Topic 2 – Gene Sequences

DNA

Biology has been an observed or empirical science for a long time

In the last 75 years, it has become more precise and quantitative: mathematical models used in physics and chemistry have become more common in biology.

1920s: 4 nucleotides and polymer nature of DNA accepted but it was not considered the basis for inheritance

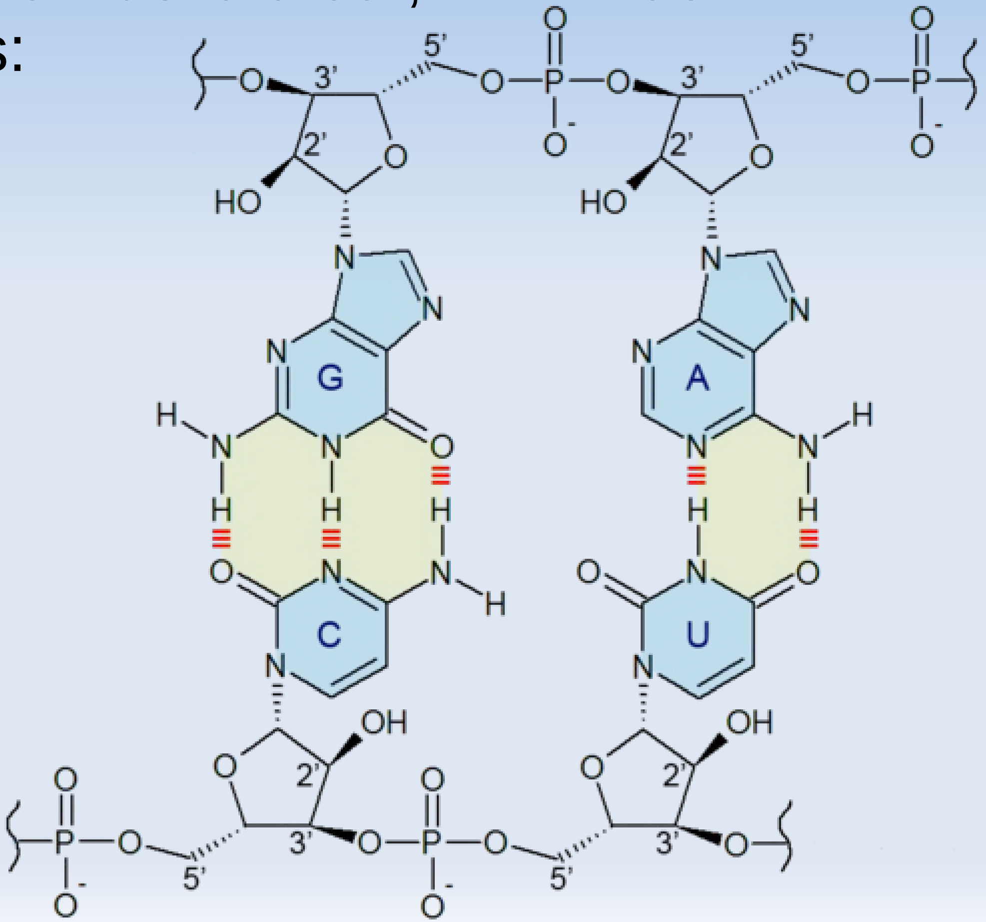
1952: Hershey-Chase experiment showed that DNA is the genetic material in T2 phage

1953: Watson-Crick's double-helix model of DNA structure

DNA bases

Even before the Watson-Crick structure was laid out, DNA was known to be made of four molecules:

- Adenine (A)
- Cytosine (C)
- Guanine (G)
- Thymine (T)

