Introduction to Bioinformatics

Goal

Goal of molecular cell biology - to understand the physiology of living cells in terms of the information that is encoded in the genome of the cell

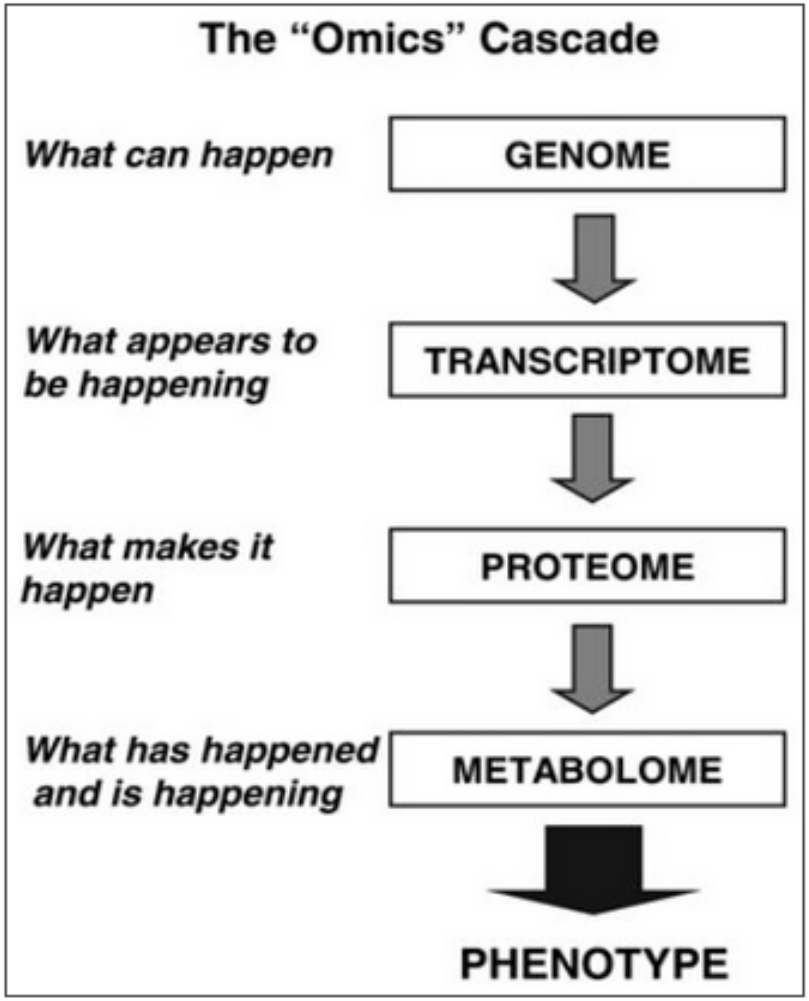
How computer science can help in achieving this goal?

⇒ Location of the genes in the genome, its function, what factors affect its expression, in normal *vs* disease state, etc.

