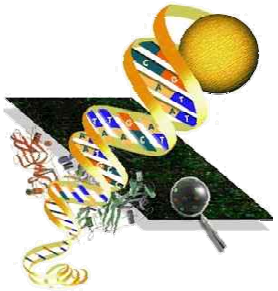
DNA/RNA overview



DNA

deoxyribo**n**ucleic **a**cid

4 bases

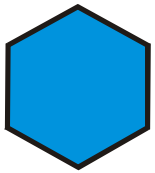
A = Adenine

T = Thymine

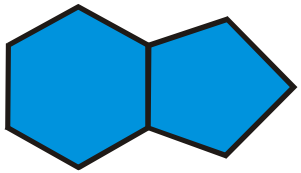
C = Cytosine

G = Guanine

Pyrimidine ($C_4N_2H_4$)

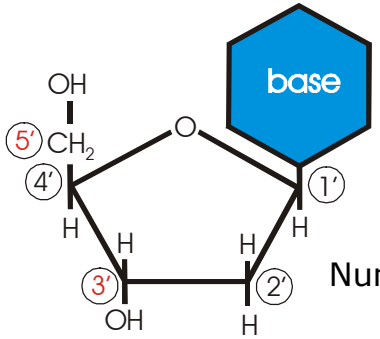


Purine ($C_5N_4H_4$)



Nucleoside

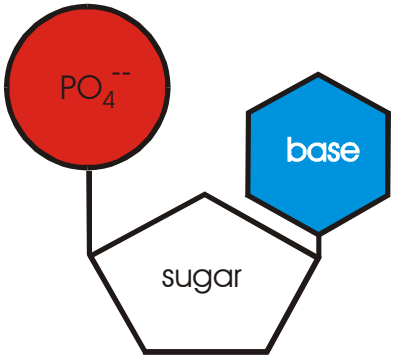
base + sugar (deoxyribose)



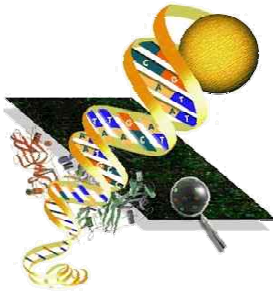
Numbering of carbons?

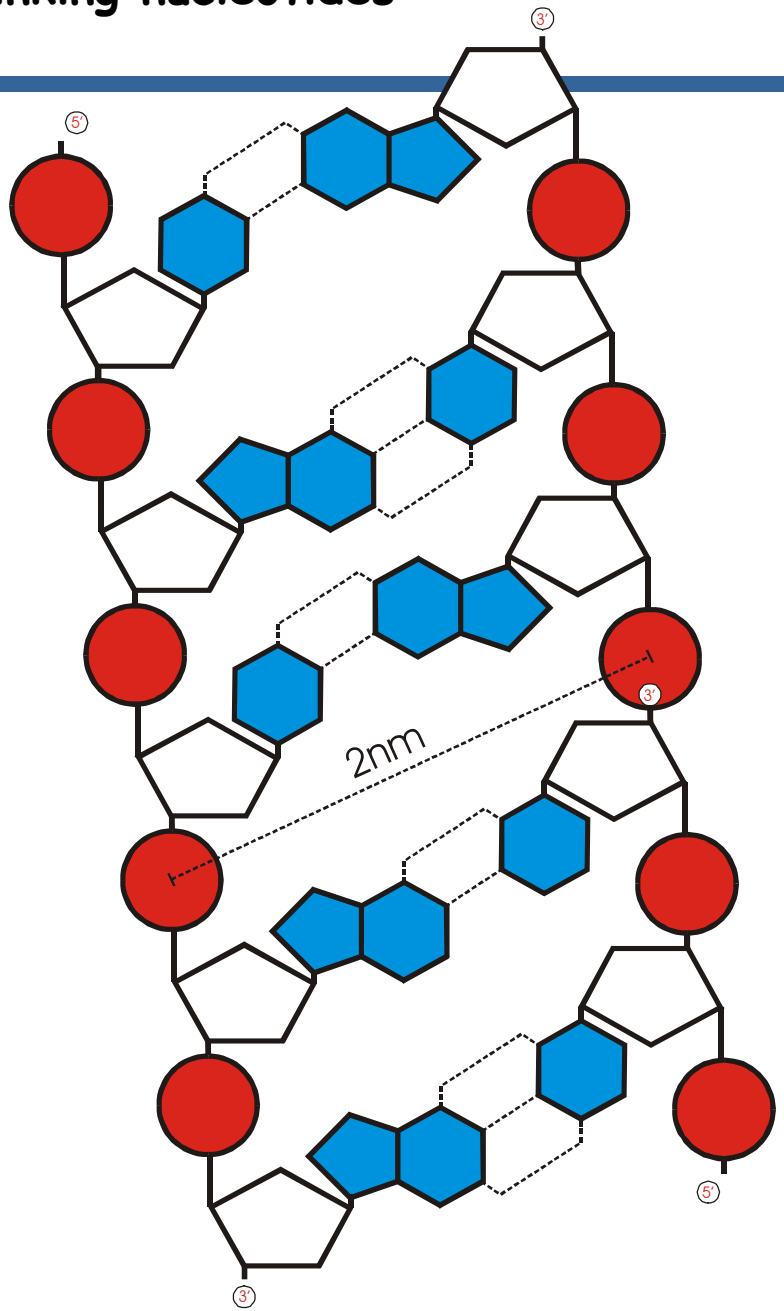
Nucleotide

base + sugar + phosphate



Linking nucleotides

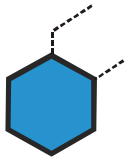




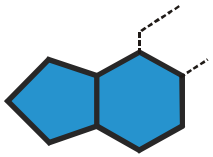
Hydrogen bonds

N-H------N

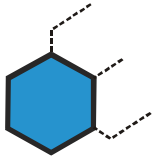
N-H------O



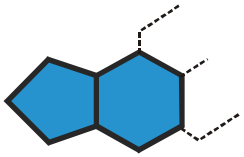
Thymine



Adenine

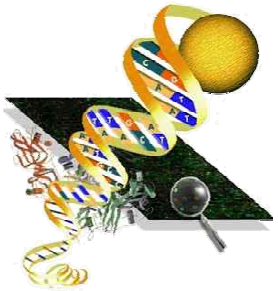


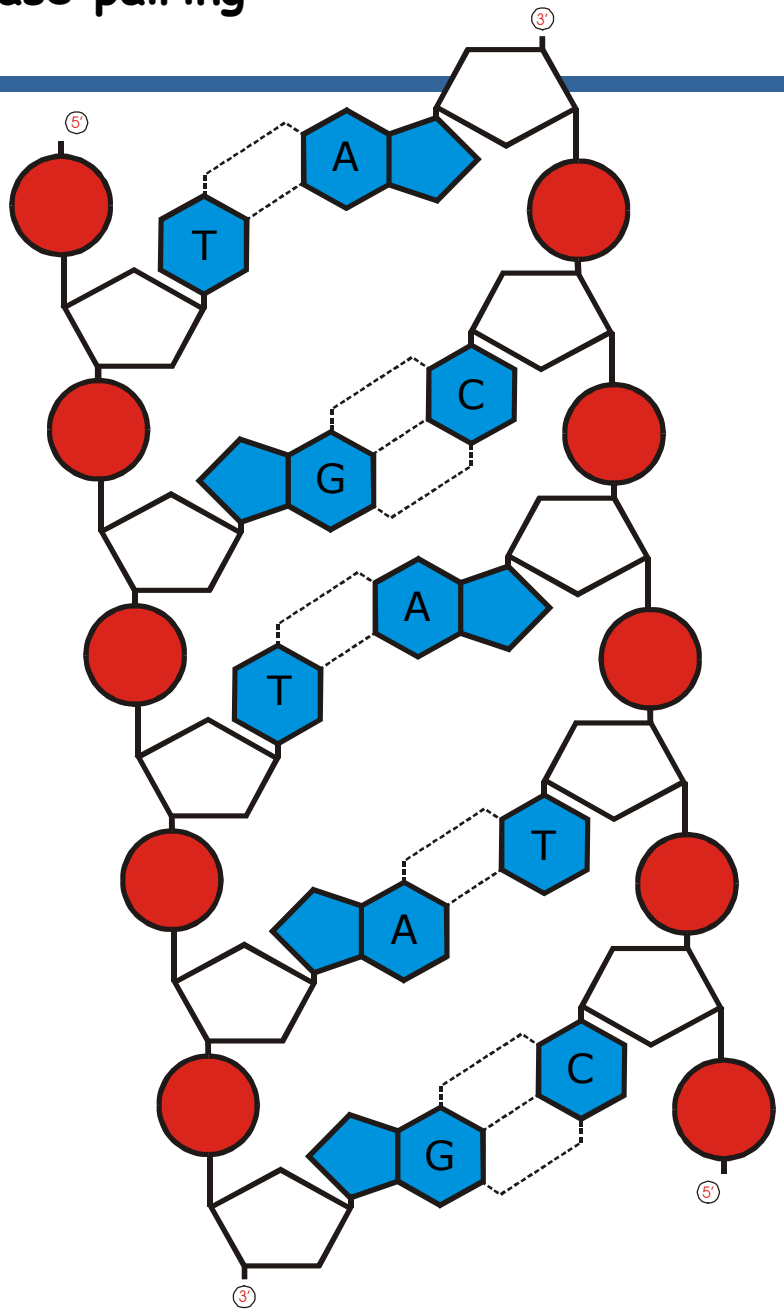
Cytosine



Guanine

Base pairing





Base pairing (Watson-Crick):

A/T (2 hydrogen bonds)

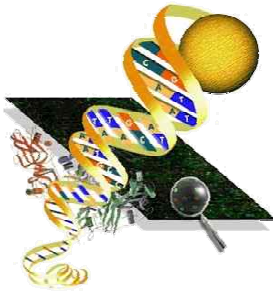
G/C (3 hydrogen bonds)