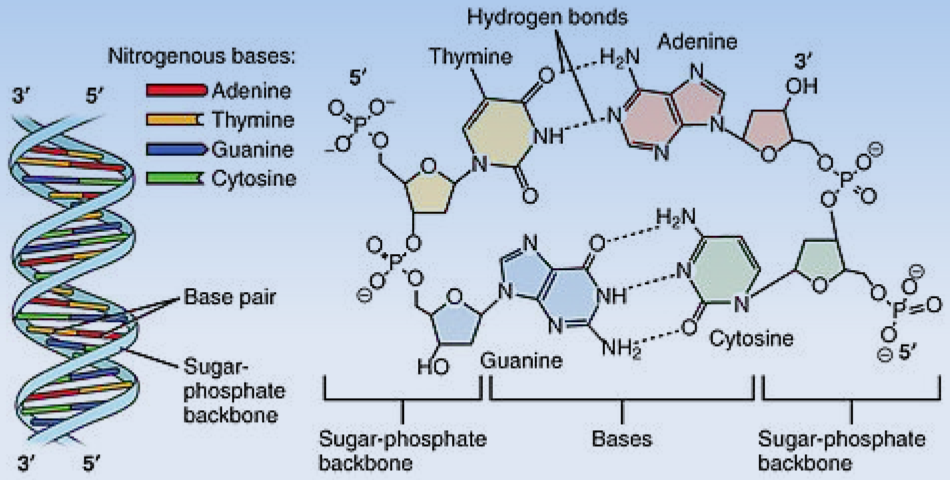


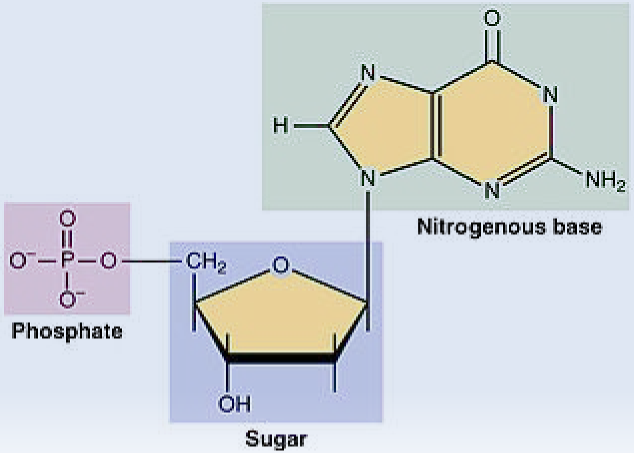
Nucleotides

These four nitrogenous "bases" (adenine, cytosine, guanine, and thymine) are chemical compounds each consisting of about 10 atoms – carbon, hydrogen, nitrogen, and oxygen.

A base bound to a 2-deoxyribose sugar is a nucleoside and a nucleoside bound to a phosphate is a **nucleotide**.

The acidity of DNA is due to the acidic phosphate group

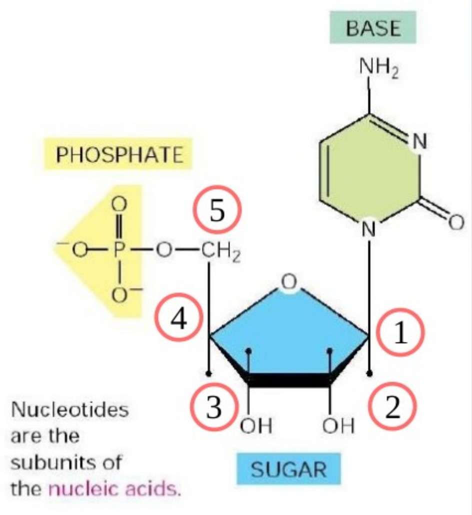




Nucleotides and sugar-phosphate backbone

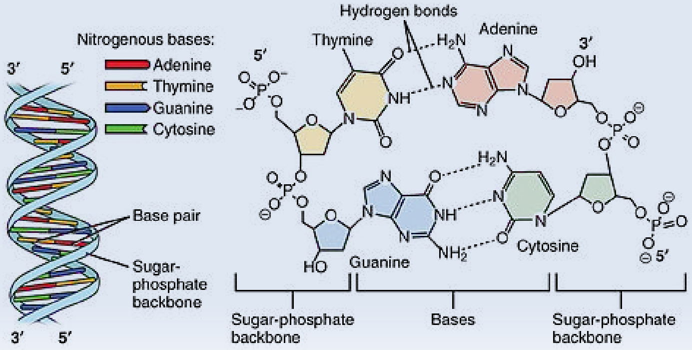
A cytosine (C) combines with a phosphate (p) and a 5-carbon sugar (s) to form the nucleotide:

```
    C
    |
p - s
```



Nucleotides are connected in a chain of alternating “sugar-phosphate” molecules:

```
    C   T   T   G   A   G   T   C   C
    |   |   |   |   |   |   |   |   |
p - s - p - s - p - s - p - s - p - s - p - s - p - s - p - s - p - s
```



Nucleotide chain

Nucleotides are connected via alternating “sugar-phosphate” molecules:

```
   C     T     T     G     A     G     T     C     C
   |     |     |     |     |     |     |     |     |
p-s-p-s-p-s-p-s-p-s-p-s-p-s-p-s-p-s
```

There is a direction to this strand: the first phosphate tells us where the DNA strand starts and is called the **5'** end. The last sugar indicates the opposite side and is called the **3'** end.