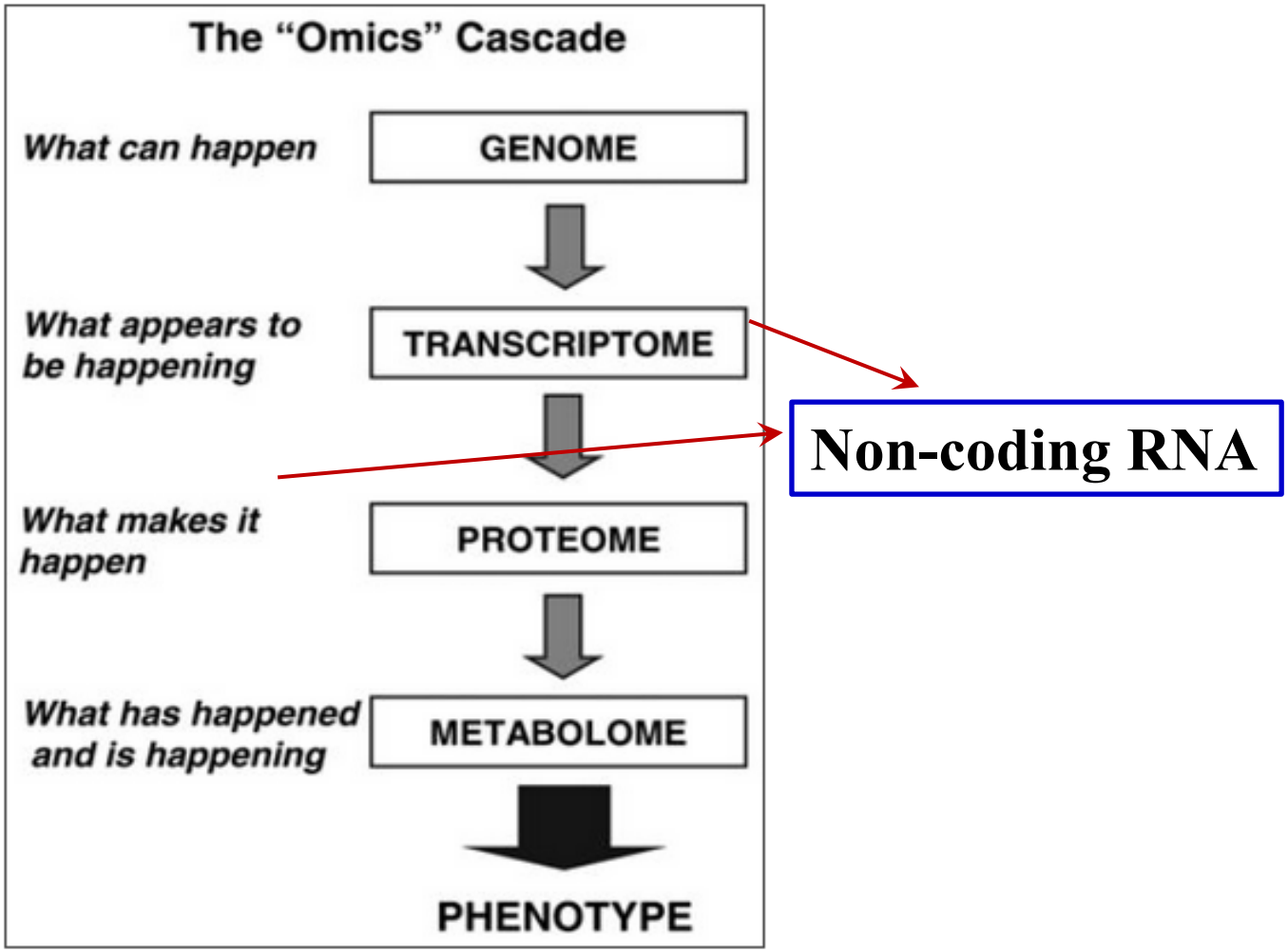
'Bioinformatics' was coined by Paulien Hogeweg in 1979, for the study of informatic processes in biological systems.

- it's a data-driven field to gain insight into what happens in the living cell of an organism using various types of biological data

Various Omics studies, *viz.*, Genomics, Transcriptomics, Proteomics & Metabolomics are data-driven fields that aim to answer the question of how genomes code for living organisms.



Major inputs from CS – develop algorithms for mining meaningful information from biological data, and develop efficient data storage and data retrieval systems for managing large volumes of data



Major inputs from CS – develop algorithms for mining meaningful information from biological data, and develop efficient data storage and data retrieval systems for managing large volumes of data

Biological Data: Levels of Organization

Central Dogma of MB

Central Dogma of Bioinformatics

Metabolic Pathways, Interacting Networks

Intercellular Interactions

Organs/Tissues Physiology

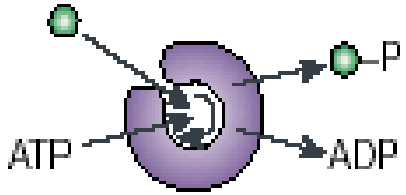
Ecological modeling
Inter-species interactions

