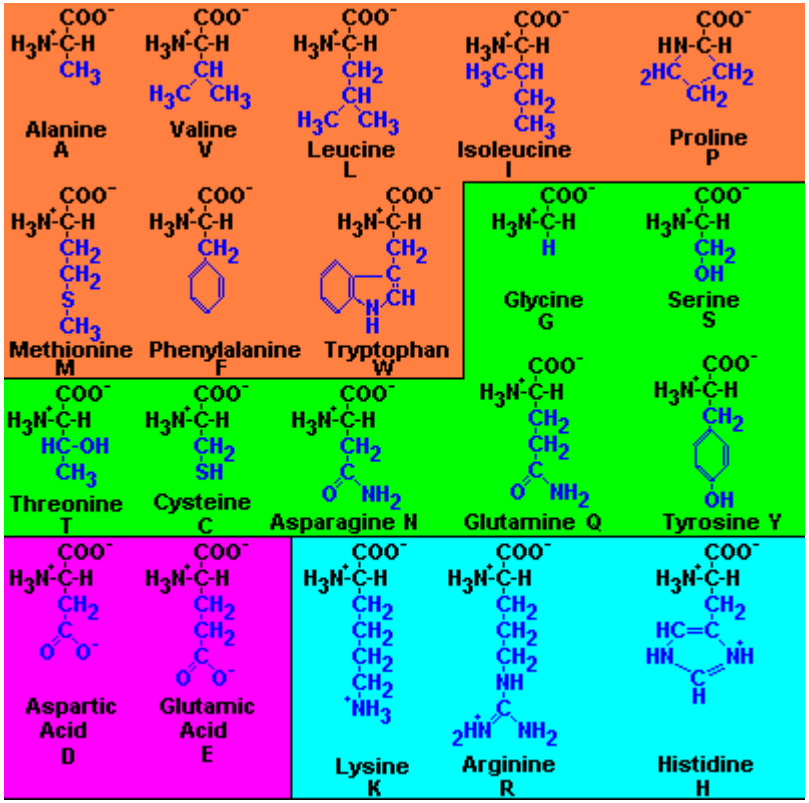






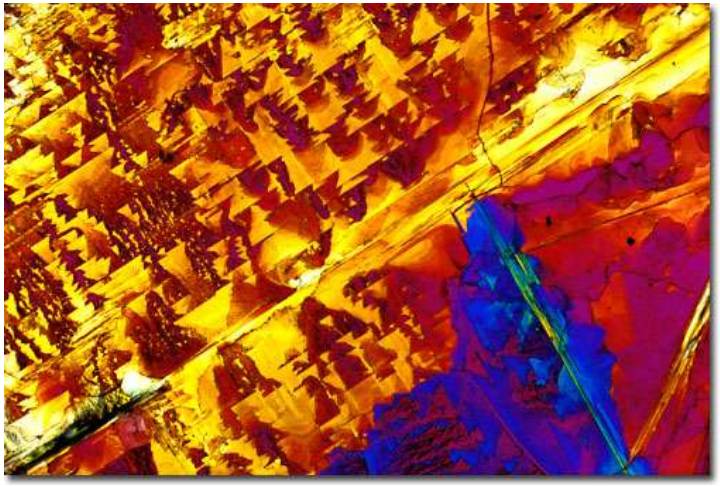
http://www.accessexcellence.org/AB/GG/aminoAcids2.html

Amino acids



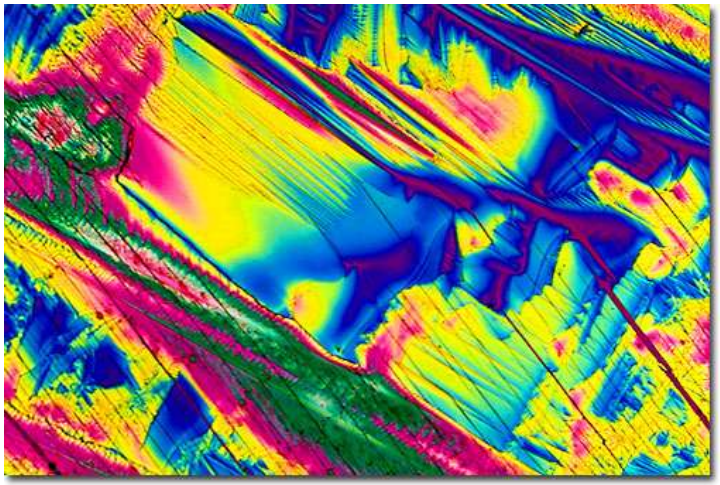
Amino acids are also often abbreviated using three-letter codes (e.g., Alanine = Ala).