Apart from the direction, the sugar-phosphate backbone does not contain any genetic information.

Thus, the above sequence of nucleotides can be represented as:

5'-C T T G A G T C C-3'

DNA Structure

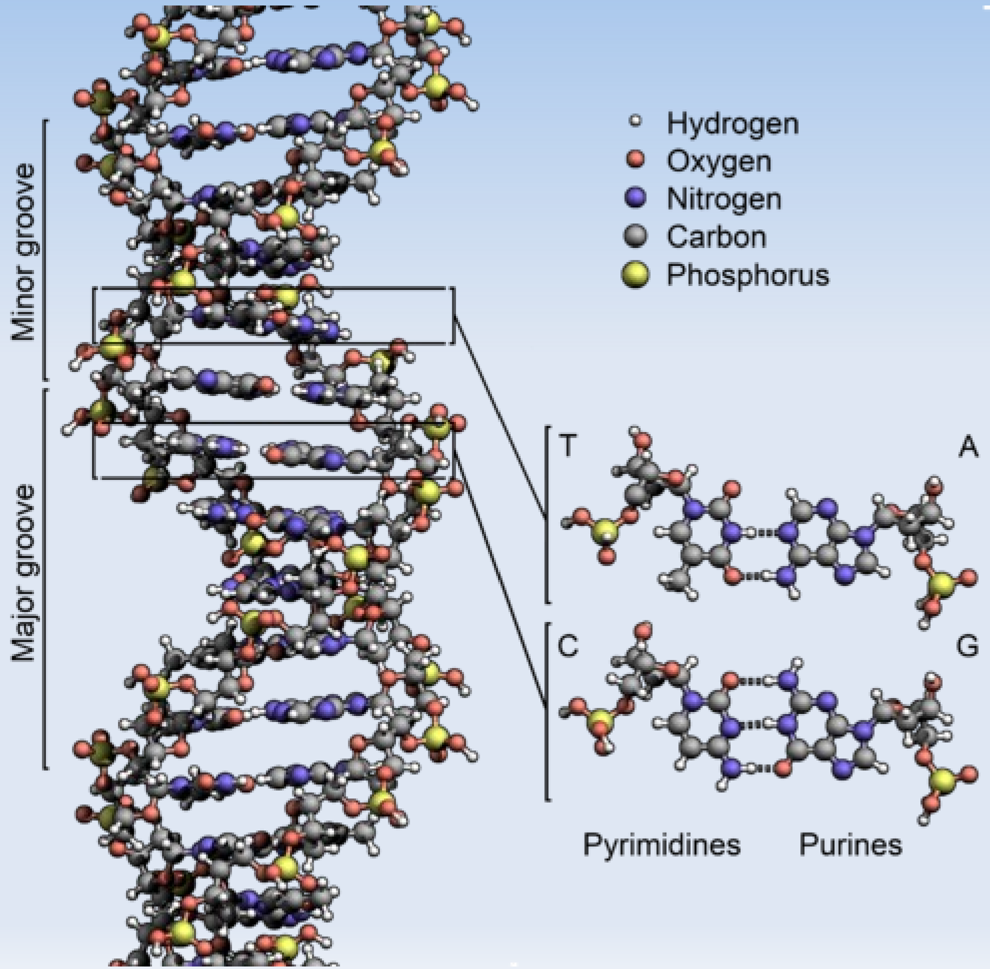
Watson-Crick double-helix model of DNA structure – showed that there is a strict base pairing rule:

Adenines bind with Thymines

Cytosine bind with Guanine

Original paper ends with this amazing understatement:

"*It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material.*"



DNA Base Pairs

The sequence of nucleotides:

```
5'-C T T G A G T C C-3'
```

is likely to exist as “double-stranded” DNA:

```
5'-C T T G A G T C C-3'
   | | | | | | | | |            9 ”base-pair” long sequence
3'-G A A C T C A G G-5'  ←  Complementary strand
```

Note that the complementary strand goes in the opposite direction. DNA is more stable in the double-helix form, i.e. with pairs of bases. Such a “double-stranded” DNA is usually specified in units of base-pairs.

DNA Base Pairs

Under ”normal” conditions, the base-pairing rules are followed and if one strand is specified:

```
5'-C T T G A G T C C-3'
```

the other (complementary) strand can be deduced using the base-pairing rules:

```
5'-C T T G A G T C C-3'
   | | | | | | | | |
3'-G A A C T C A G G-5'  ←  Complementary strand
```

Since the information in the two strands is redundant, usually only one is shown. The other strand is assumed to be present.