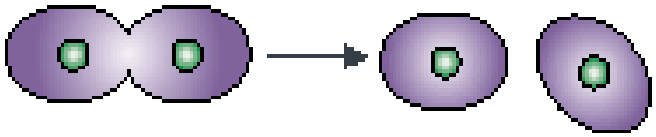
DNA → mRNA → Protein → Enzyme → Reaction network → Cell physiology

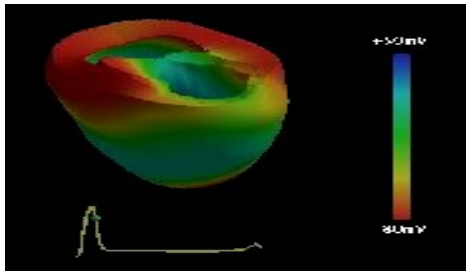
...TACCCGATGGCGAAATGC...

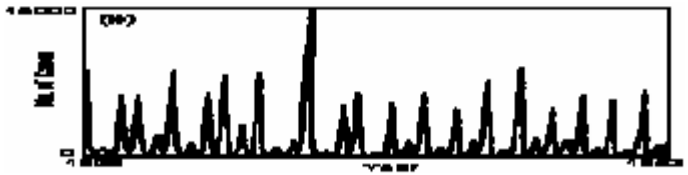
...AUGGGCUACCGCUUUACG...

...Met-Gly-Tyr-Arg-Phe-Thr...









Sequence/Structure Alignment, Pattern Recognition, Molecular Modeling

Network Modeling, Dynamical Systems Modeling, Differential Eqns, Graph Theory, Chaos Theory, Pattern Formation, Characterization, Time-series data analysis

If it were required, in a single model, to span all the scales from

molecular motions ⇒ **cell responses**

(nanometers/picosecs) **(micrometers/secs)**

- theoretical approach to cell physiology would be beyond grasp, both computationally & intellectually

Fortunately, considerable progress can be made – at any given level of hierarchy – independent of the successes or failures at levels above/below

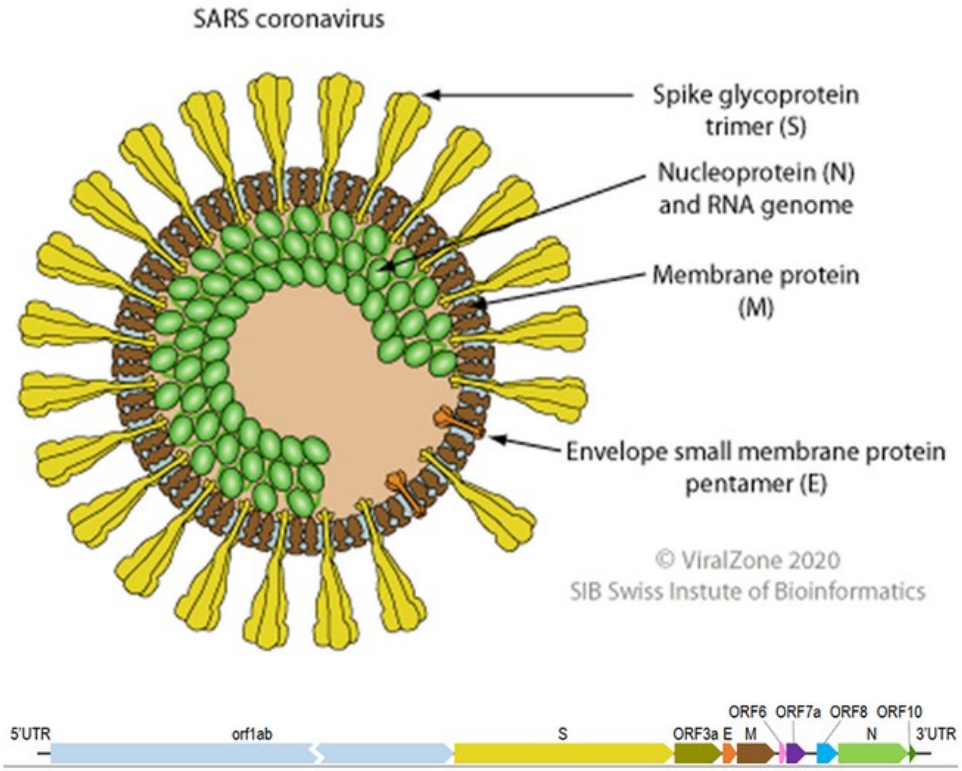
Systems biology is aiming towards achieving this goal of understanding the functional behaviour of a cell/tissue as a whole.