http://www.people.virginia.edu/~rjh9u/aminacid.html



Photomicrograph of Alanine

http://micro.magnet.fsu.edu/aminoacids/pages/alanine.html

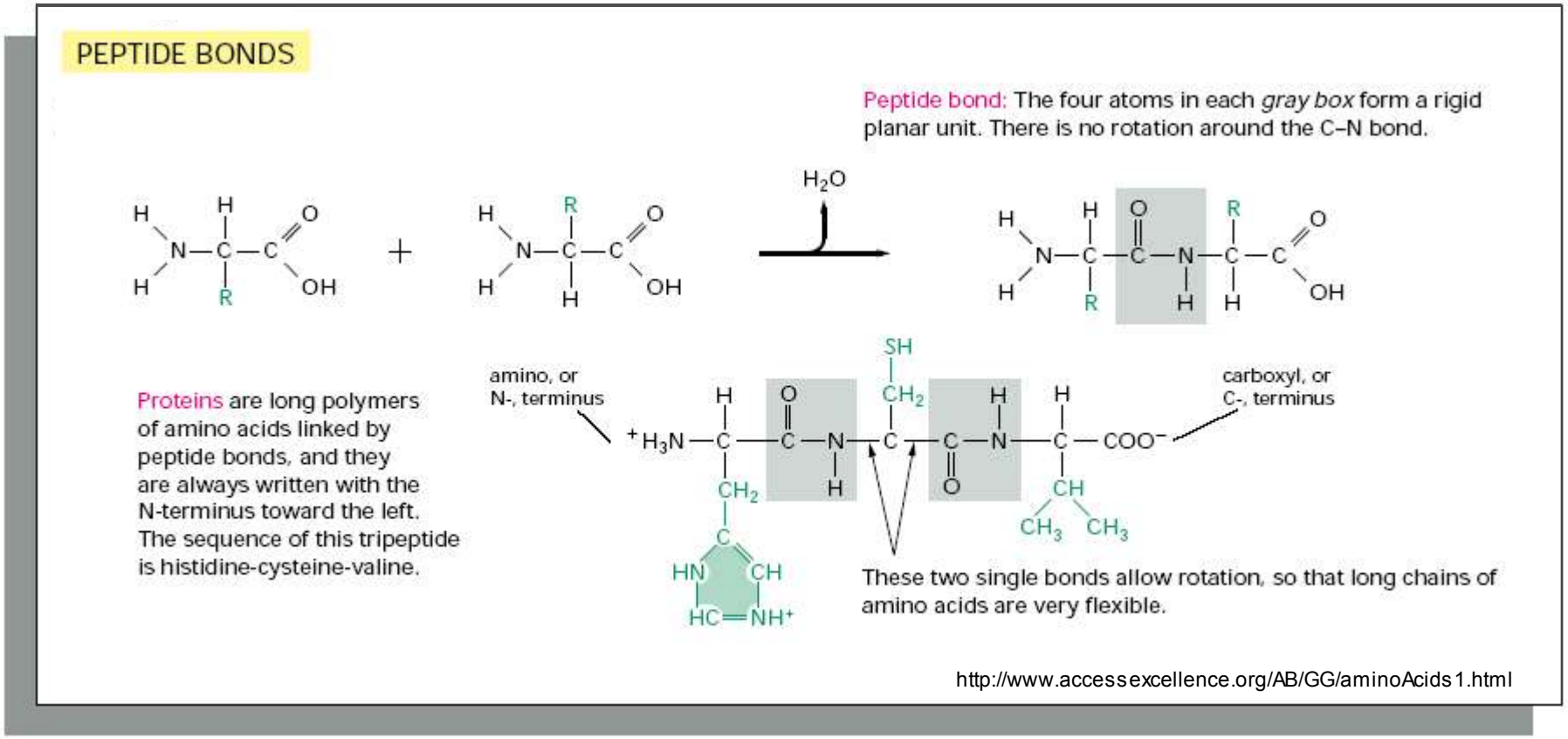


Photomicrograph of Glutamine

http://micro.magnet.fsu.edu/aminoacids/pages/glutamine.html

Amino acid bonding

Amino acids join together to form proteins via peptide bonds.

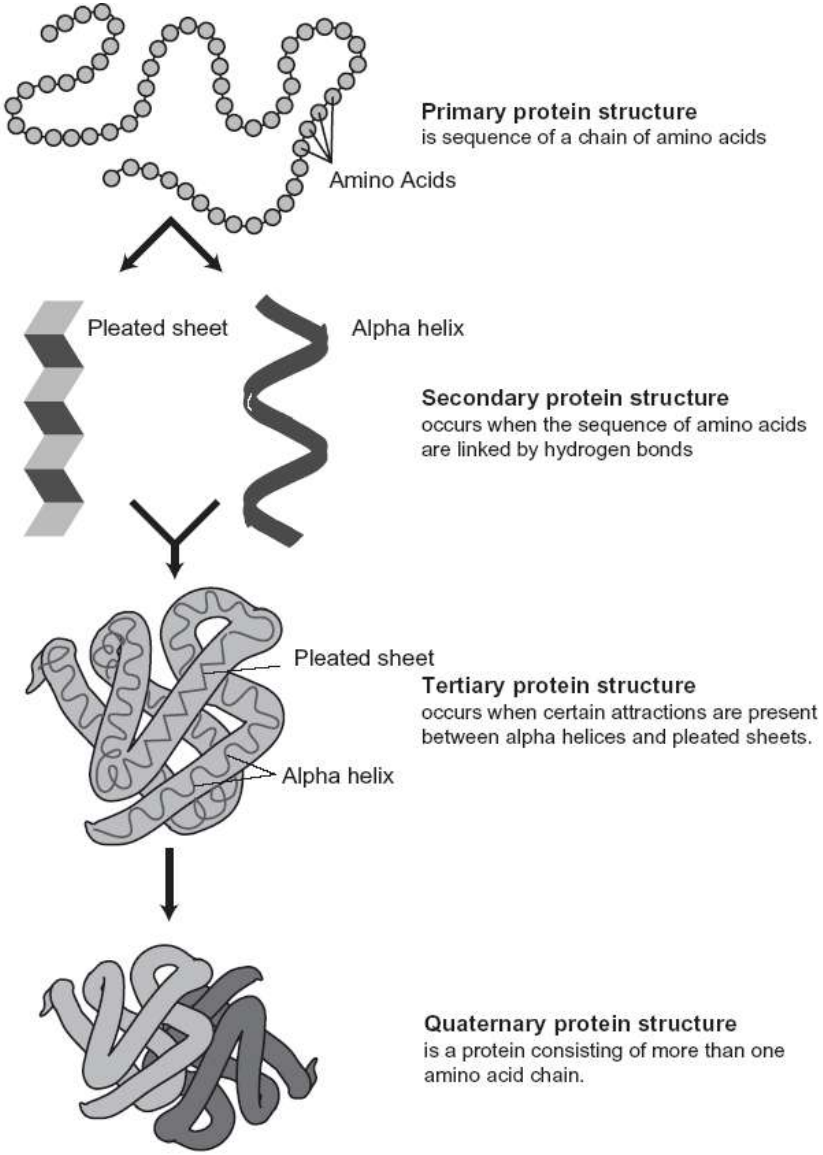


Since a water molecule is liberated, the chain is really composed of *residues* of amino acids.

Protein structure

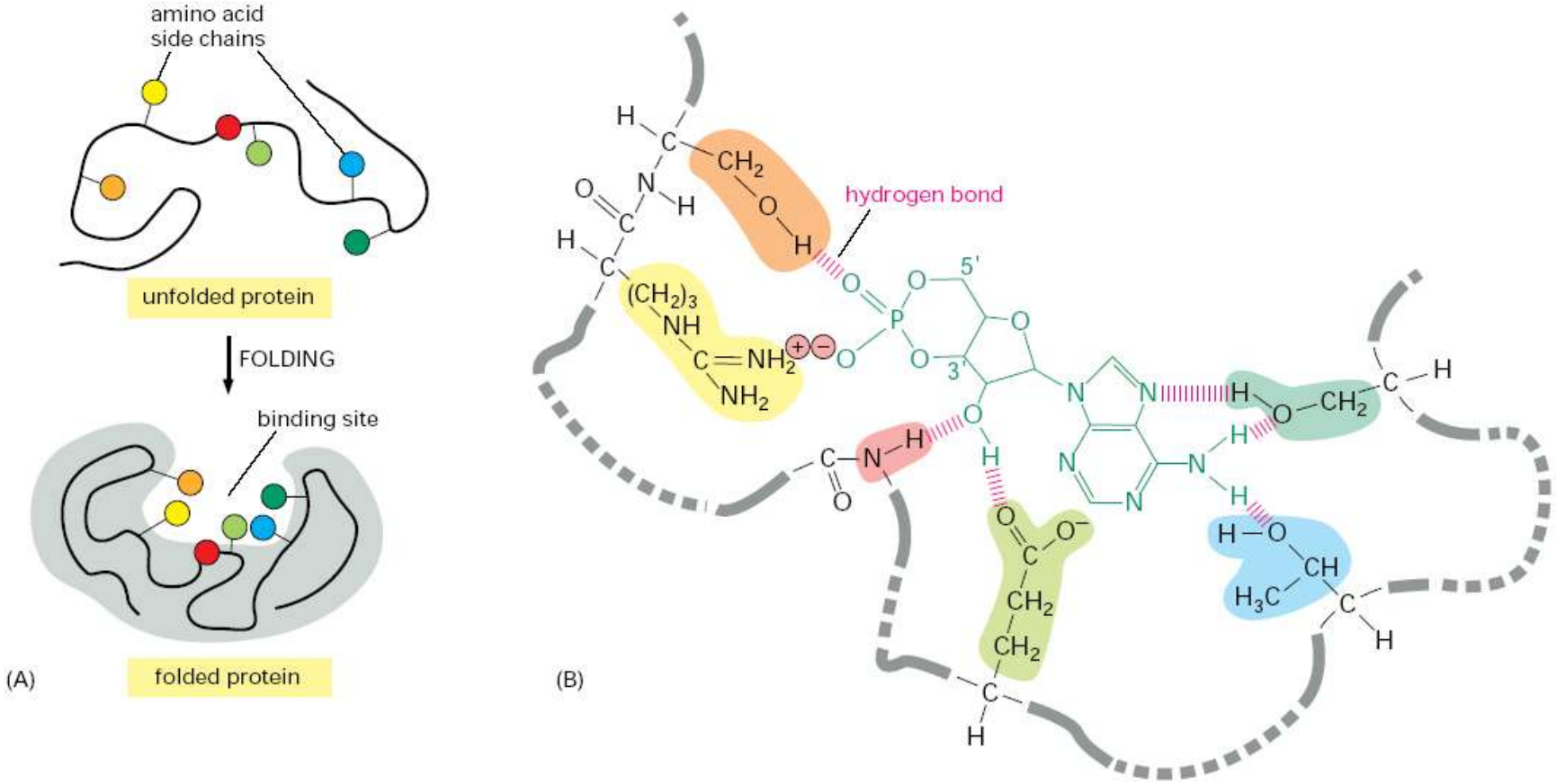
While we will often employ a one-dimensional abstraction in this course, real life is considerably more complex.