RNA

We saw earlier that a gene is used to make a protein or RNA product. RNA is similar to DNA but has a different sugar: ribose.

The ribose-phosphate backbone makes RNA relatively stable as a **single strand**. Furthermore, RNA has Uracils instead of the Thymines seen in DNA.

A Uracil can bind to an Adenine in DNA or RNA with less energy than Thymine – good for efficiency in making copies but results in RNA being less stable than DNA.

A process called “transcription” generates RNA from a DNA strand.

DNA Transcription

A process called “transcription” generates RNA from a DNA strand.

- The DNA molecule unwinds enough to let certain proteins come into contact with the 5' end of a gene - the ”start” location and create an RNA copy of the ”coding” strand.
- These proteins ”read” the bases from the 5' end to the 3' end of the ”coding” strand by binding to the non-coding strand and following the base-pairing rules: an Adenine in the DNA generates a Uracil in the RNA strand and a Guanine in the DNA generates a Cytosine in the RNA, and so on.

DNA Transcription

Suppose we have the following dsDNA:

```
5'-C T T G A G T C C-3'  ← Coding or "sense" strand
   | | | | | | | | |
3'-G A A C T C A G G-5'  ← Reverse complement
```

This will get transcribed to the following RNA single strand

```
5'-C U U G A G U C C-3'
```

Protein

If the product of a gene is a protein molecule, the transcribed RNA is used to make the protein.

Information in a gene goes in one direction:

DNA → RNA → Protein

This was thought to be a universal law for all living things until the discovery of RNA viruses which have RNA as their genetic material. For example: retroviruses like the HIV virus.

In eukaryotic cells, messenger RNA goes through a number of sequence ”editing” steps before being translated to protein sequences

Protein

DNA → RNA → Protein

Proteins, like DNA and RNA, are linear sequences but instead of nucleotides, proteins are made of amino acids.

The carboxyl group of one amino acid can bind to the amine group of the next, forming a stable chain of amino acids – a protein.

There are 22 amino acids that occur naturally – proteins have a much larger alphabet than RNA or DNA. Two of these 22 are considered ”non-standard”

Three consecutive RNA bases (a **codon**) are used to make a single amino acid.

Counting Possibilities

DNA → RNA → Protein

Let's count the number of possible codons. Remember that a codon is any set of 3 (RNA) nucleotides.

This means there are 4 possibilities for the first of the 3, another 4 possibilities for the second and 4 for the third nucleotide in the codon

→ 4 * 4 * 4 different codons

→ 64 codons

These 64 codons are converted to just 20-22 amino acids – there is redundancy in coding: there may be many codons that are translatedd to the same amino acid.

The Genetic Code

The rules used to translate an RNA codon to an amino acid in a protein are called the Genetic code:

Our RNA single strand

5'-CUU GAG UCC-3'

will be translated to the amino acid sequence:

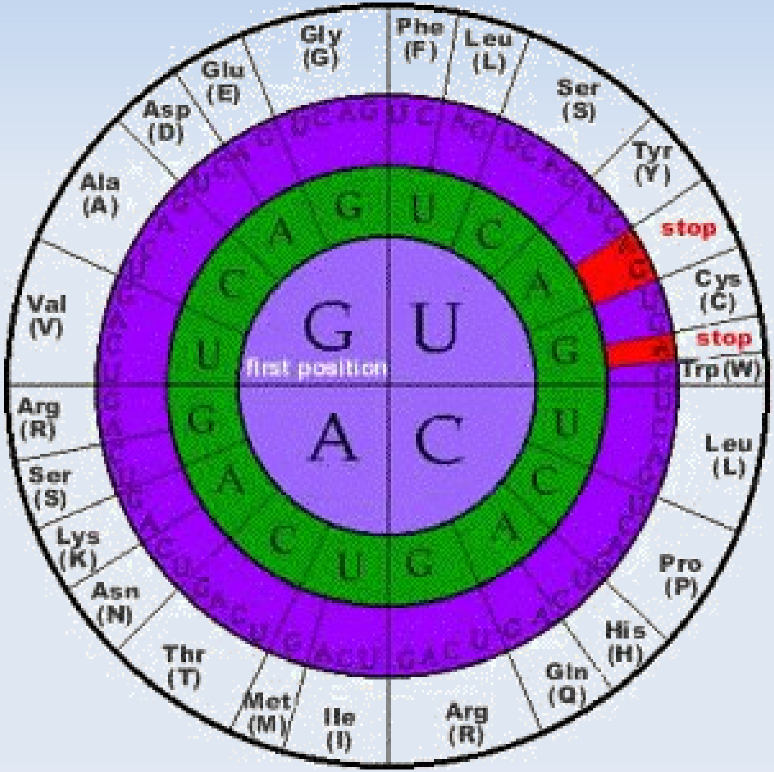
Leu-Glu-Ser

Second Letter

1st 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

The Genetic Code

The standard Genetic code is used in most organisms but there are alternative genetic codes.