Disease - COVID19

- When a new virus strikes the population, such as we saw in the case of COVID-19, no specific treatment is available

What kind of sequence analysis can help in combating the disease?

SARS-CoV-2



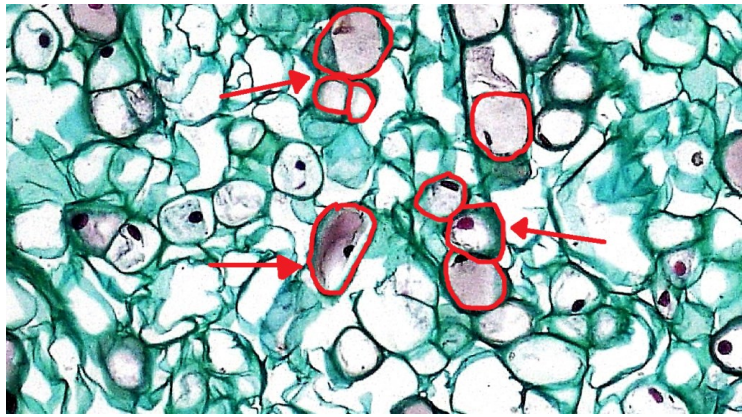
What kind of Bioinformatics analysis can we carry out to know about the virus causing COVID-19?

- How to identify if a person is infected with SAR-COV-2?
- Is it the only known human coronavirus?
- Comparing its genome with other viral genomes – to identify its closest relative
- What proteins aid in its transmission and infection?
- Identifying drug targets and develop vaccines
- What organs/tissues are affected by its infection?
- Its rate of propagation
- Is it mutating and becoming more virulent, or milder with time
- etc.

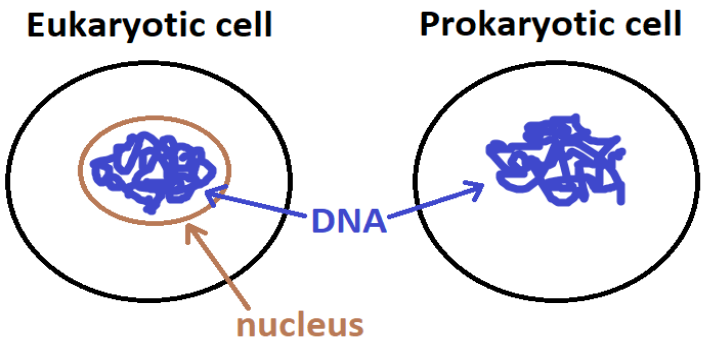
The Cell

Cell - the basic building block of all living creatures.

All forms of "omics" measure large quantities of something found inside cells.



Cells are of two types:

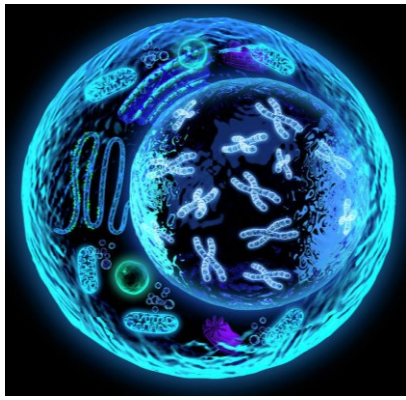


Original image

Apart from prokaryotes and eukaryotes, there is a third category called Viruses, acellular entities. May contain DNA or RNA as their nuclear material.