Proteins *fold* in three dimensions, which determines how they bind to other molecules and, hence, what function they perform.



http://www.accessexcellence.org/AB/GG/protein.html

Protein binding sites



The folding of the chain creates a crevice on the protein surface. This contains a set of amino acid side chains disposed in such a way as to bond only with certain molecules.

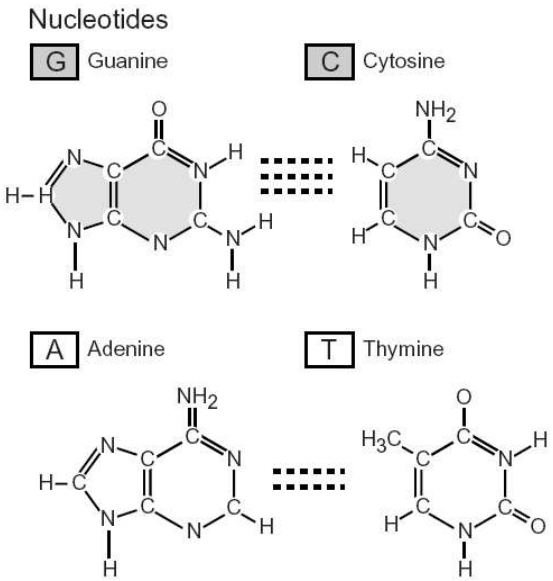
http://www.accessexcellence.org/AB/GG/garland_PDFs/Fig_5.24.pdf

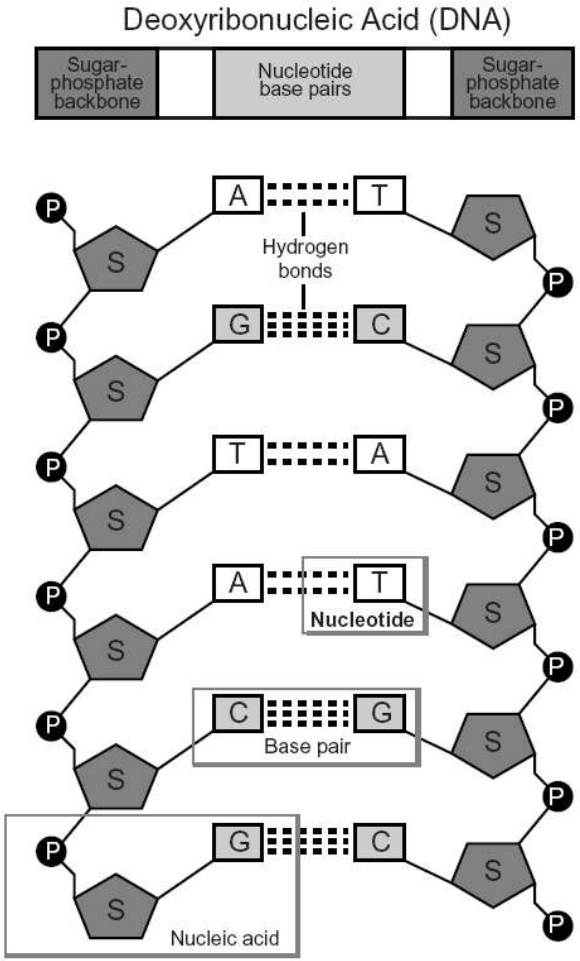
Deoxyribonucleic acid (DNA)

Like proteins, DNA is a macromolecular chain. DNA is double-stranded, however.

Each building block consists of:

- sugar molecule,
- phosphate residue,
- one of four bases (nucleotides).





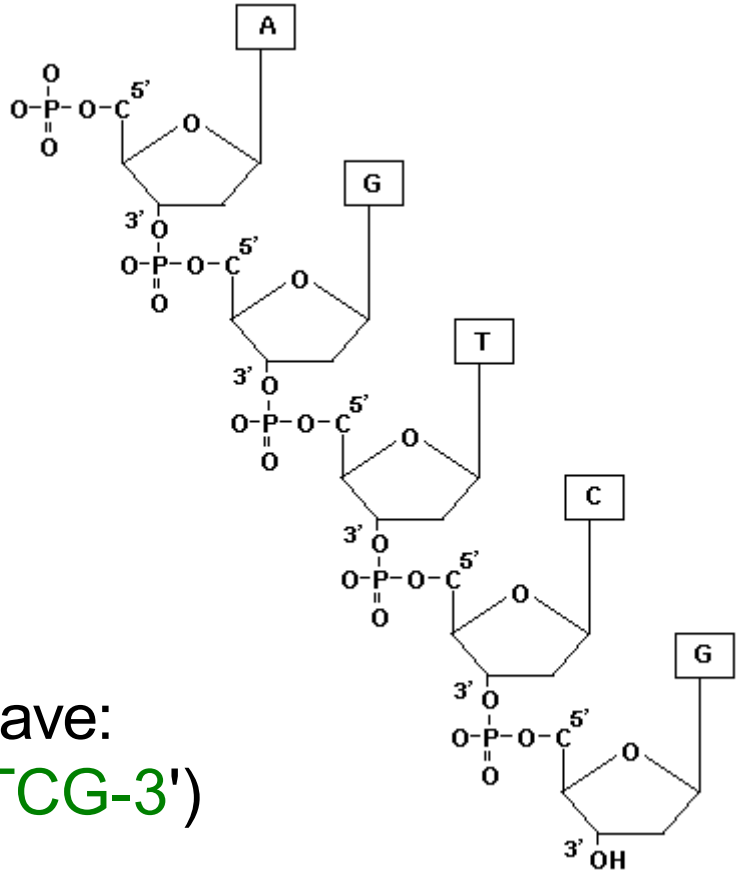
http://www.accessexcellence.org/AB/GG/nhgri_PDFs/nucleotide2.pdf

DNA reading order

The sugar backbone molecule in DNA consists of five carbon atoms. These are numbered from 1' through 5'.

On one end of the DNA macromolecule, there is a free phosphate (P) group. This is the end that corresponds to the 5' carbon atom.

By convention, DNA sequences are recorded in this order, reading from the 5' end to the 3' end.



So here, for example, we have:

AGTCG (or 5'-AGTCG-3')

http://avery.rutgers.edu/WSSP/StudentScholars/project/archives/onions/orien.html

DNA structure

The two DNA strands bind together to form a *double helix* structure first described by Watson and Crick in 1953. This is a right handed double helix, with about 10 nucleotide pairs per helical turn.