UUU Phe	UCU Ser	UAU Tyr	UGU Cys
UUC Phe	UCC Ser	UAC Tyr	UGC Cys
UUA Leu	UCA Ser	UAA Stp, Gln[3]	UGA Stp,Trp[4,5],Cys[6],SeC[7]
UUG Leu	UCG Ser	UAG Stp, Gln[3]	UGG Trp
CUU Leu	CCU Pro	CAU His	CGU Arg
CUC Leu	CCC Pro	CAC His	CGC Arg
CUA Leu	CCA Pro	CAA Gln	CGA Arg
CUG Leu, Ser[1]	CCG Pro	CAG Gln	CGG Arg, Usp[8]
AUU Ile	ACU Thr	AAU Asn	AGU Ser
AUC Ile	ACC Thr	AAC Asn	AGC Ser
AUA Ile, Usp[2]	ACA Thr	AAA Lys	AGA Arg, Usp[9]
AUG Met	ACG Thr	AAG Lys	AGG Arg
GUU Val	GCU Ala	GAU Asp	GGU Gly
GUC Val	GCC Ala	GAC Asp	GGC Gly
GUA Val	GCA Ala Res[10]	GAA Glu	GGA Gly
GUG Val	GCG Ala	GAG Glu	GGG Gly

1–Candida, 2–Micrococcus, 3–ciliated protozoans and green algae, 4–Mycoplasma, 5–Bacteria, 6–Euplotes, 7–Selenocysteine, 8–spiroplasma, 9–Micrococcus, 10 – resume ssrA RNA codon

Reading Frames

A DNA strand has a direction but where does translation start?

Remember, codons occur as consecutive non-overlapping nucleotide triplets. A reading frame is how nucleotide sequences are divided up into codon triplets:

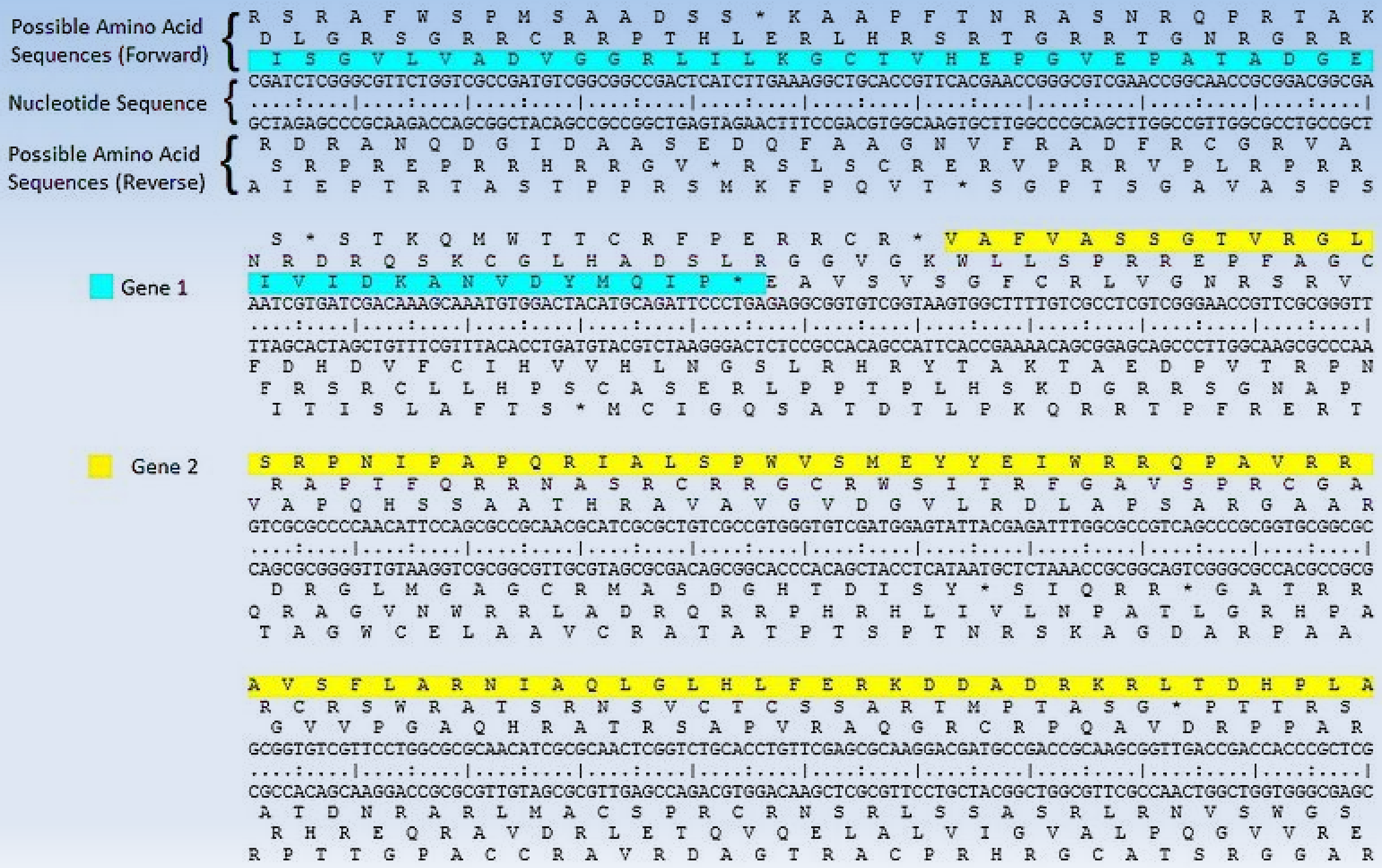
AGG·TGA·CAC·CGC·AAG·CCT·TAT·ATT·AGC

A·**GGT·GAC·ACC·GCA·AGC·CTT·ATA·TTA**·GC

AG·**GTG·ACA·CCG·CAA·GCC·TTA·TAT·TAG**·C

Reading Frames

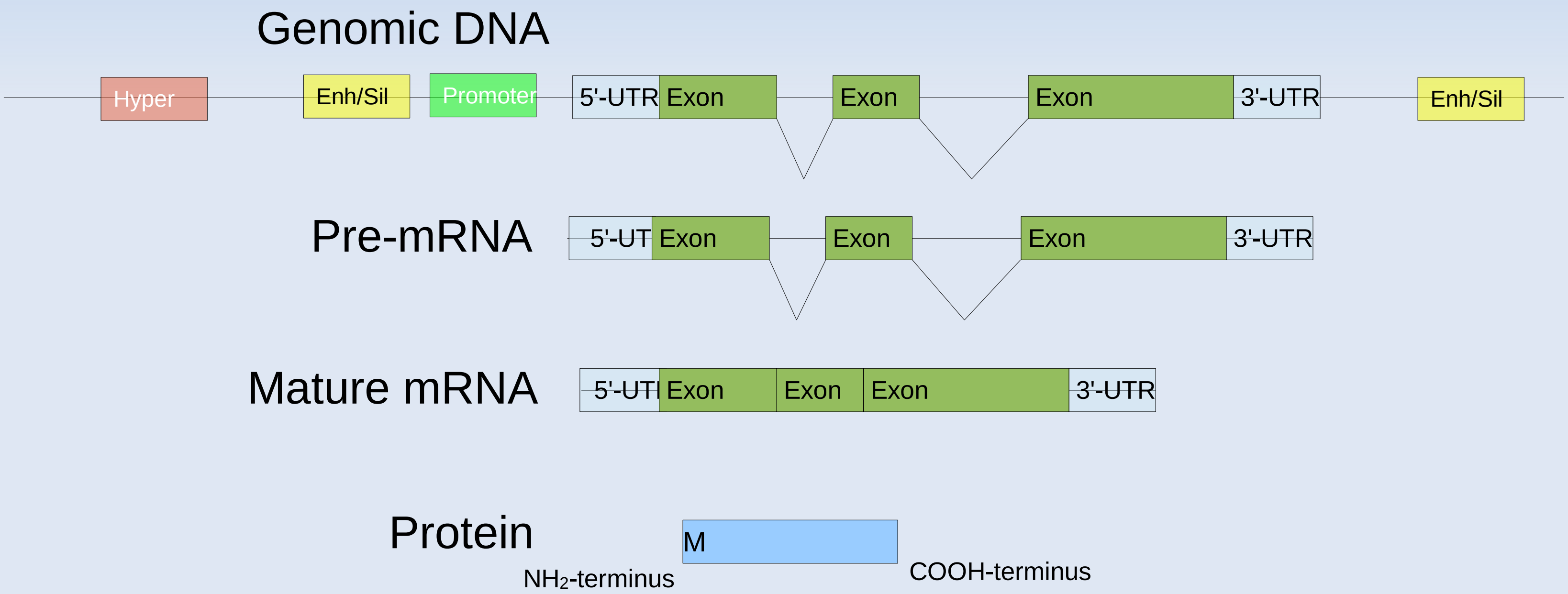
Remember, that a DNA strand is usually assumed to have a reverse complement strand as well. The complementary strand has 3 reading frames as well:



Gene Expression and Regulation

The expression of genes – whether they are

Regulation: promoters, upstream and downstream enhancers/silencers or DNase hypersensitive sites.



Extras

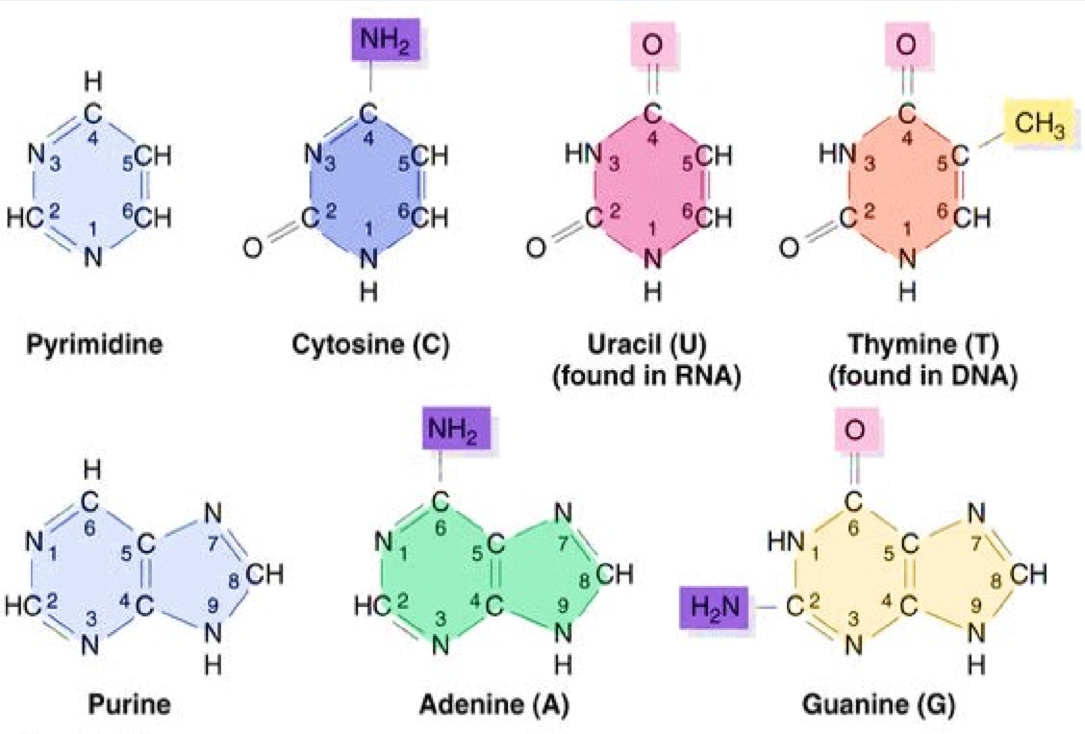
Chemical properties of DNA

- Deoxyribose or Ribose == 5-carbon sugar
- Nucleotides == Nityrogenous bases
- Acid == phosphate group