Cells and Chromosomes

E. B. Wilson: "the key to every biological problem must be sought in the cell; for every living organism is, or at some time has been, a cell."



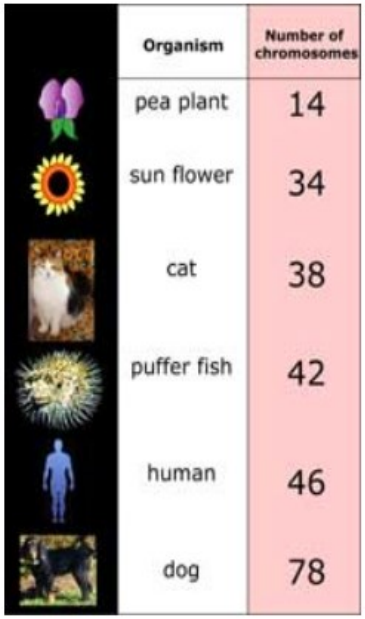
~ 10^{13} cells that form a human body, the whole organism has been generated by cell divisions from a single cell

Cells are the fundamental units of life - the vehicle for all the hereditary information that defines each species.

Genome – the total DNA content of an organism

Chromosomes – are physically separate molecules that range in length from ~ 50 - 250Mbp

In mammals and many other eukaryotes, the chromosomes occur in homologous pairs, called diploids, except for sex chromosomes.



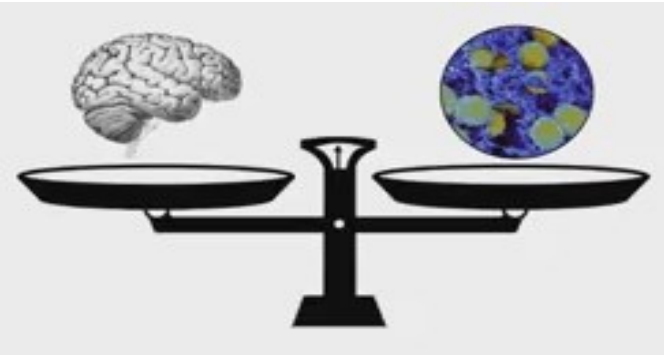
Organism	Number of chromosomes
pea plant	14
sun flower	34
cat	38
puffer fish	42
human	46
dog	78

Cells and Genomes

Do we carry cells of any other organism within us, apart from human cells?

How human are we?

- **We have 10 trillion human cells and 100 trillion microbial cells: with respect to cell count we are just 10% human**
- **Our genome has 20-30K genes, our microbiome has 2-20M genes: with respect to genes we are 0.1-1% human**
- **Our microbiome weighs ~ 3 pounds, about the same weight as our brain, and maybe as important to our well being, if not more!**
- **We share 99.9% of our genome with other individuals, but we share only 10% of our microbiome**
- **Microbiota include bacteria, archaea, viruses, eukaryota**