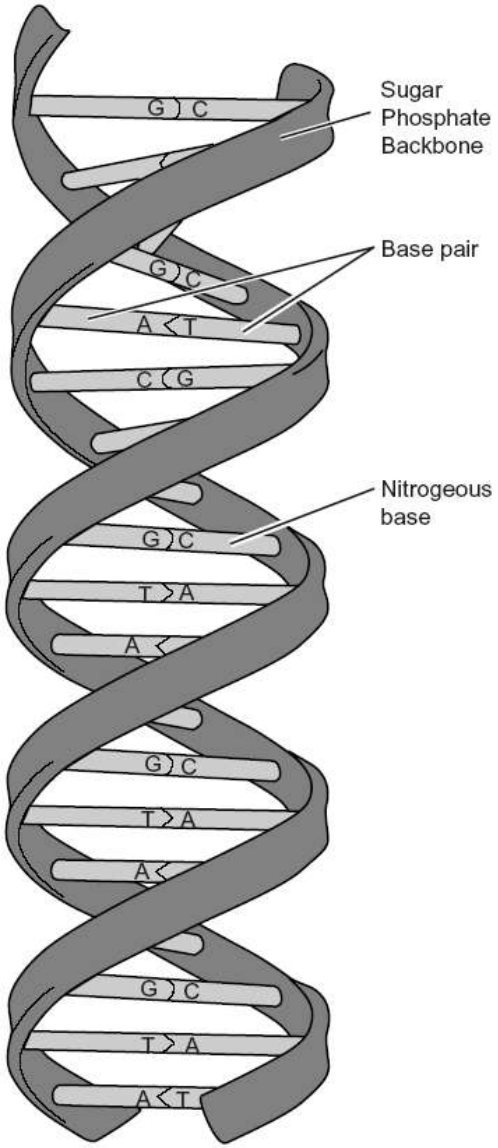
The bases always bind in complementary fashion: adenine (A) with thymine (T), and guanine (G) with cytosine (C).

Hence, one strand will be the *reverse complement* of the other:

GGACTAGTA →

reverse ... ATGATACAGG →

complement ... TACTATGTCC



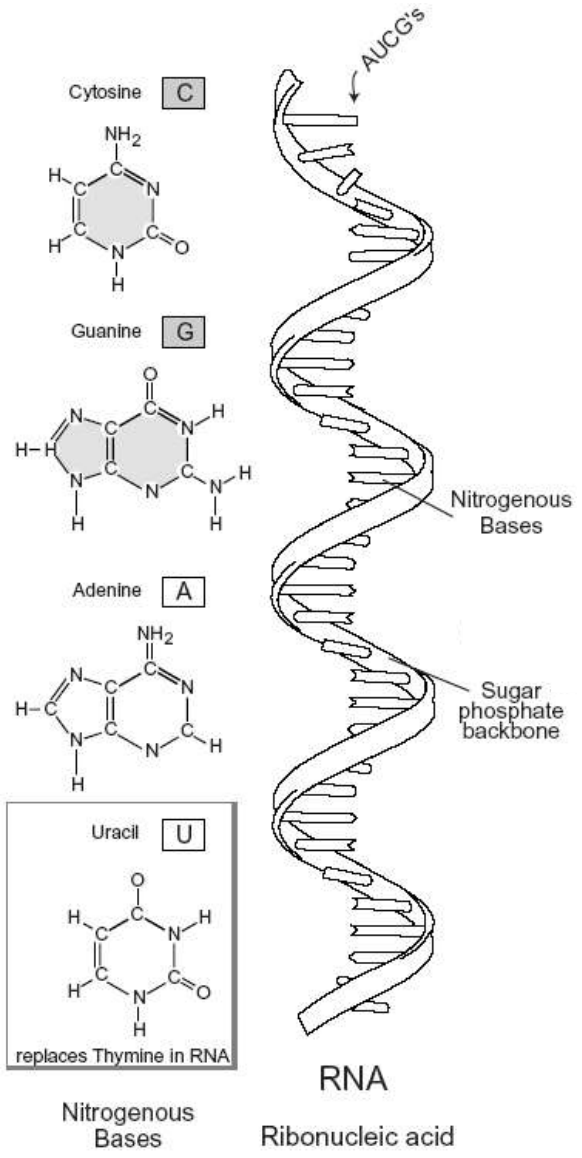
http://www.accessexcellence.org/AB/GG/nhgri_PDFs/dna.pdf

DNA and RNA

The fact that the two DNA strands are reverse complements isn't just a curious coincidence. This is the mechanism DNA uses to replicate itself, so it is, in fact, the basis of life.

RNA is somewhat similar to DNA with several notable differences:

- backbone sugar is ribose instead of 2'-deoxyribose,
- uracil (U) replaces thymine (T),
- usually found single-stranded; it does not form double helix,
- performs different function.



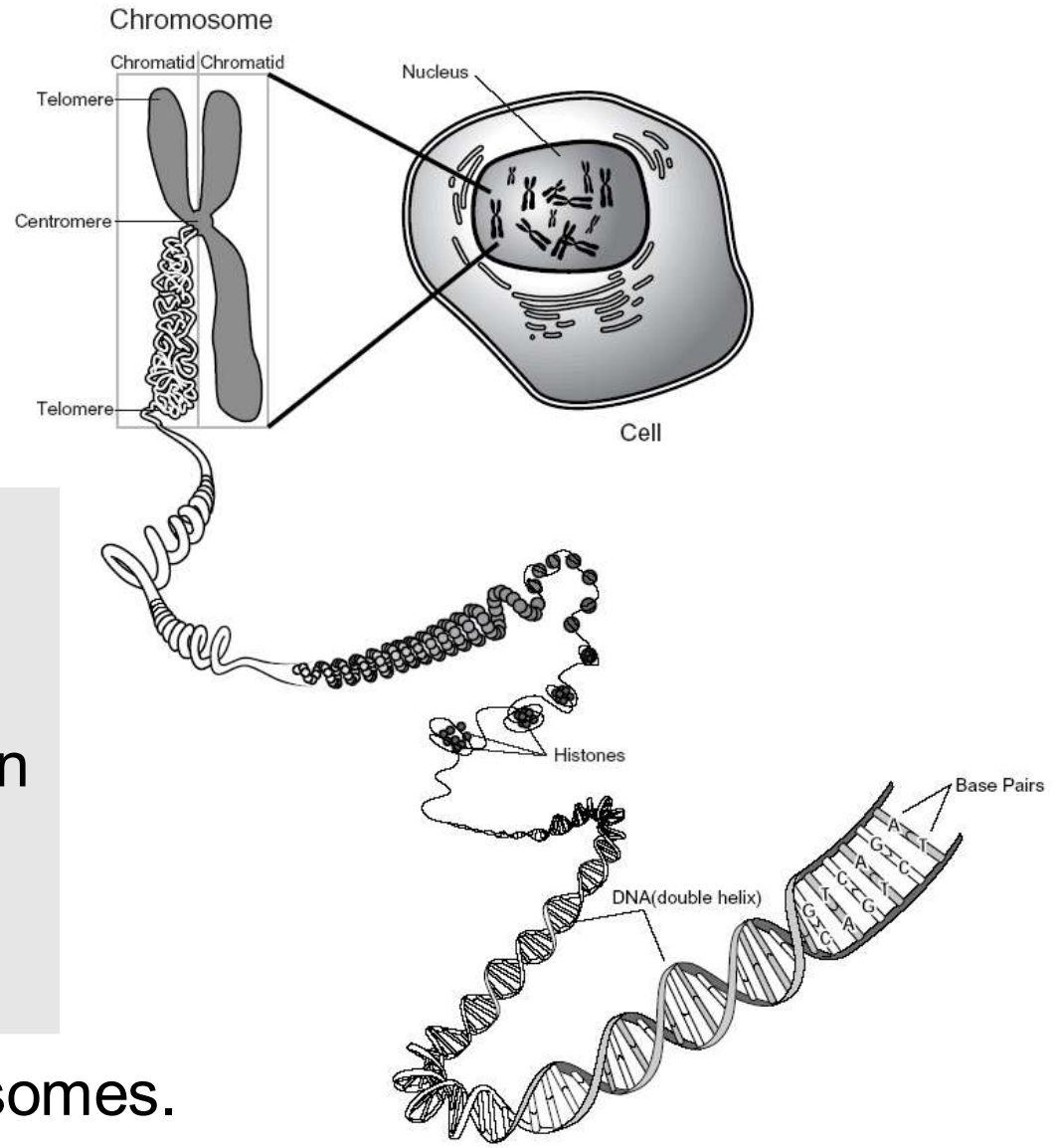
http://www.accessexcellence.org/AB/GG/nhgri_PDFs/rna2.pdf