Always pairing a purine and a pyrimidine yields a constant width

DNA base composition:

A + G = T + C (Chargaff's rule)

DNA conventions



1. DNA is a right-handed helix
2. The 5′ end is to the left by convention

5′-ATCGCAATCAGCTAGGTT-3′	sense (forward)
3′-TAGCGTTAGTCGATCCAA-5′	antisense (reverse)

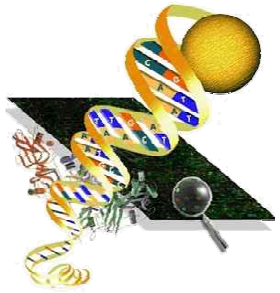
,5 -AACCTAGCGTTAGTCGATCCAA- 5′

,3 -TTGGTAGCTAGTCAATCGCAT- 3′

,3 - TTGGTAGCTAGTCAATCGCAT - 5′

,5 - AACCTAGCGTTAGTCGATCCAA - 3′

DNA structure



Some more facts:

1. Forces stabilizing DNA structure: **Watson-Crick-H-bonding** and **base stacking** (planar aromatic bases overlap geometrically and electronically → energy gain)

2. Genomic DNAs are large molecules:
   Eschericia coli: 4.7 x 10^6 bp; ~ 1 mm contour length
   Human: 3.2 x 10^9 bp; ~ 1 m contour length

3. Some DNA molecules (plasmids) are circular and have no free ends:

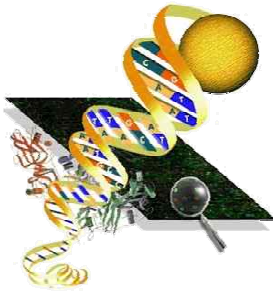
   mtDNA
   bacterial DNA (only one circular chromosome)

4. Average gene of 1000 bp can code for average protein of about 330 amino acids

5. Percentage of non-coding DNA varies greatly among organisms

Organism	# Base pairs	# Genes	Non-coding DNA
small virus	4 x 10^3	3	very little
small virus	3 x 10^5	200	very little
bacterium	5 x 10^6	3000	10 - 20%
yeast	1 x 10^7	6000	> 50%
human	3.2 x 10^9	35,000?	99%
amphibians	< 80 x 10^9	?	?
plants	< 900 x 10^9	23,000 - >50,000	> 99%

RNA structure



RNA

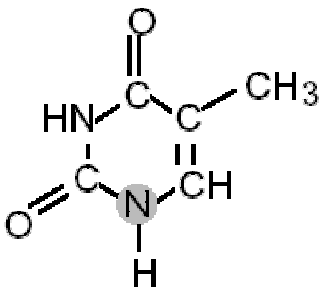
ribo**n**ucleic **a**cid

3 major types of RNA

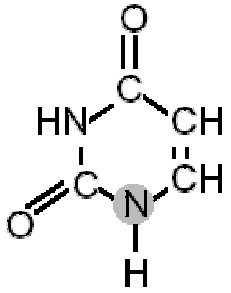
messenger RNA (mRNA); template for protein synthesis
transfer RNA (tRNA); adaptor molecules that decode the genetic code
ribosomal RNA (rRNA); catalyzing the synthesis of proteins

4 bases

A = Adenine
U = Uracil
C = Cytosine
G = Guanine



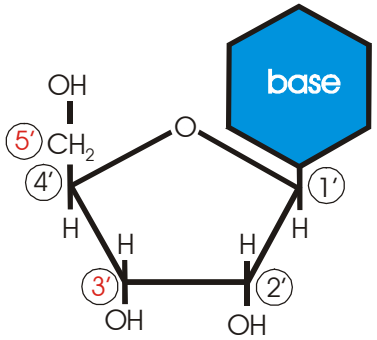
Thymine (DNA)



Uracil (RNA)

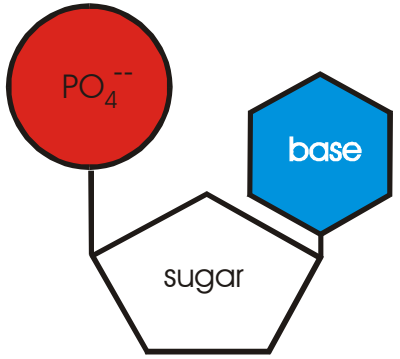
Nucleoside

base + sugar (ribose)



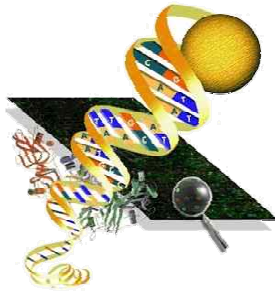
Nucleotide

base + sugar + phosphate



DNA
RNA
Amino acids/ proteins
PCR, sequencing
Mutations

Base interactions in RNA



Base pairing:

U/A/(T) (2 hydrogen bonds)

G/C (3 hydrogen bonds)

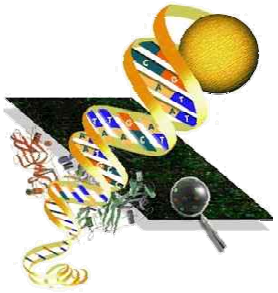
RNA base composition:

A + G ≠ U + C **Chargaff's rule does not apply (RNA usually prevails as single strand)**

RNA structure:

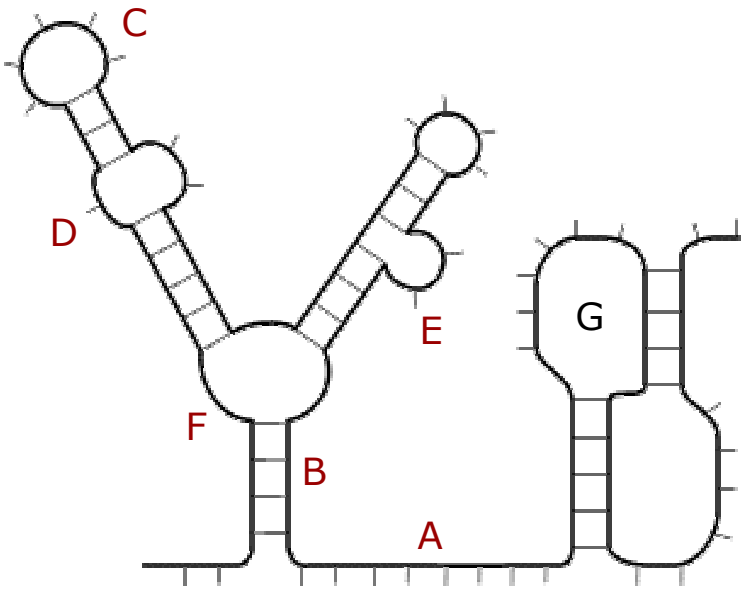
- usually single stranded
- many self-complementary regions → RNA commonly exhibits an intricate secondary structure (relatively short, double helical segments alternated with single stranded regions)
- complex tertiary interactions fold the RNA in its final three dimensional form
- the folded RNA molecule is stabilized by interactions (e.g. hydrogen bonds and base stacking)

RNA structure



Primary structure

Secondary structure



A) single stranded regions

formed by unpaired nucleotides

B) duplex

double helical RNA (A-form with 11 bp per turn)

C) hairpin

duplex bridged by a loop of unpaired nucleotides

D) internal loop

nucleotides not forming Watson-Crick base pairs

E) bulge loop

unpaired nucleotides in one strand,
other strand has contiguous base pairing

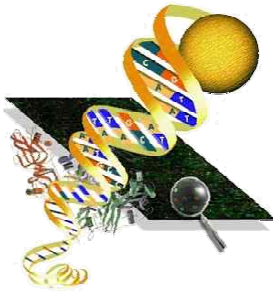
F) junction

three or more duplexes separated by single stranded regions

G) pseudoknot

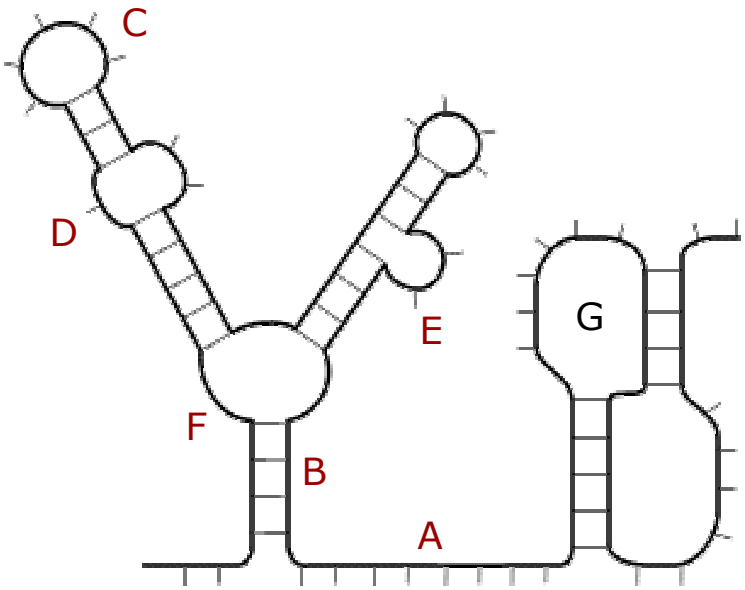
tertiary interaction between bases of hairpin loop and outside bases

RNA structure

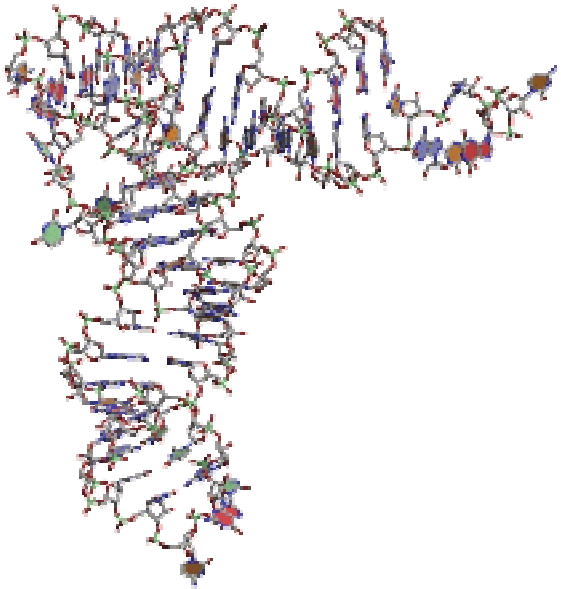


Primary structure

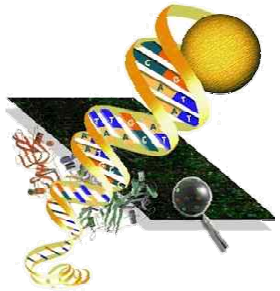
Secondary structure



Tertiary structure



RNA structure



How to predict RNA secondary/tertiary structure?