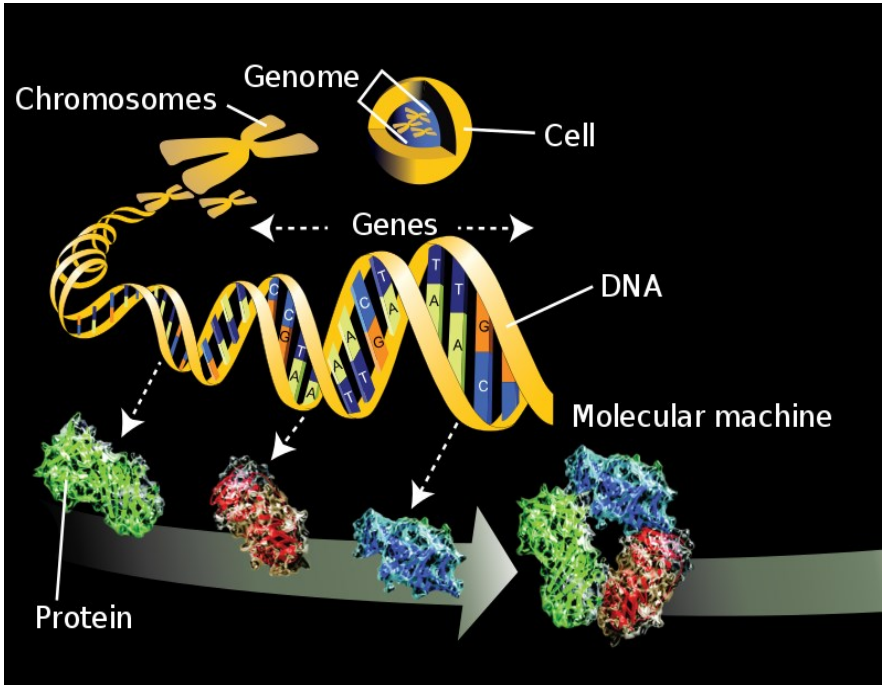
DNA and Genes

Human DNA - a long sequence of 3 billion letters, inside every cell in our body. There are ~ 37.2 trillion cells in our body.

A "gene" is a particular segment of DNA that encodes instructions for making a protein, hence genes are referred to as the "coding" part of the genome.

~ 25,000 genes covering ~ 2-3% of the human genome.

Function of remaining 98% of "non-coding" part of the genome?



DNA

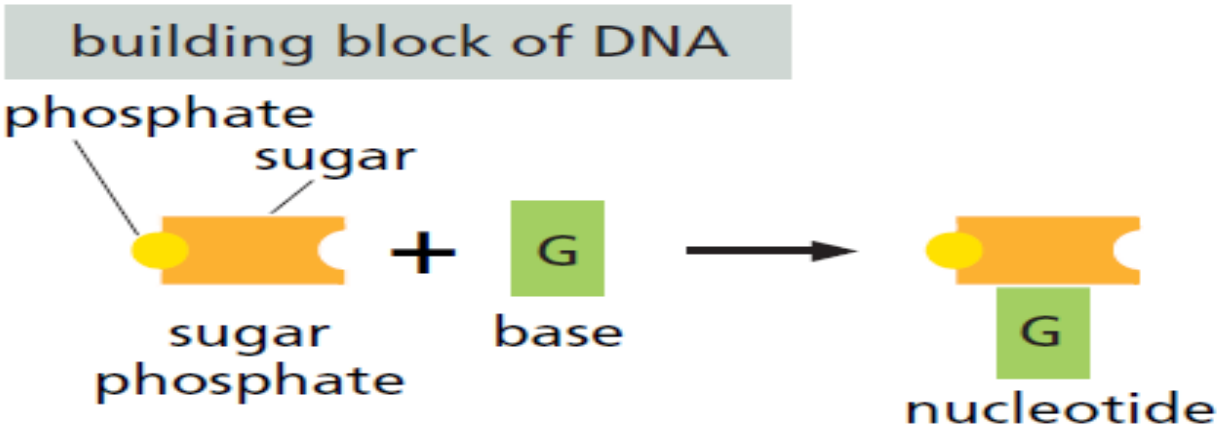
DNA (Deoxyribonucleic acid):

Composed of four basic units - called nucleotides

Each nucleotide contains - a sugar, a phosphate and one of the four bases:

Adenine (A), Thymine (T),

Guanine (G), Cytosine (C).