Chromosomes

As we noted earlier, DNA encodes the information needed to manufacture proteins. Now we will see how that works.

A *chromosome* is a long DNA molecule. Most organisms have relatively few chromosomes (tens) in comparison to the total number of base pairs in their DNA (billions).

Humans have 46 chromosomes.



http://www.accessexcellence.org/AB/GG/chromosome.html

Genes

Genes are contiguous stretches along a DNA molecule that encodes the information necessary to build one protein (or, in some cases, one RNA).

What would this entail?

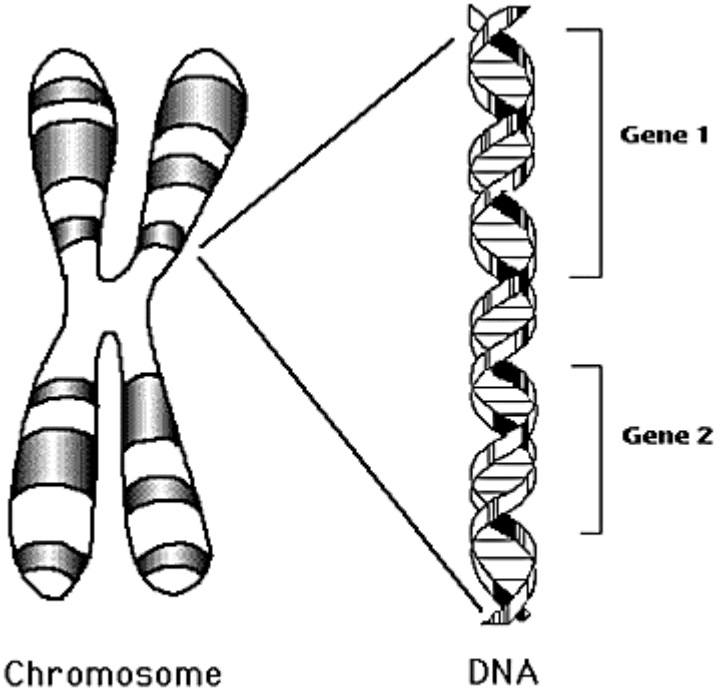
Well, we need a way to specify which one of the 20 amino acids to use at each position in the protein.

But there are only 4 nucleotides in DNA. How many, at a minimum, do we need to code for an amino acid?

$4^1 = 4$

$4^2 = 16$

$4^3 = 64$ ... that's enough!



Genes

http://www.accessexcellence.org/AB/GG/genes.html

The Genetic Code

Each group of 3 consecutive nucleotides is called a *codon*.

Since the actual synthesis of proteins is handled by an RNA molecular known as *messenger RNA* (or *mRNA*), this mapping is usually written using U instead of T.

So, for example,

CCA → Pro (Prolene)

AAA → Lys (Lysine)

This is known as the *genetic code*. It is called "universal" because it covers nearly all life forms we know of, including humans, plants, fungi, archaea, bacteria, and viruses.

http://www.people.virginia.edu/~rjh9u/code.html

The Genetic Code

Some points to ponder:

- There are 64 possible codons, but only 20 amino acids. Hence, the genetic code is redundant.

  This might seem inefficient, but in truth it is a brilliant piece of evolution (think of error correcting codes).

- There appear to be several kinds of ambiguity in mapping from DNA sequences to protein sequences:

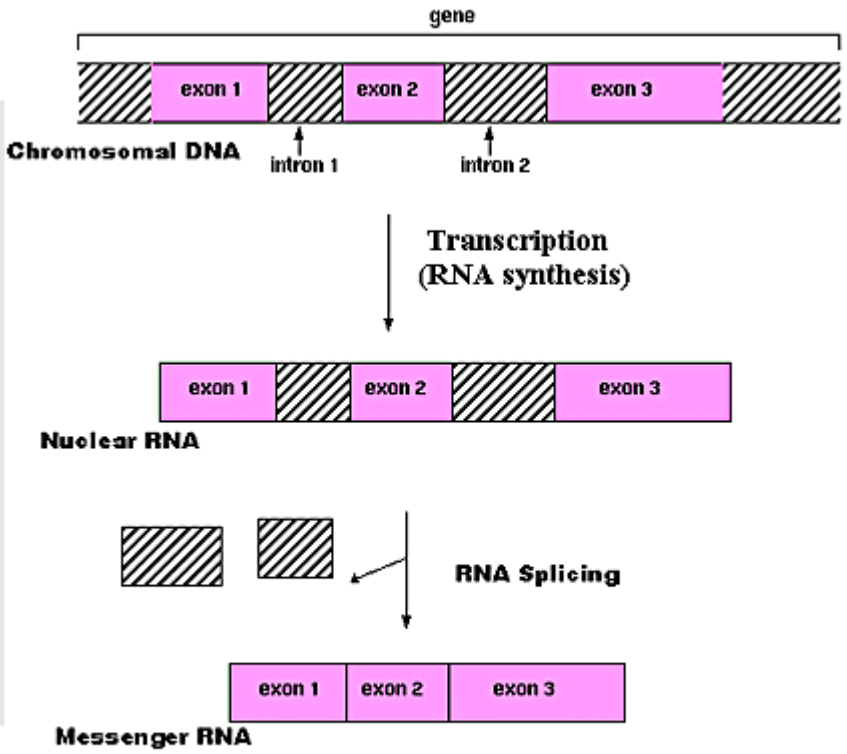
  (1) The two DNA strands are reverse complements. Which do we read?

  (2) It takes 3 consecutive nucleotides to code for an amino acid. How do we know which nucleotide is the first one in the first group of 3 (i.e., where do we start decoding)?

Transcribing DNA to RNA

To use the information encoded in DNA to make a protein, it is first *transcribed* into RNA.

Process begins with designated marker (short nucleotide sequence) that lies before gene in question known as a *promoter*. This sequence works as promoter, but reverse complement won't, so only one strand is transcribed.