Probing RNA structure experimentally:

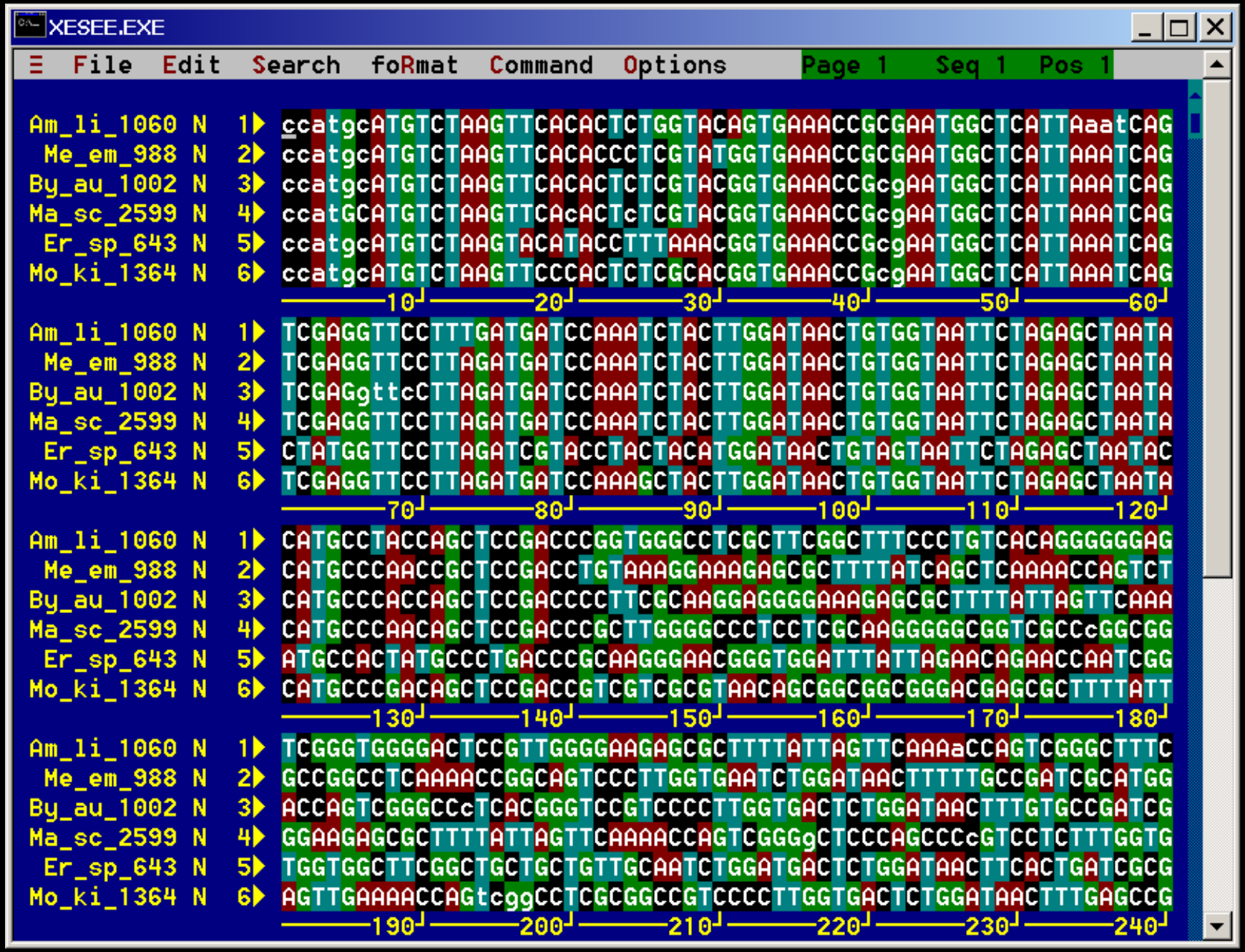
- physical methods (single crystal X-ray diffraction, electron microscopy)
- chemical and enzymatic methods
- mutational analysis (introduction of specific mutations to test change in some function or protein-RNA interaction)

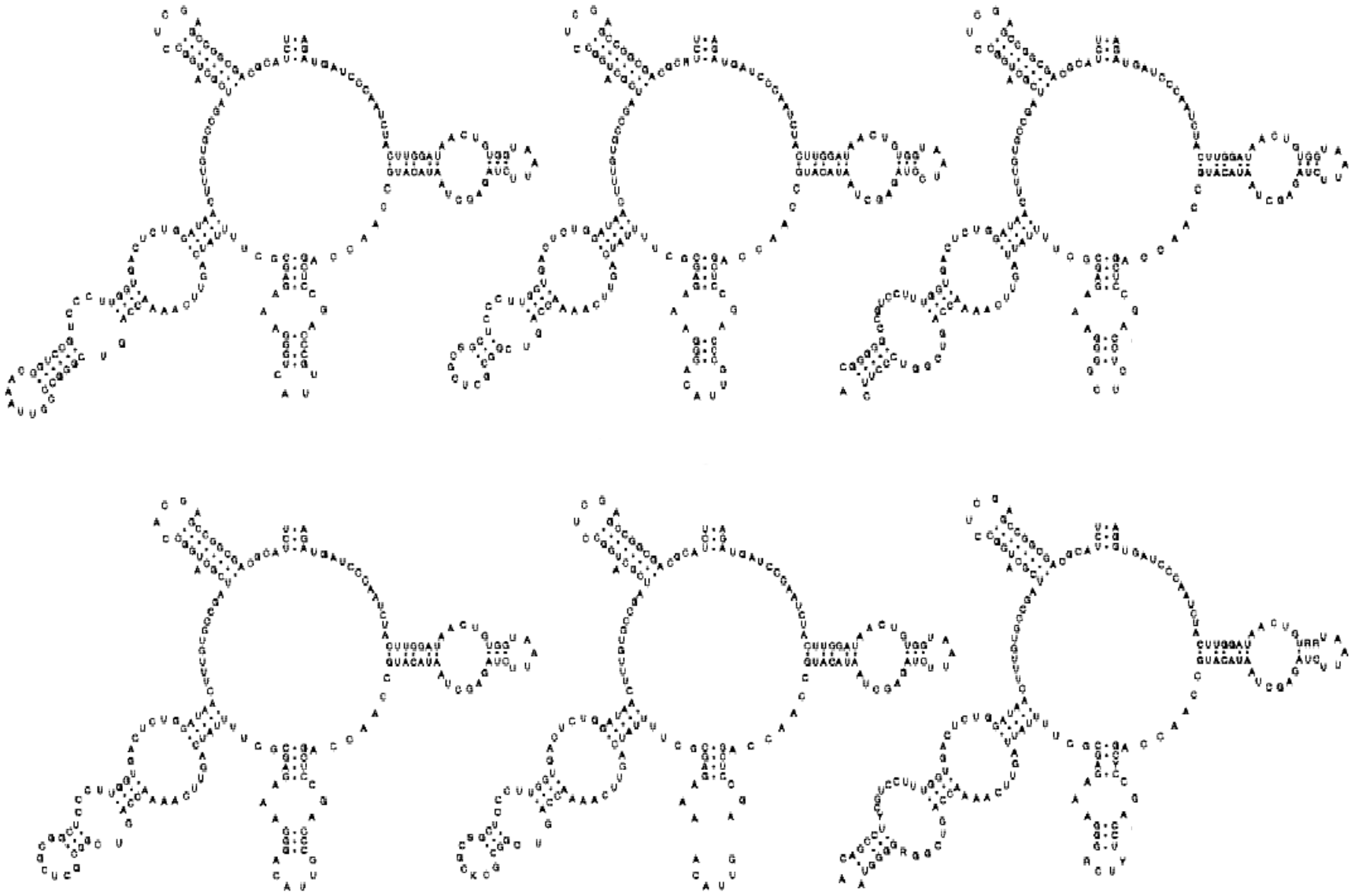
Thermodynamic prediction of RNA structure:

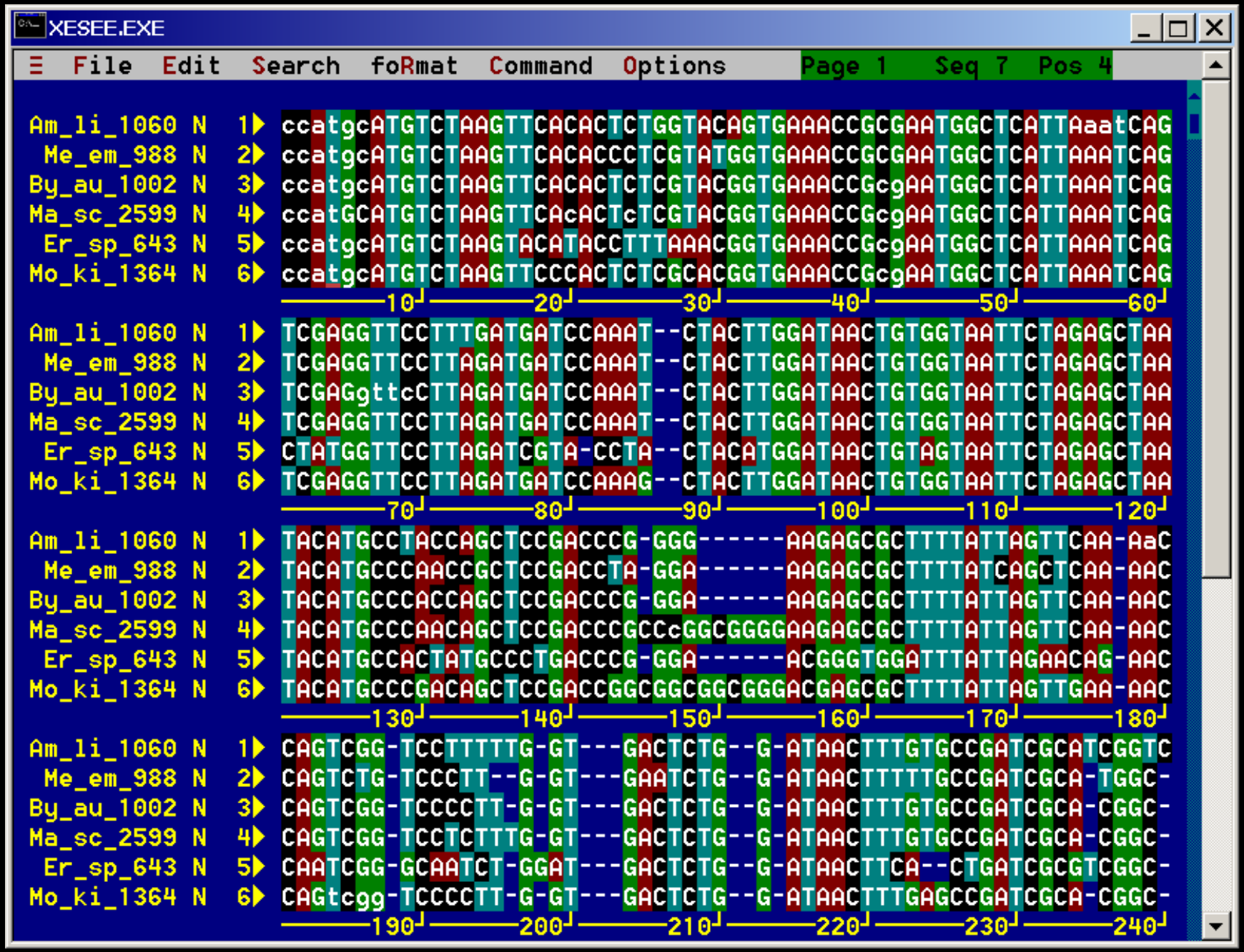
- RNA molecules comply to the laws of thermodynamics, therefore it should be possible to deduce RNA structure from its sequence by finding the conformation with the lowest free energy
- Pros: only one sequence required; no difficult experiments; does not rely on alignments
- Cons: thermodynamic data experimentally determined, but not always accurate; possible interactions of RNA with solvent, ions, and proteins

Comparative determination of RNA structure:

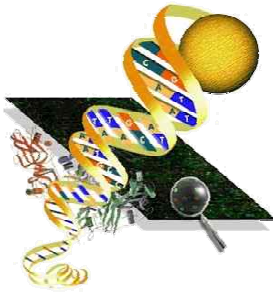
- basic assumption: secondary structure of a functional RNA will be conserved in the evolution of the molecule (at least more conserved than the primary structure); when a set of homologous sequences has a certain structure in common, this structure can be deduced by comparing the structures possible from their sequences
- Pros: very powerful in finding secondary structure, relatively easy to use, only sequences required, not affected by interactions of the RNA and other molecules
- Cons: large number of sequences to study preferred, structure constrains in fully conserved regions cannot be inferred, extremely variable regions cause problems with alignment


XESEE.EXE
File Edit Search foRmat Command Options Page 1 Seq 1 Pos 1
Am_li_1060 N 1 ccatgcATGTCTAAGTTCACACTCTGGTACAGTGAAACCGCGAATGGCTCATTAaatCAG
Me_em_988 N 2 ccatgcATGTCTAAGTTCACACCCTCGTATGGTGAAACCGCGAATGGCTCATTAAATCAG
By_au_1002 N 3 ccatgcATGTCTAAGTTCACACTCTCGTACGGTGAAACCGcgAATGGCTCATTAAATCAG
Ma_sc_2599 N 4 ccatGCATGTCTAAGTTCAcACTcTCGTACGGTGAAACCGcgAATGGCTCATTAAATCAG
Er_sp_643 N 5 ccatgcATGTCTAAGTACATACCTTTAAACGGTGAAACCGcgAATGGCTCATTAAATCAG
Mo_ki_1364 N 6 ccatgcATGTCTAAGTTCCCACTCTCGCACGGTGAAACCGcgAATGGCTCATTAAATCAG
10 20 30 40 50 60
Am_li_1060 N 1 TCGAGGTTCCTTTGATGATCCAAATCTACTTGGATAACTGTGGTAATTCTAGAGCTAATA
Me_em_988 N 2 TCGAGGTTCCTTAGATGATCCAAATCTACTTGGATAACTGTGGTAATTCTAGAGCTAATA
By_au_1002 N 3 TCGAGgttcCTTAGATGATCCAAATCTACTTGGATAACTGTGGTAATTCTAGAGCTAATA
Ma_sc_2599 N 4 TCGAGGTTCCTTAGATGATCCAAATCTACTTGGATAACTGTGGTAATTCTAGAGCTAATA
Er_sp_643 N 5 CTATGGTTCCTTAGATCGTACCTACTACATGGATAACTGTAGTAATTCTAGAGCTAATAC
Mo_ki_1364 N 6 TCGAGGTTCCTTAGATGATCCAAAGCTACTTGGATAACTGTGGTAATTCTAGAGCTAATA
70 80 90 100 110 120
Am_li_1060 N 1 CATGCCTACCAGCTCCGACCCGGTGGGCCTCGCTTCGGCTTTCCCTGTCACAGGGGGGAG
Me_em_988 N 2 CATGCCCAACCGCTCCGACCTGTAAAGGAAAGAGCGCTTTTATCAGCTCAAAACCAGTCT
By_au_1002 N 3 CATGCCCACCAGCTCCGACCCCTTCGCAAGGAGGGGAAAGAGCGCTTTTATTAGTTCAAA
Ma_sc_2599 N 4 CATGCCCAACAGCTCCGACCCGCTTGGGGCCCTCCTCGCAAGGGGGCGGTCGCCcGGCGG
Er_sp_643 N 5 ATGCCACTATGCCCTGACCCGCAAGGGAACGGGTGGATTTATTAGAACAGAACCAATCGG
Mo_ki_1364 N 6 CATGCCCGACAGCTCCGACCGTCGTCGCGTAACAGCGGCGGCGGGACGAGCGCTTTTATT
130 140 150 160 170 180
Am_li_1060 N 1 TCGGGTGGGGACTCCGTTGGGGAAGAGCGCTTTTATTAGTTCAAAaCCAGTCGGGCTTTC
Me_em_988 N 2 GCCGGCCTCAAAACCGGCAGTCCCTTGGTGAATCTGGATAACTTTTTGCCGATCGCATGG
By_au_1002 N 3 ACCAGTCGGGCCcTCACGGGTCCGTCCCCTTGGTGACTCTGGATAACTTTGTGCCGATCG
Ma_sc_2599 N 4 GGAAGAGCGCTTTTATTAGTTCAAAACCAGTCGGGgCTCCCAGCCCcGTCCTCTTTGGTG
Er_sp_643 N 5 TGGTGGCTTCGGCTGCTGCTGTTGCAATCTGGATGACTCTGGATAACTTCACTGATCGCG
Mo_ki_1364 N 6 AGTTGAAAACCAGtcggCCTCGCGGCCGTCCCCTTGGTGACTCTGGATAACTTTGAGCCG
190 200 210 220 230 240