a) Pyrimidines (Y): C, U, T

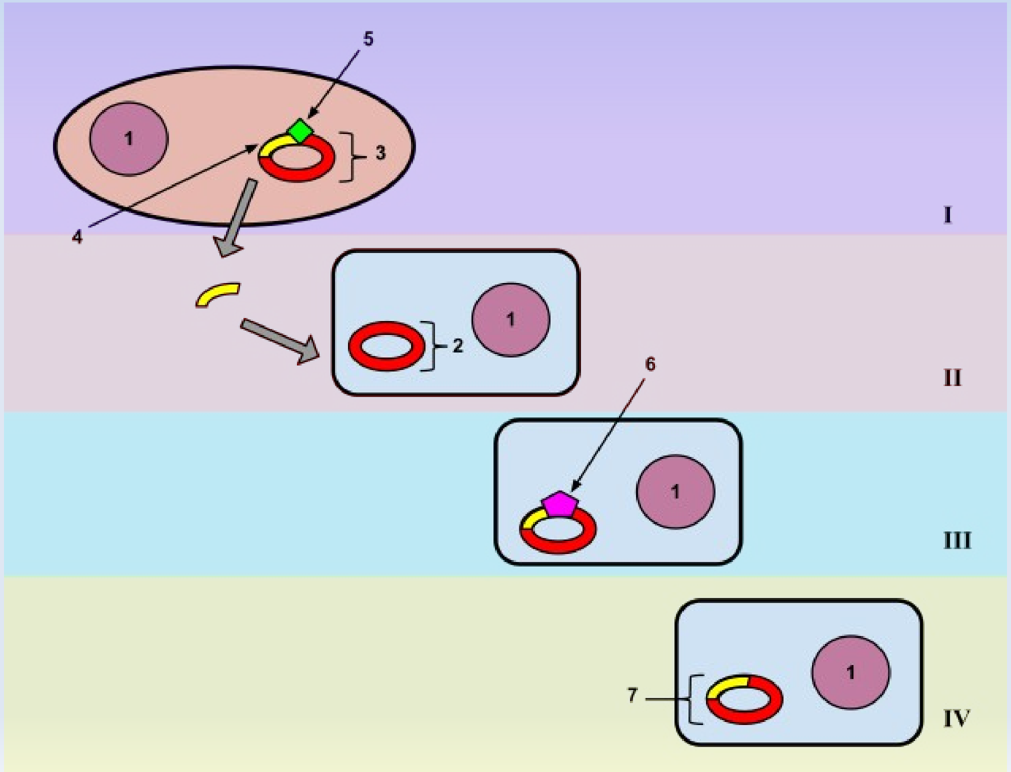
b) Purines (R): A, G

Ambiguity code



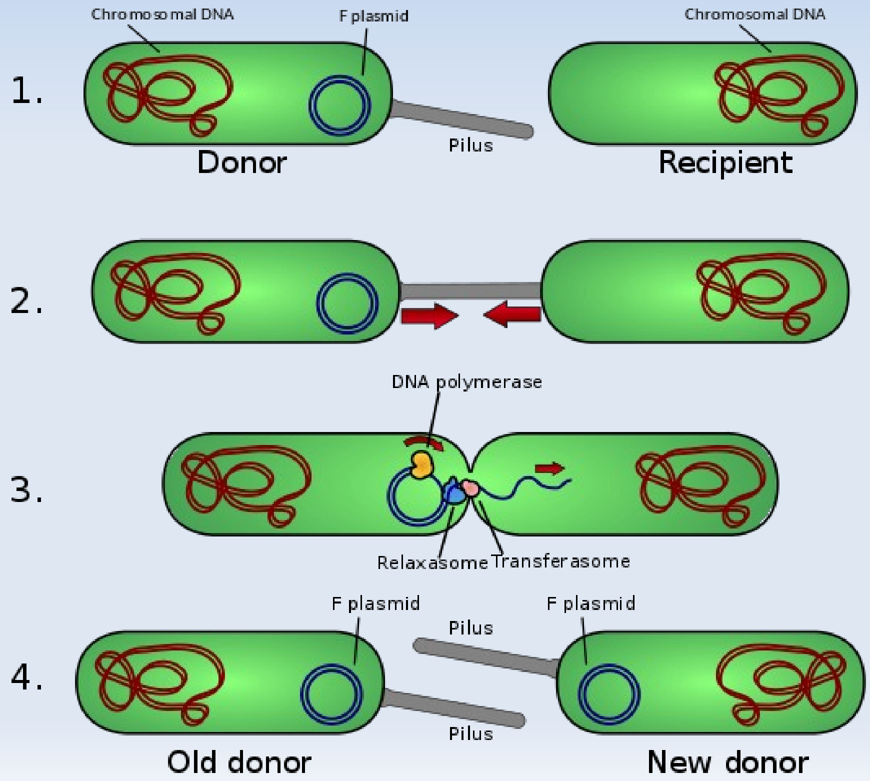
Evidence: DNA holds genetic information

- 1944 experiments with bacteria – Transformation
- gene from one bacterium (I) can go to another (II-IV)
- 1953: Double-helix



Evidence: DNA holds genetic information

- 1946 experiments also in bacteria – Conjugation
- gene from one bacterium (1)
  can go to another (2-4)



Evidence: DNA holds genetic information

- 1952 experiments in bacteria/virus – Transduction
- gene from one bacterium (1) can go to another (5-7) via a phage

Expts in Madison
in 1952