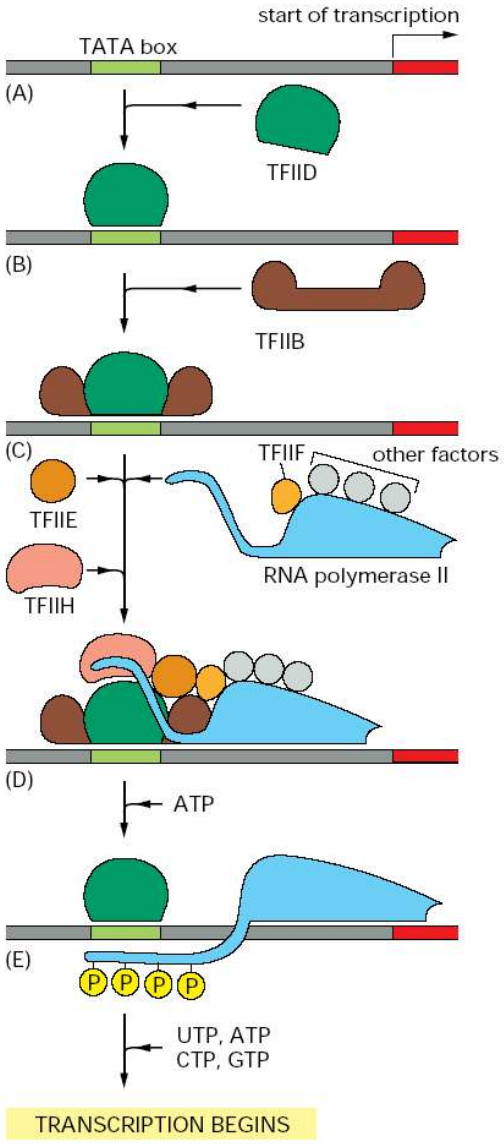


Transcribed strand = *sense, anticoding, template*

Other strand (looks like RNA) = *antisense, coding*

http://www.accessexcellence.org/AB/GG/rna_synth.html

Transcribing DNA to RNA



(A) Promoter contains sequence called *TATA box*.

(B) TATA box bound by transcription factor, TFIID.

(C) This enables binding by another transcription factor, TFIIB.

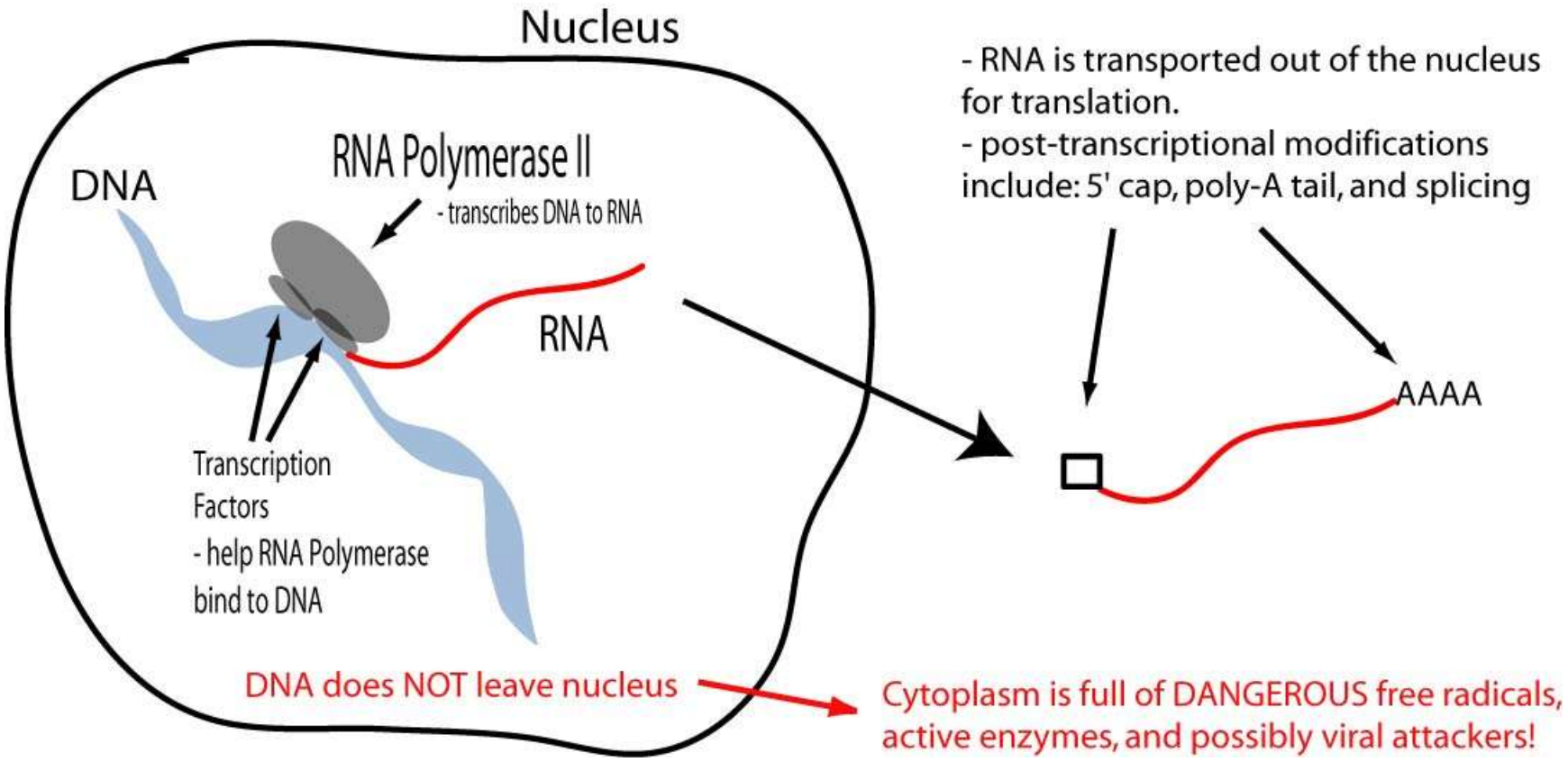
(D) Other transcription factors and RNA polymerase attach (*polymerase* is an enzyme used to assemble DNA and RNA sequences).

(E) RNA polymerase is modified by TFIIF to starting transcribing DNA sequence.