XESEE.EXE
File Edit Search foRmat Command Options
Page 1 Seq 7 Pos 4
Am_li_1060 N 1 ccatgcATGTCTAAGTTCACACTCTGGTACAGTGAAACCGCGAATGGCTCATTAaatCAG
Me_em_988 N 2 ccatgcATGTCTAAGTTCACACCCTCGTATGGTGAAACCGCGAATGGCTCATTAAATCAG
By_au_1002 N 3 ccatgcATGTCTAAGTTCACACTCTCGTACGGTGAAACCGcgAATGGCTCATTAAATCAG
Ma_sc_2599 N 4 ccatGCATGTCTAAGTTCAcACTcTCGTACGGTGAAACCGcgAATGGCTCATTAAATCAG
Er_sp_643 N 5 ccatgcATGTCTAAGTACATACCTTTAAACGGTGAAACCGcgAATGGCTCATTAAATCAG
Mo_ki_1364 N 6 ccatgcATGTCTAAGTTCCCACTCTCGCACGGTGAAACCGcgAATGGCTCATTAAATCAG
10 20 30 40 50 60
Am_li_1060 N 1 TCGAGGTTCCTTTGATGATCCAAAT--CTACTTGGATAACTGTGGTAATTCTAGAGCTAA
Me_em_988 N 2 TCGAGGTTCCTTAGATGATCCAAAT--CTACTTGGATAACTGTGGTAATTCTAGAGCTAA
By_au_1002 N 3 TCGAGgttcCTTAGATGATCCAAAT--CTACTTGGATAACTGTGGTAATTCTAGAGCTAA
Ma_sc_2599 N 4 TCGAGGTTCCTTAGATGATCCAAAT--CTACTTGGATAACTGTGGTAATTCTAGAGCTAA
Er_sp_643 N 5 CTATGGTTCCTTAGATCGTA-CCTA--CTACATGGATAACTGTAGTAATTCTAGAGCTAA
Mo_ki_1364 N 6 TCGAGGTTCCTTAGATGATCCAAAG--CTACTTGGATAACTGTGGTAATTCTAGAGCTAA
70 80 90 100 110 120
Am_li_1060 N 1 TACATGCCTACCAGCTCCGACCCG-GGG------AAGAGCGCTTTTATTAGTTCAA-AaC
Me_em_988 N 2 TACATGCCCAACCGCTCCGACCTA-GGA------AAGAGCGCTTTTATCAGCTCAA-AAC
By_au_1002 N 3 TACATGCCCACCAGCTCCGACCCG-GGA------AAGAGCGCTTTTATTAGTTCAA-AAC
Ma_sc_2599 N 4 TACATGCCCAACAGCTCCGACCCGCCcGGCGGGGAAGAGCGCTTTTATTAGTTCAA-AAC
Er_sp_643 N 5 TACATGCCACTATGCCCTGACCCG-GGA------ACGGGTGGATTTATTAGAACAG-AAC
Mo_ki_1364 N 6 TACATGCCCGACAGCTCCGACCGGCGGCGGCGGGACGAGCGCTTTTATTAGTTGAA-AAC
130 140 150 160 170 180
Am_li_1060 N 1 CAGTCGG-TCCTTTTTG-GT---GACTCTG--G-ATAACTTTGTGCCGATCGCATCGGTC
Me_em_988 N 2 CAGTCTG-TCCCTT--G-GT---GAATCTG--G-ATAACTTTTTGCCGATCGCA-TGGC-
By_au_1002 N 3 CAGTCGG-TCCCCTT-G-GT---GACTCTG--G-ATAACTTTGTGCCGATCGCA-CGGC-
Ma_sc_2599 N 4 CAGTCGG-TCCTCTTTG-GT---GACTCTG--G-ATAACTTTGTGCCGATCGCA-CGGC-
Er_sp_643 N 5 CAATCGG-GCAATCT-GGAT---GACTCTG--G-ATAACTTCA--CTGATCGCGTCGGC-
Mo_ki_1364 N 6 CAGtcgg-TCCCCTT-G-GT---GACTCTG--G-ATAACTTTGAGCCGATCGCA-CGGC-
190 200 210 220 230 240

Amino acids/proteins



The "central dogma" of modern biology: DNA → RNA → protein

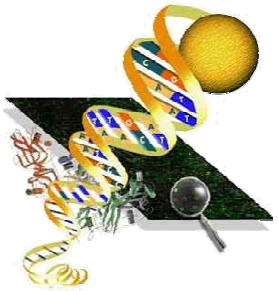
Getting from DNA to protein:

Two parts: 1. **Transcription** in which a short portion of chromosomal DNA is used to make a RNA molecule small enough to leave the nucleus.
2. **Translation** in which the RNA code is used to assemble the protein at the ribosome

The genetic code

- The code problem: 4 nucleotides in RNA, but 20 amino acids in proteins
- Bases are read in groups of 3 (= a **codon**)
- The code consists of **64** codons **(4^3 = 64)**
- All codons are used in protein synthesis:
  - 20 amino acids
  - 3 stop codons
- AUG (methionine) is the start codon (also used internally)
- The code is non-overlapping and punctuation-free
- The code is degenerate (but **NOT** ambiguous): each amino acid is specified by at least one codon
- The code is universal (virtually all organisms use the same code)

The genetic code



1	2 T	2 C	2 A	2 G		3
T	Phenylalanine **F**	Serine **S**	Tyrosine **Y**	Cysteine **C**	T	
	Leucine **L**		**STOP**	**STOP**	A	
				Tryptophan **W**	G	
C	Leucine **L**	Proline **P**	Histidine **H**	Arginine **R**	T	
			Glutamine **Q**		A	
					G	
A	Isoleucine **I**	Threonine **T**	Asparagine **N**	Serine **S**	T	
					C	
			Lysine **K**	Arginine **R**	A	
	Methionine **M**				G	
G	Valine **V**	Alanine **A**	Aspartate **B**	Glycine **G**	T	
					C	
			Glutamate **Z**		A	
					G	

In-class exercise

1. Which amino acids are specified by single codons?

   methionine and tryptophan

2. How many amino acids are specified by the first two nucleotides only?

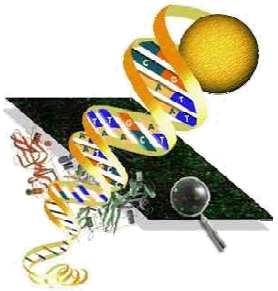
   five: proline, threonine, valine, alanine, glycine

3. What is the RNA code for the start codon?

   AUG

DNA
RNA
Amino acids/ proteins
PCR, sequencing
Mutations

Reading frames



Reading frame (also open reading frame):

The stretch of triplet sequence of DNA that potentially encodes a protein. The reading frame is designated by the initiation or start codon and is terminated by a stop codon.

- a reading frame is not always easily recognizable
- each strand of RNA/DNA has three possible starting points (position one, two, or three):

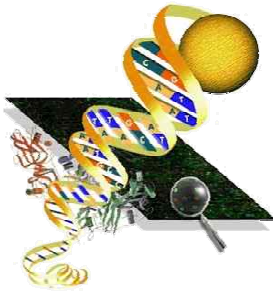
Position 1	CAG AUG AGG UCA GGC AUA	
	gln met arg ser gly ile	
Position 2	C AGA UGA GGU CAG GCA UA	
	arg trp gly gln ala	
Position 3	CA GAU GAG GUC AGG CAU A	
	asp glu val arg his	

- mutations within an open reading frame that delete or add nucleotides can disrupt the reading frame (**frameshift mutation**):

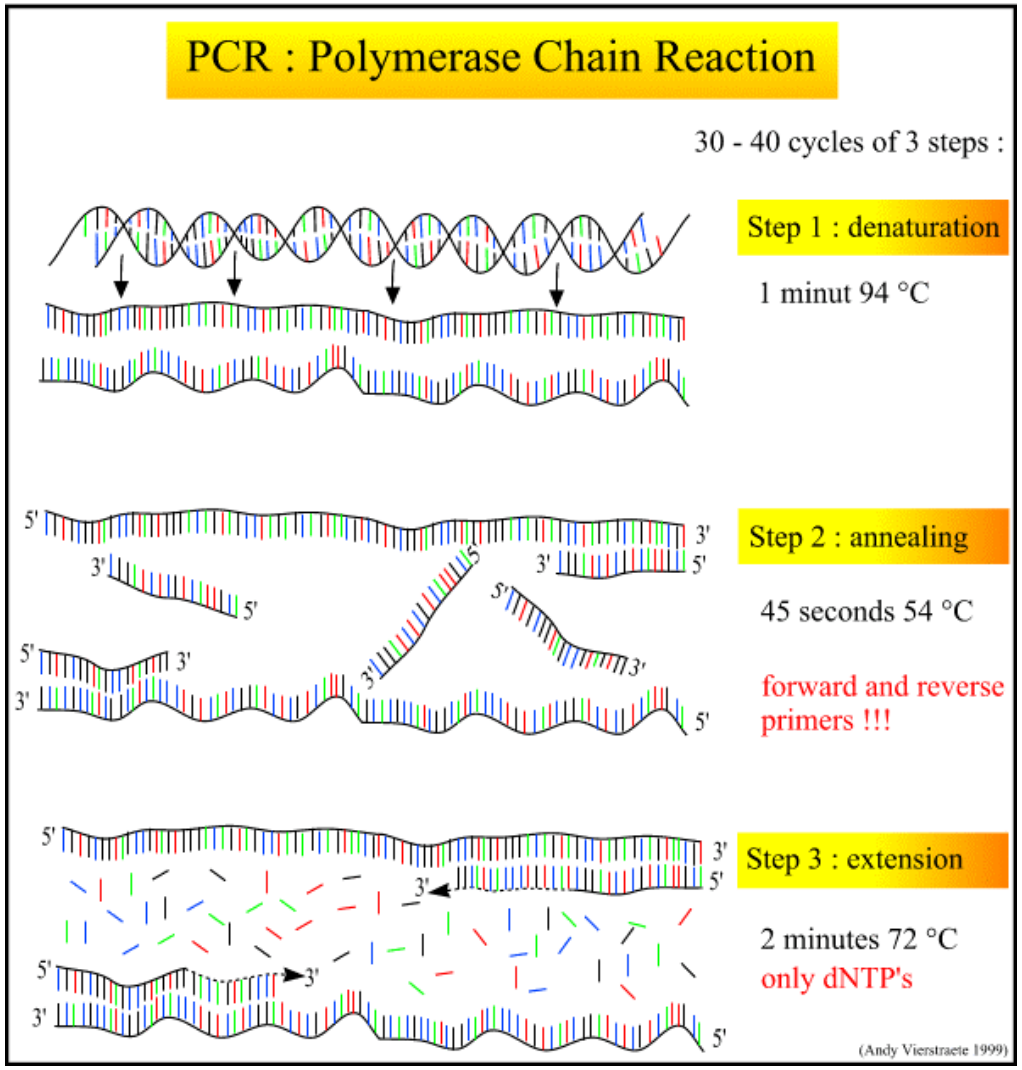
Wildtype	CAG AUG AG**G** UCA GGC AUA GAG	
	gln met arg ser gly ile glu	
Mutant	CAG AUG AG**U** CAG GCA UAG AG	
	gln met ser gln ala	

Up to 30% of mutations causing humane disease are due to premature termination of translation (nonsense mutations or frameshift)

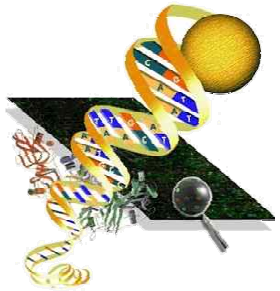
PCR



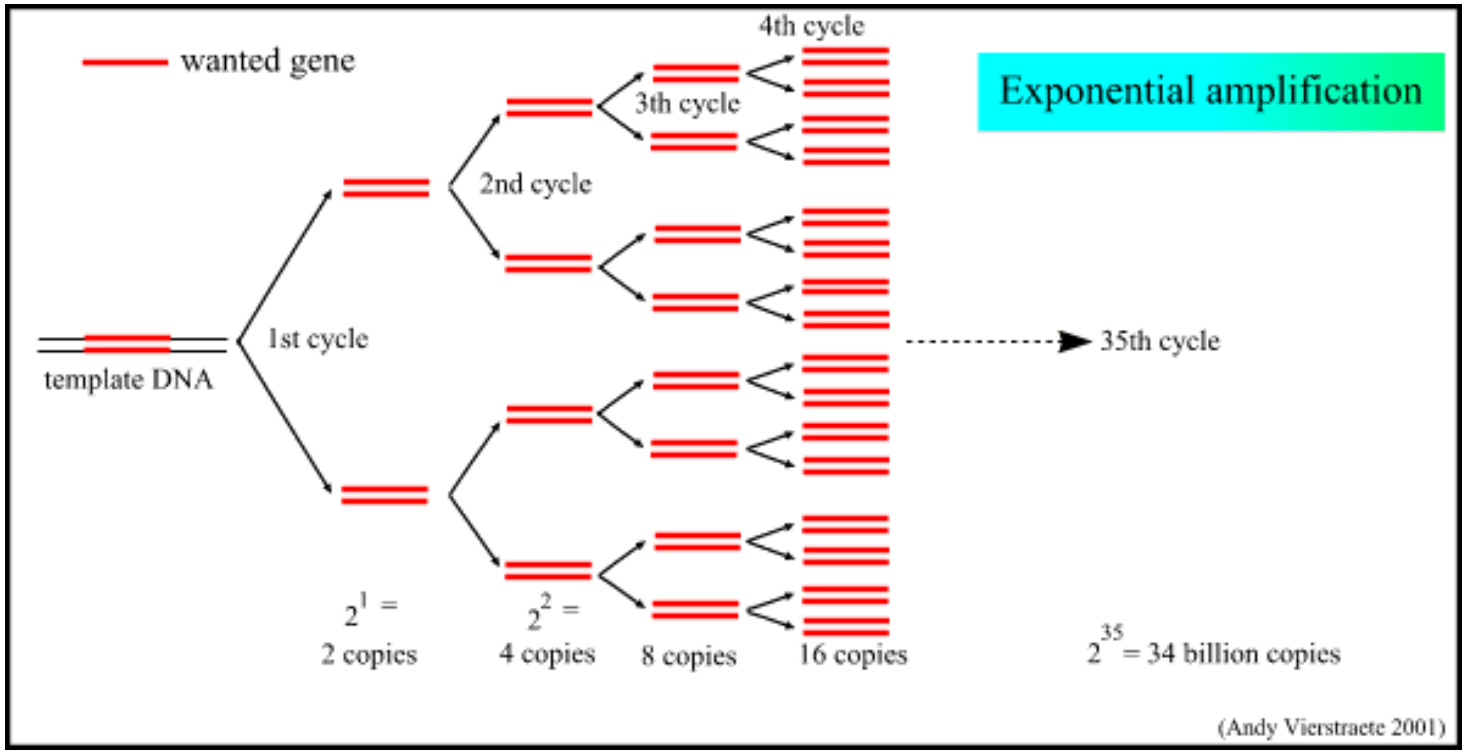
The different steps of PCR



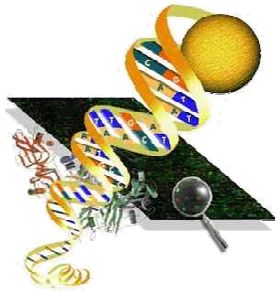
PCR



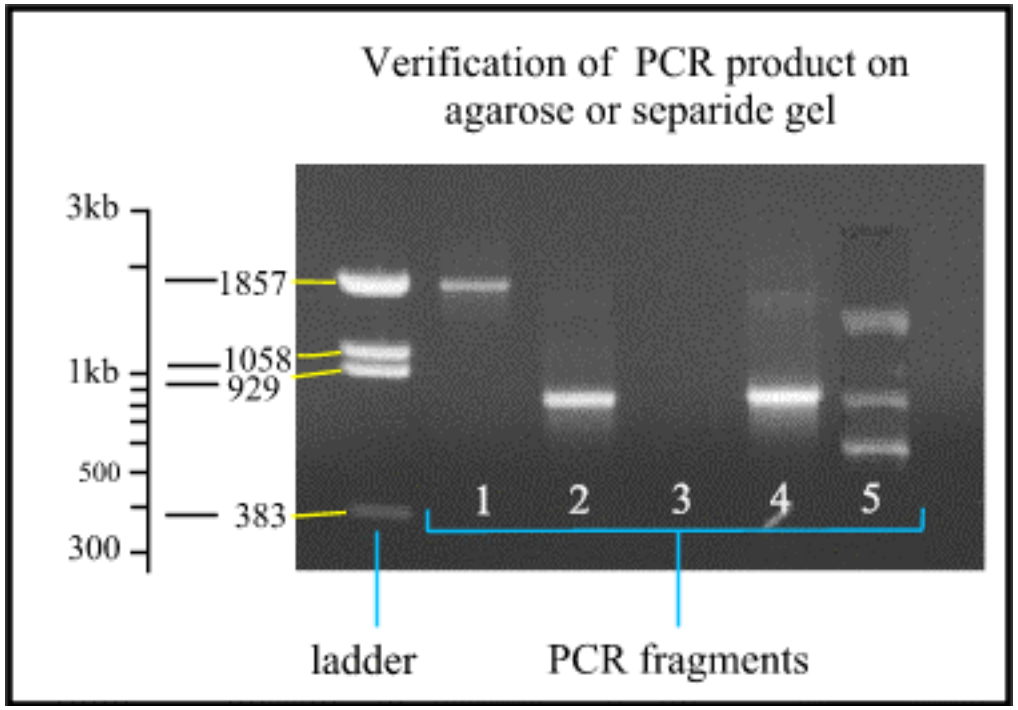
Exponential increase of the number of copies during PCR



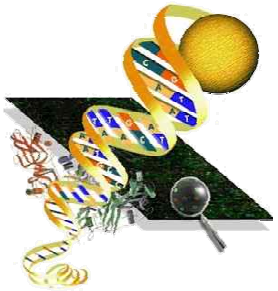
PCR



Verification of PCR product on gel



PCR



General notes on primer design in PCR

- Perhaps the most critical parameter for successful PCR is the design of primers

Primer selection

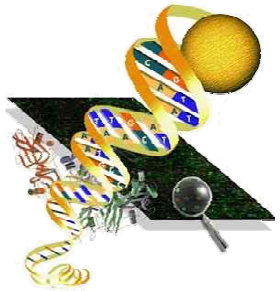
Critical variables are:

- primer length
- melting temperature (T_m)
- specificity
- complementary primer sequences
- G/C content
- 3'-end sequence

Primer length

- specificity and the temperature of annealing are at least partly dependent on primer length
- oligonucleotides between 18 and 24 bases are highly sequence specific
- primer length is proportional to annealing efficiency: in general, the longer the primer, the more inefficient the annealing
- the primers should not be too short as specificity decreases

Primer design



Melting temperature (T_m)

- the goal should be to design a primer with an annealing temperature of at least 50°C
- the relationship between annealing temperature and melting temperature is one of the "Black Boxes" of PCR
- a general rule-of-thumb is to use an annealing temperature that is 5°C lower than the melting temperature
- the melting temperatures of oligos are most accurately calculated using nearest neighbor thermodynamic calculations with the formula:

$$T_m = H\,[S + R \ln (c/4)] - 273.15\ °C + 16.6 \log_{10} [K^+]$$

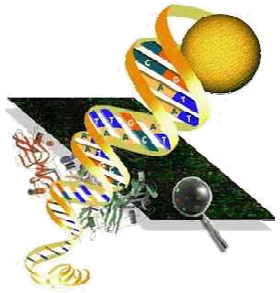
(H is the enthalpy, S is the entropy for helix formation, R is the molar gas constant and c is the concentration of primer)

- a good working approximation of this value can be calculated using the **Wallace** formula:

$$T_m = 4x\ (\#C + \#G) + 2x\ (\#A + \#T)\ °C$$

- both of the primers should be designed such that they have similar melting temperatures. If primers are mismatched in terms of Tm, amplification will be less efficient or may not work: the primer with the higher Tm will mis-prime at lower temperatures; the primer with the lower Tm may not work at higher temperatures.