http://www.accessexcellence.org/AB/GG/garland_PDFs/Fig_8.23.pdf

Transcribing DNA to RNA

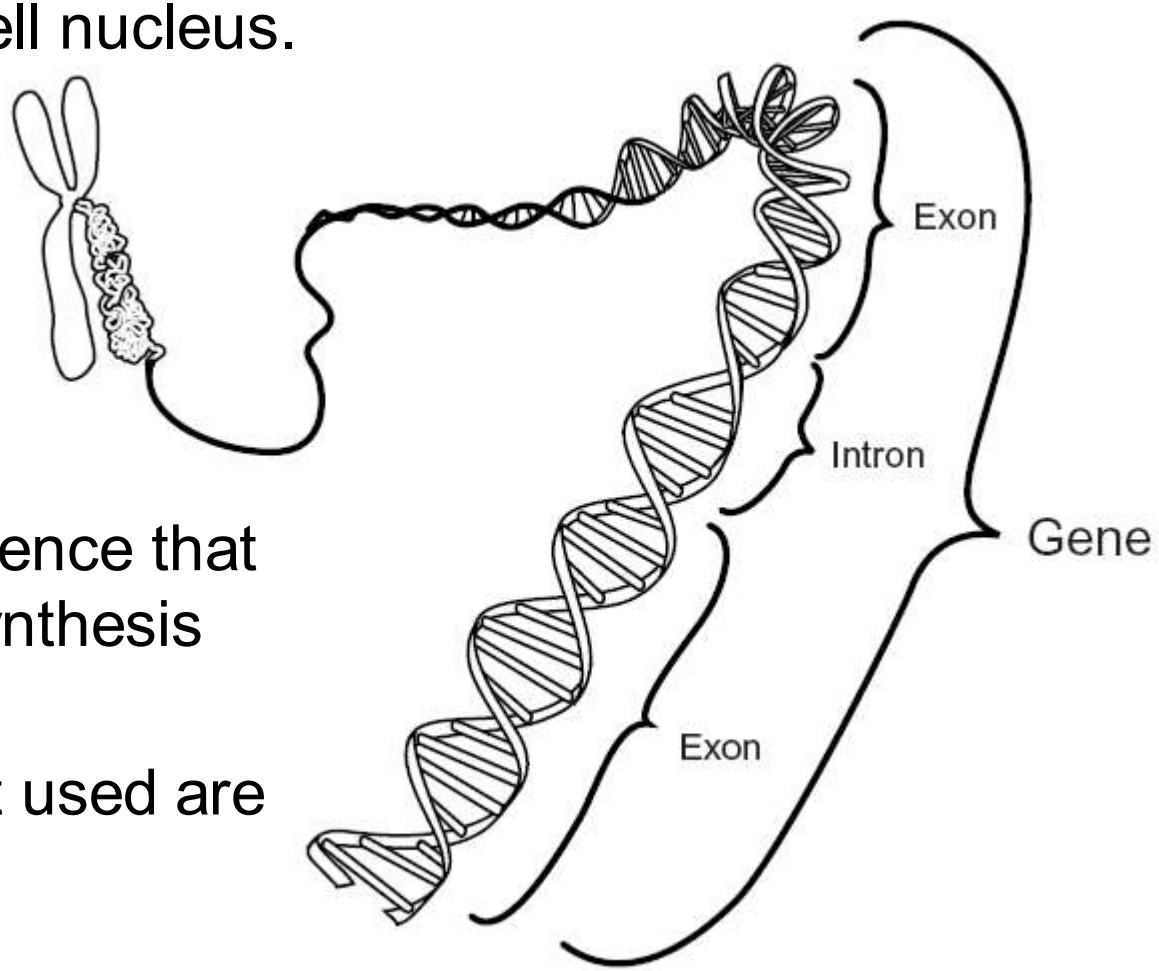
Why go through all of this? To protect the information encoded in DNA.



http://cnx.rice.edu/content/m11415/latest/

Introns and exons

Not all of a gene sequence ends up playing a role in making a protein. Some regions may be spliced out of the mRNA before it leaves the cell nucleus.



The parts of the sequence that are used in protein synthesis are known as *exons*.

The parts that are not used are known as *introns*.

http://www.accessexcellence.org/AB/GG/nhgri_PDFs/exon.pdf

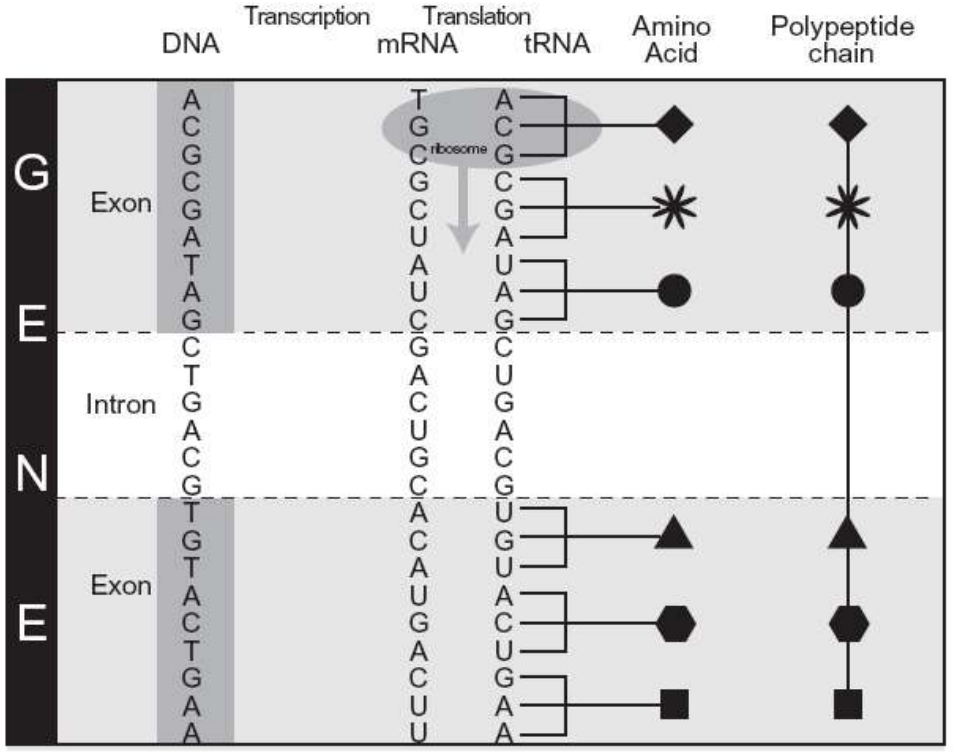
Protein synthesis (translation)

Protein synthesis takes place in a cellular structure known as a *ribosome*. Here the mRNA is acted upon by another form of RNA called *transfer RNA* (*tRNA*).

Each tRNA molecule makes the connection between a specific codon and the corresponding amino acid.

This process is known as *translation*.

Translation ends when a stop codon is encountered.



http://www.accessexcellence.org/AB/GG/nhgri_PDFs/gene2.pdf

Reading frames

Once a gene has been sequenced it is important to determine the *correct open reading frame* (*ORF*). Every region of DNA has six possible reading frames, three in each direction. Typically only one reading frame is used in translating a gene, and this is often the longest open reading frame.