Primer design



Specificity

- primer specificity is at least partly dependent on primer length: there are many more unique 24 base oligos than there are 15 base pair oligos

Complementary primer sequences

- primers need to be designed with absolutely no intra-primer homology beyond 3 base pairs. If a primer has such a region of self-homology, “snap back” can occur
- another related danger is inter-primer homology: partial homology in the middle regions of two primers can interfere with hybridization. If the homology should occur at the 3' end of either primer, primer dimer formation will occur

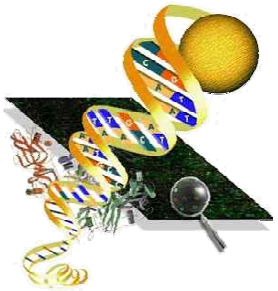
G/C content

- ideally a primer should have a near random mix of nucleotides, a 50% GC content
- there should be no PolyG or PolyC stretches that can promote non-specific annealing

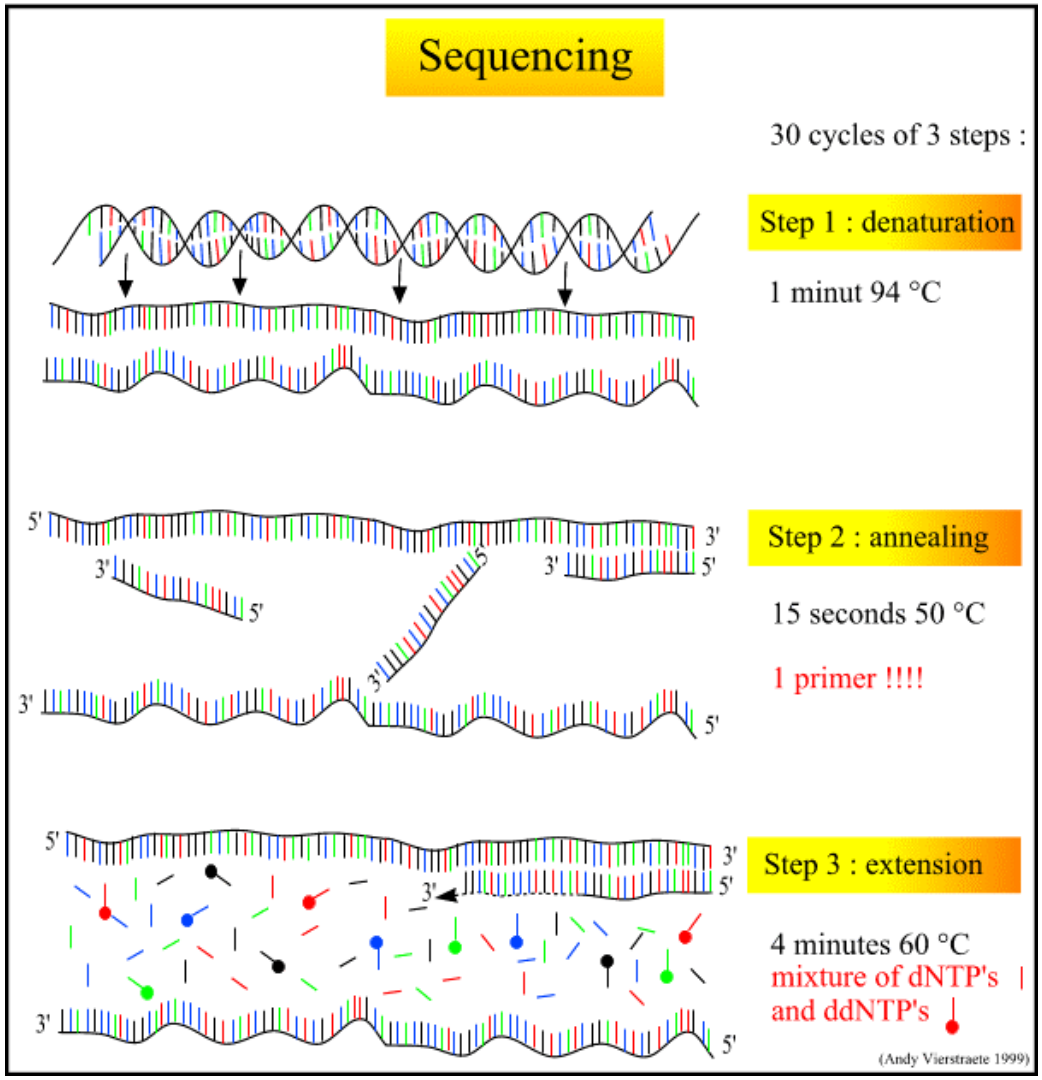
3’-end sequence

- the 3' terminal position in PCR primers is essential for the control of mis-priming
- inclusion of a G or C residue at the 3' end of primers helps to ensure correct binding (stronger hydrogen bonding of G/C residues)