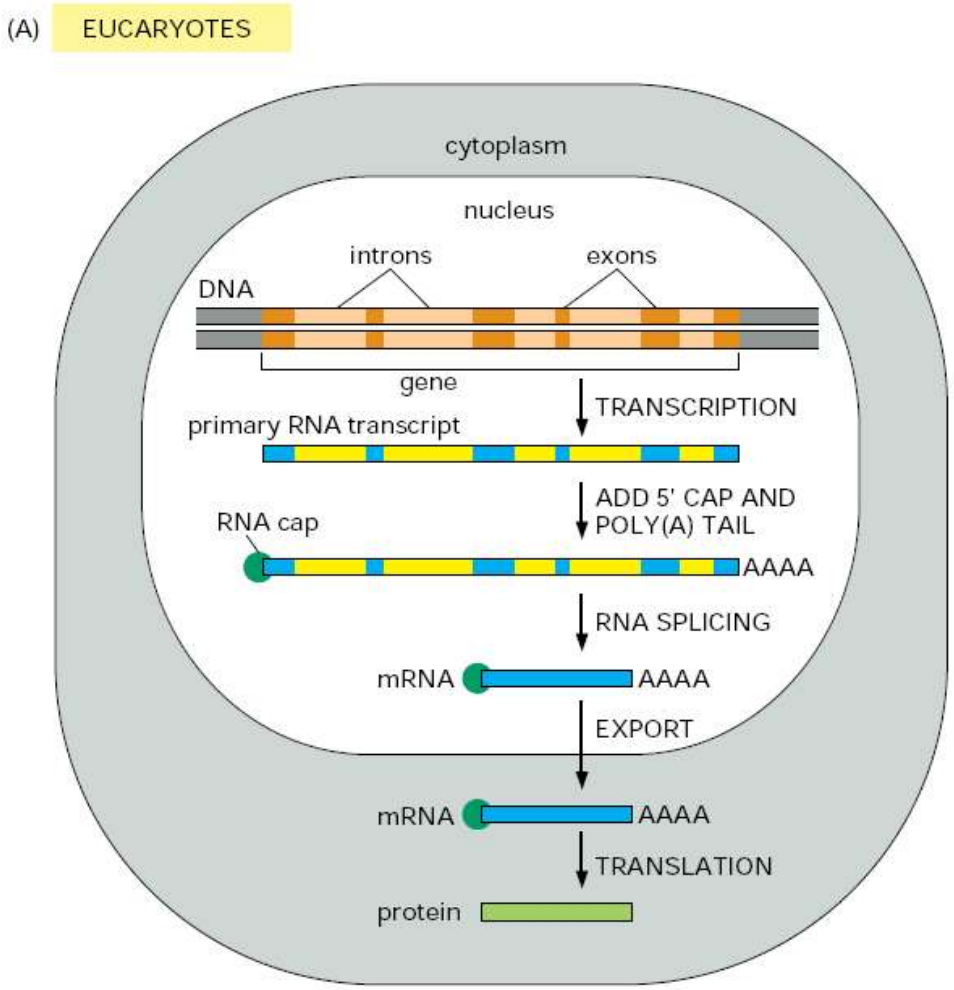
```
5'                                                    3'
atgcccaagctgaatagcgtagaggggttttcatcatttgaggacgatgtataa

1 atg ccc aag ctg aat agc gta gag ggg ttt tca tca ttt gag gac gat gta taa
   M   P   K   L   N   S   V   E   G   F   S   S   F   E   D   D   V   *
2  tgc cca agc tga ata gcg tag agg ggt ttt cat cat ttg agg acg atg tat
    C   P   S   *   I   A   *   R   G   F   H   H   L   R   T   M   Y
3   gcc caa gct gaa tag cgt aga ggg gtt ttc atc att tga gga cga tgt ata
     A   Q   A   E   *   R   R   G   V   F   I   I   *   G   R   C   I
```

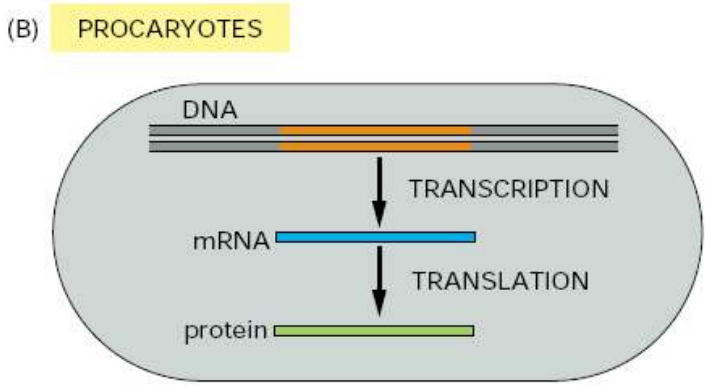
In this case, the first open reading frame is the longest.

http://bioweb.uwlax.edu/GenWeb/Molecular/Seq_Anal/Translation/translation.html

To recap

Steps leading from gene to protein:



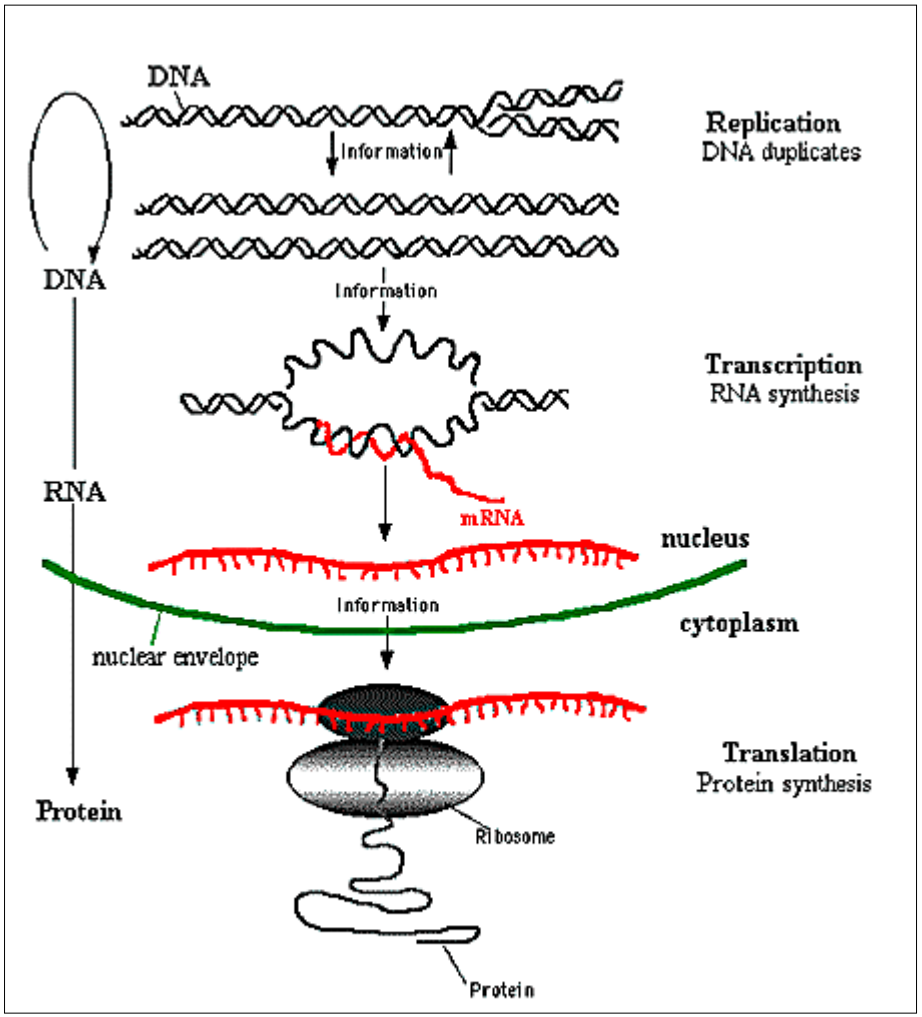


eucaryotes = organisms with cells that contain nuclei which hold DNA.

procaryotes = organisms with cells that lack nuclei, so DNA floats freely in cell.

http://www.accessexcellence.org/AB/GG/nhgri_PDFs/gene2.pdf

The Central Dogma of Molecular Biology



1. DNA copies its information in process involving many enzymes (replication).
2. DNA codes for production of mRNA during transcription.
3. mRNA is processed (by splicing) and migrates from nucleus to cytoplasm.
4. mRNA carries coded information to ribosomes which "read" it and use it for protein synthesis (translation).

http://www.accessexcellence.org/AB/GG/central.html

Wrap-up

Readings for next time (Tuesday, January 27):

- Chapter 1 in your textbook.
- "The Blueprint for Life?" by D.G. Feitelson and M. Treinin, *IEEE Computer*, July 2002, pp. 34-40. (Available via our library's e-journal subscription or online in Blackboard as BlueprintforLife.pdf in the "General Background" folder).
- You should already know the material in Chapter 2 of your textbook. If not, read that as well.

Remember:

- Come to class prepared to discuss what you have read.
- Check Blackboard regularly for updates.