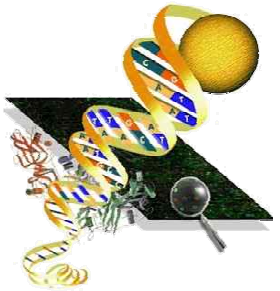
Sequencing



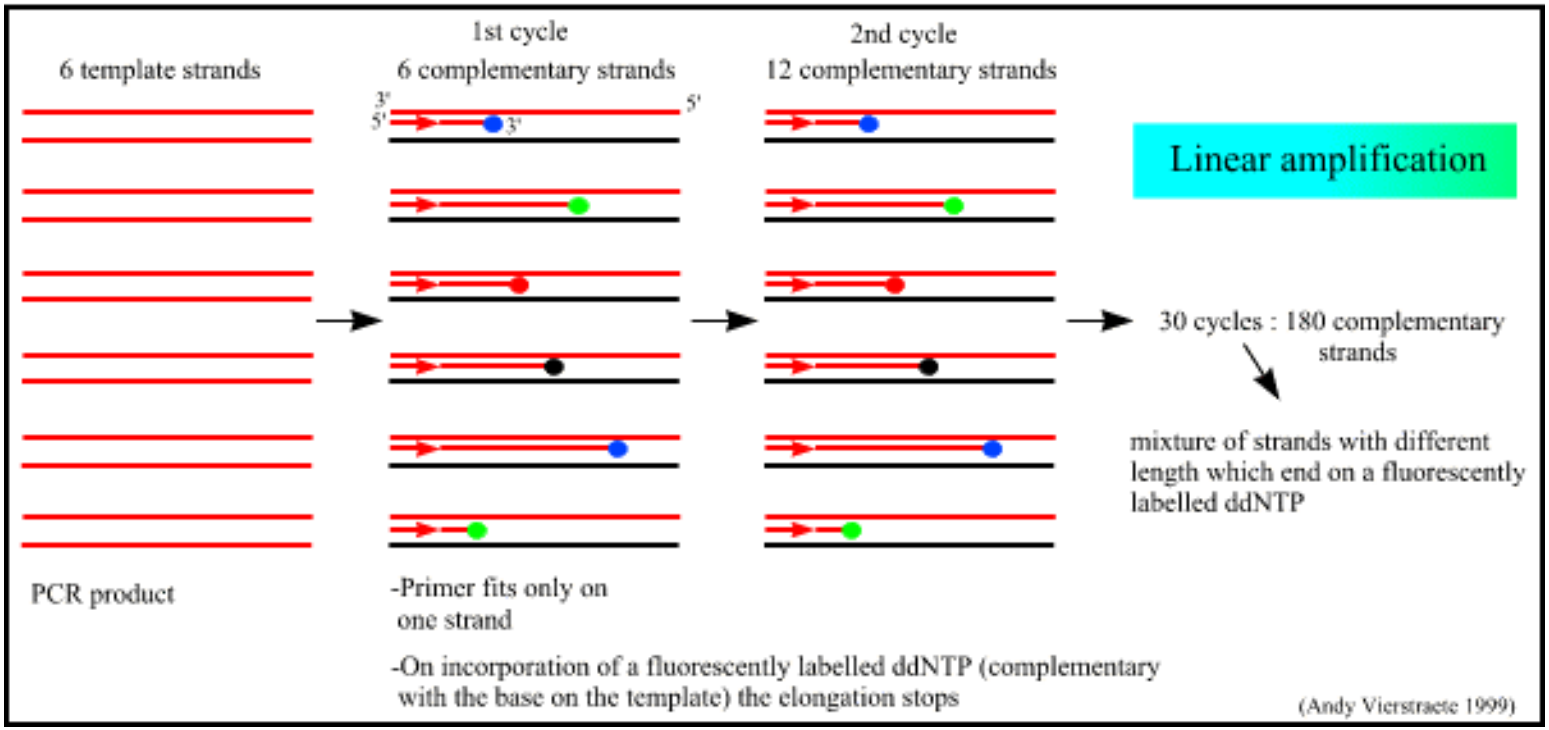
The different steps in sequencing



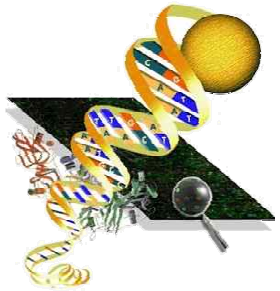
Sequencing



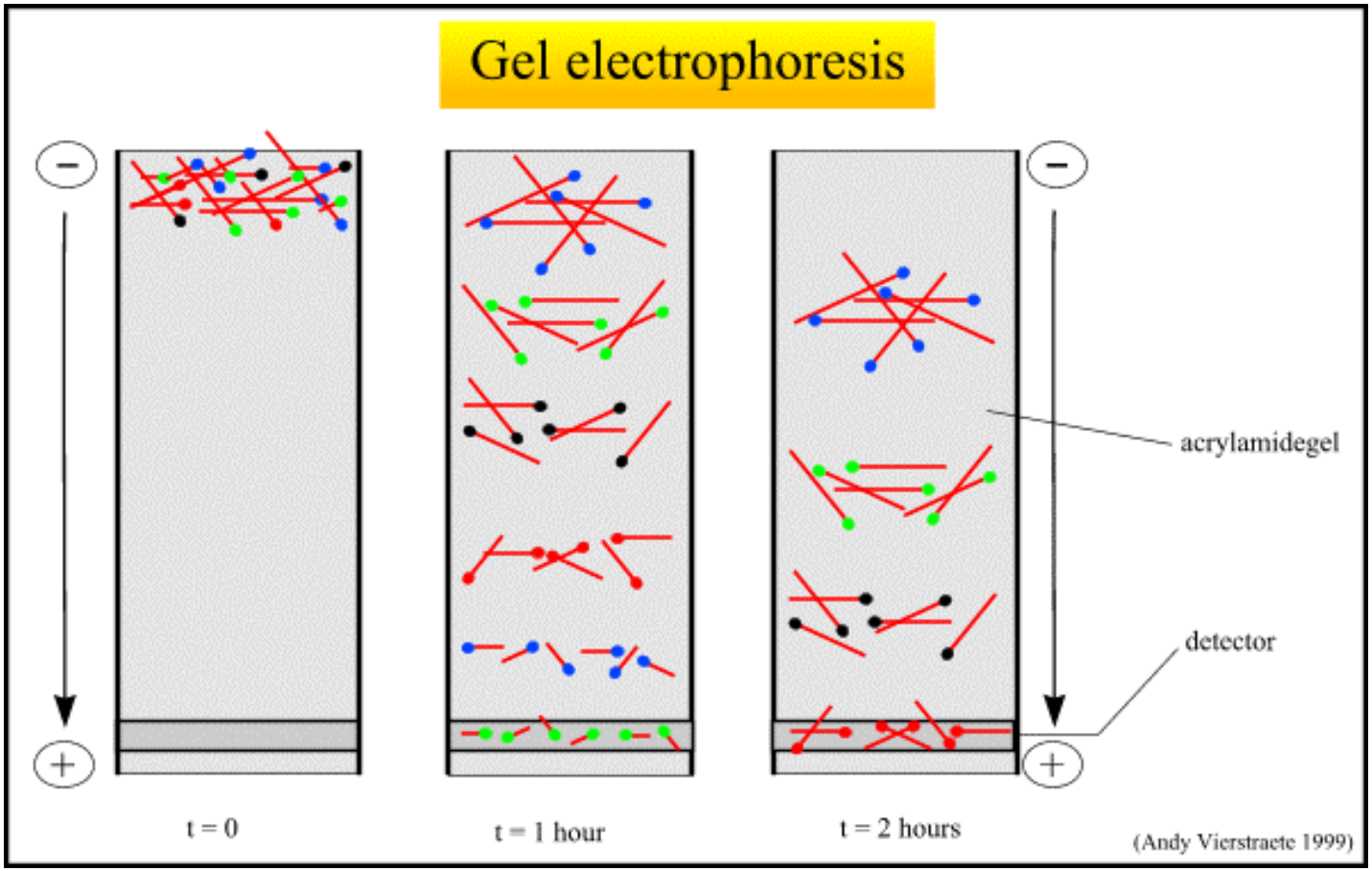
The linear amplification of the gene in sequencing



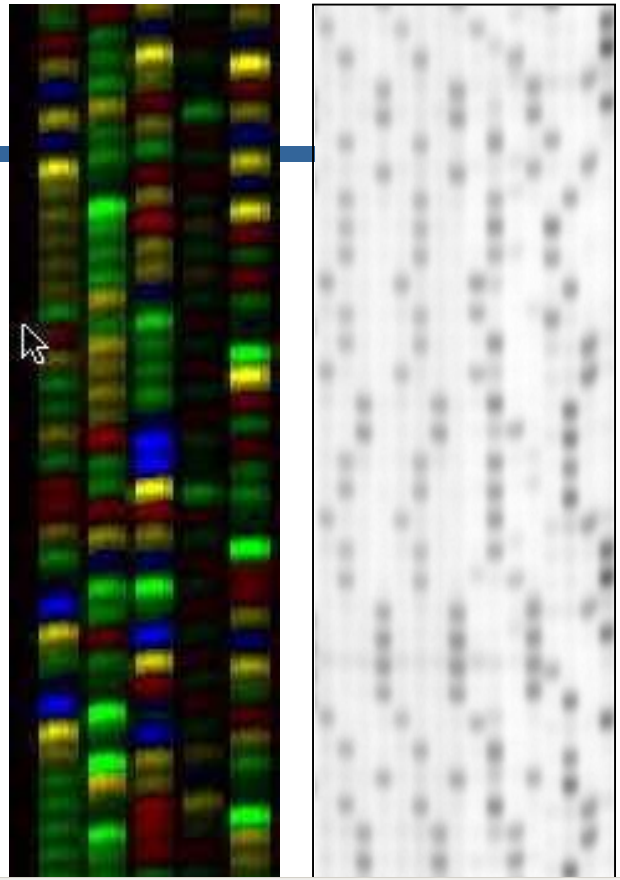
Sequencing

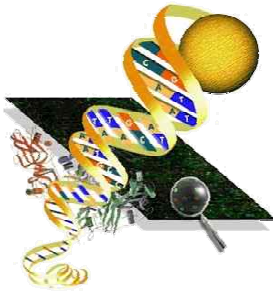


The separation of the sequencing fragments

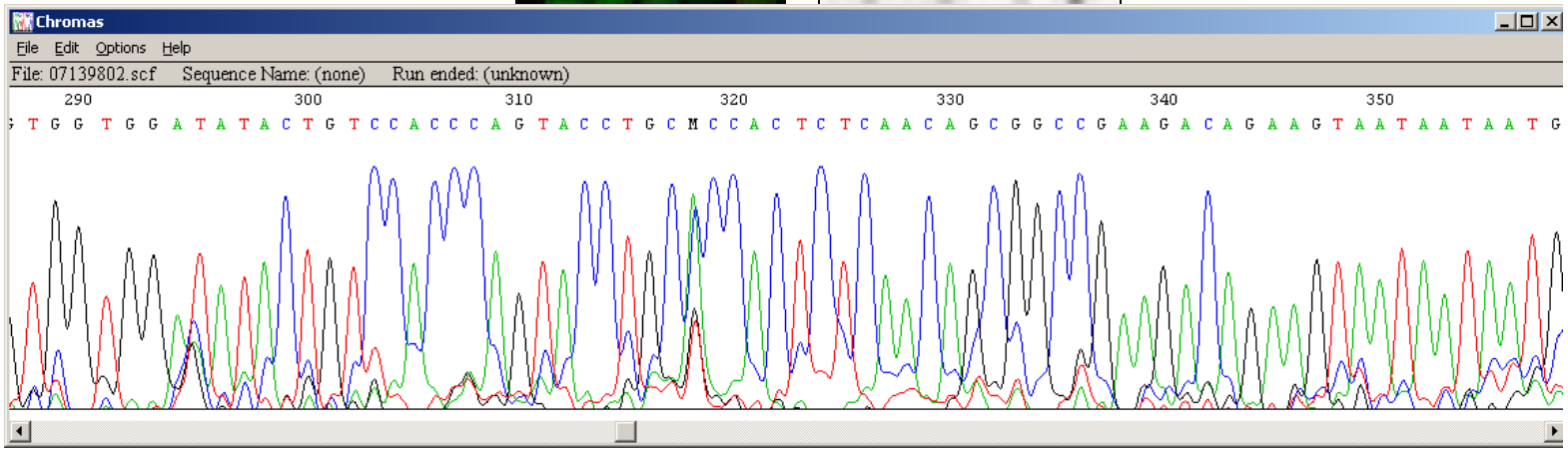


Sequencing

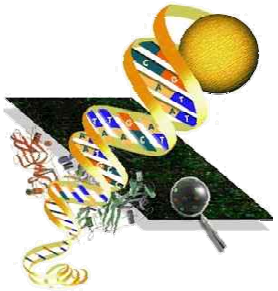




Chromatogram file



Mutations



Mutation: any heritable change in DNA

Sources of mutation:

Spontaneous mutations: mutations occur for unknown reasons

Induced mutations: exposure to substance (mutagen) known to cause mutations, e.g. X-rays, UV light, free radicals

Mutations may influence one or several base pairs

a) Nucleotide substitutions (point mutation)

1) Transitions (Pu ↔ Pu; Py ↔ Py)
2) Transversions (Pu ↔ Py)

b) Insertion or deletion ("indels")

- one to many bases can be involved
- frequently associated with repeated sequences ("hot spots")
- lead to frameshift in protein-coding genes, except when N = 3X
- also caused by insertion of transposable elements into genes

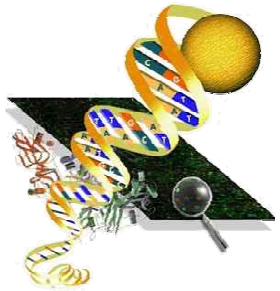
In-class exercise

How many transition and transversion events are possible?

2 transitions: T ↔ C; A ↔ G
4 transversions: T ↔ A; T ↔ G
C ↔ A; C ↔ G

"Weighting" of mutation events plays important role for phylogenetic analyses (model of sequence evolution)

Mutations



Mutations may influence phenotype

a) Silent (or synonymous) substitution

- nucleotide substitution without amino acid change
- no effect on phenotype
- mostly third codon position
- other possible silent substitutions: changes in non-coding DNA

b) Replacement substitution

- causes amino acid change
- **neutral**: protein still functions normally
- **missense**: protein loses some functions (e.g. sickle cell anemia: mutation in ß-globin)

c) Sense/nonsense substitution

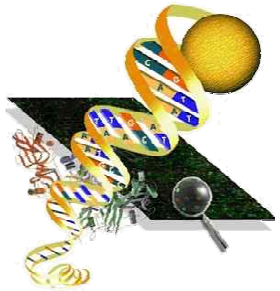
- **sense:** involves a change from a termination codon to one that codes for an amino acid
- **nonsense:** creates premature termination codon

Mutation rates

= a measure of the frequency of a given mutation per generation

- mutation rates are usually given for specific loci (e.g. sickle cell anemia)
- the rate of nucleotide substitutions in humans is on the order of 1 per 100,000,000
- range varies from 1 in 10,000 to 1 in 10,000,000,000
- every human has about 30 new mutations involving nucleotide substitutions
- mutation rate is about twice as high in male as in female meiosis

Next week



First quiz

Lecture 1

- Bioinformatics definitions
- The human genome project

Lecture 2

- Amino acids
- Proteins

Lecture 3

- DNA structure
- RNA structure
- Basics of PCR and sequencing
- Mutations

DNA

RNA

Amino acids/ proteins

PCR, sequencing

Mutations