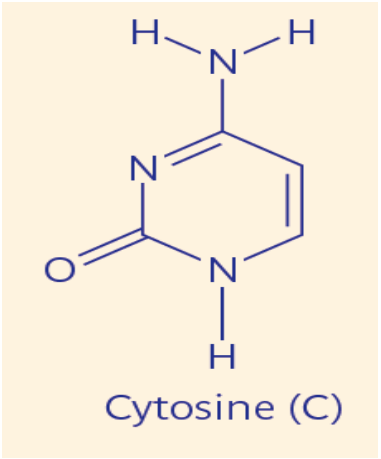


DNA

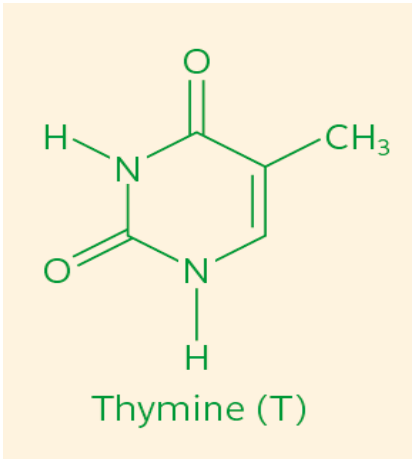
Bases: are ring-shaped and come in four types which fit together in pairs - this pairing forms the basis of information carrying capacity of DNA.

These are categorized as:

Pyrimidines

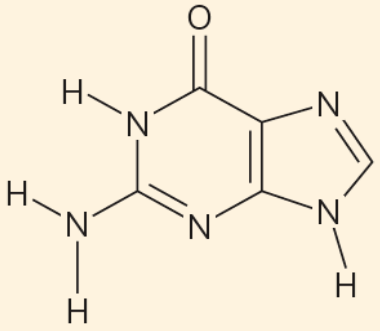


Cytosine (C)

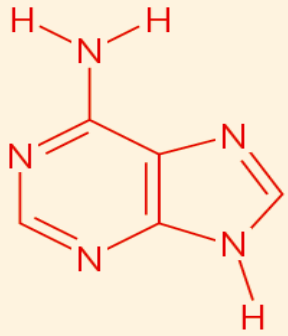


Thymine (T)

Purines



Guanine (G)



Adenine (A)

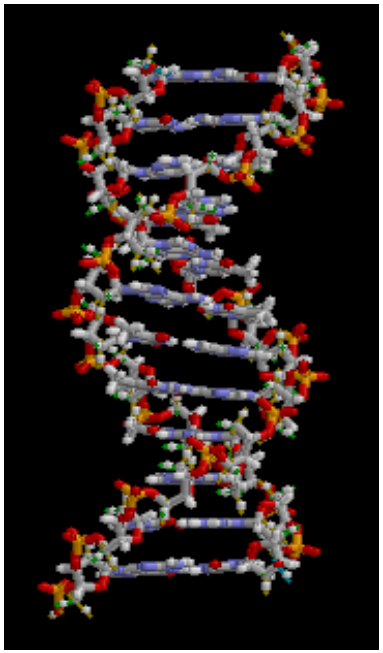
Which of these form base-pairs?

DNA

DNA is double-stranded - the two strands of DNA wind around each other to form a double helix.

Information in one strand is a "mirror copy" of the information in the other strand, achieved by base-pairing:

$$A \Leftrightarrow T, \qquad G \Leftrightarrow C$$



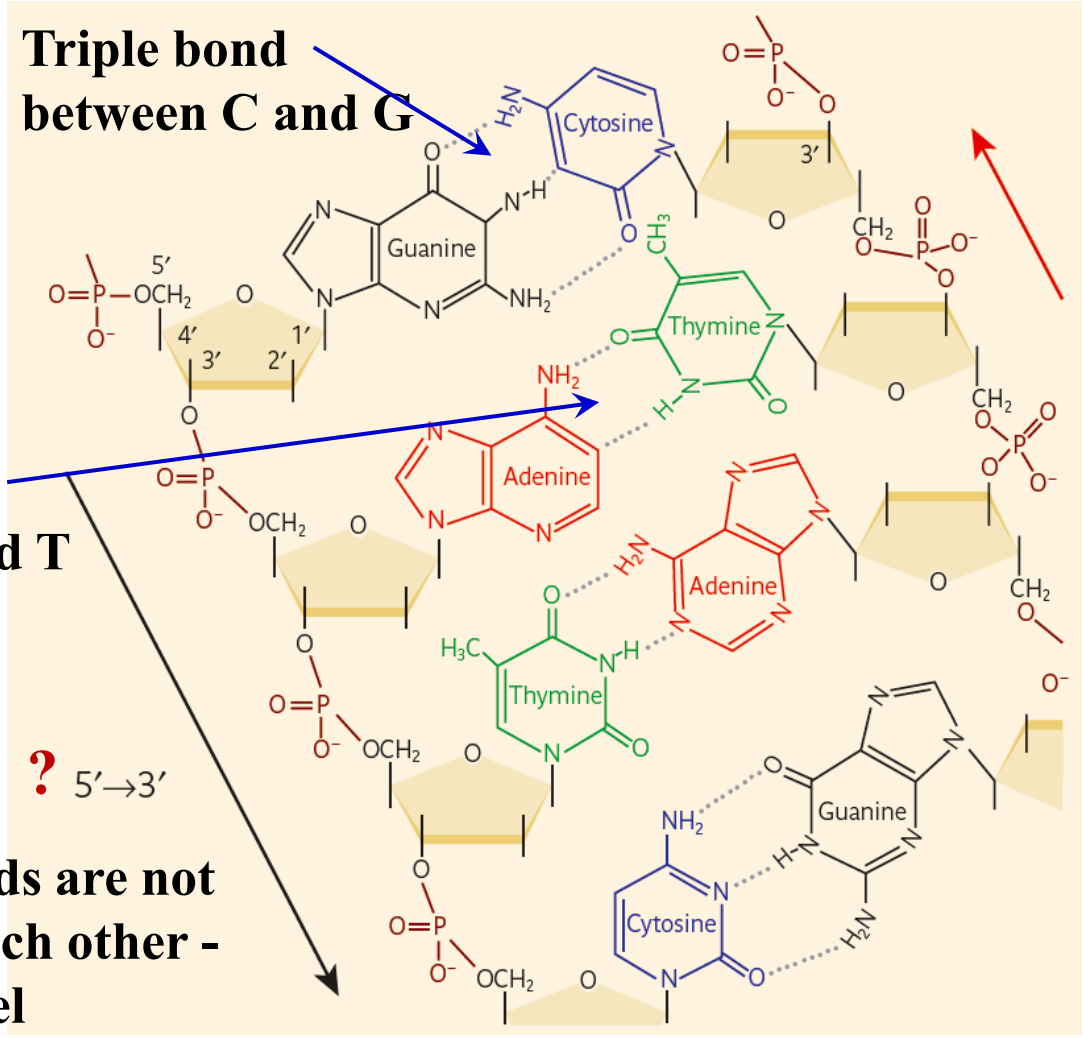
So, if the sequence on one strand is GATTACA, what is the sequence of the other strand?
Importance of double-stranded nature of DNA:

- **Facilitates DNA replication**
- **Error-correct the genome**
- **Provides stability**

From a computational perspective, the sequence of only one strand is needed.

DNA

Base Pairing: **If two polynucleotide strands face each other, sugar-phosphate backbone runs down each side, and complementary pairs of bases in the middle spontaneously form hydrogen bonds:**



Note: the two strands are not mirror images of each other - they run antiparallel

Double-Stranded DNA: **If the sequence in the forward strand in 5′ to 3′ direction is:**

5′ CATTGCCAGT 3′

Then what is the sequence on the reverse strand when read in 5′ to 3′ orientation?

DNA

If the sequence in the forward strand in 5′ to 3′ direction is:

5′ CATTGCCAGT 3′

Then what is the sequence on the reverse strand when read in 5′ to 3′ orientation?

First write its complement:

5′ CATTGCCAGT 3′

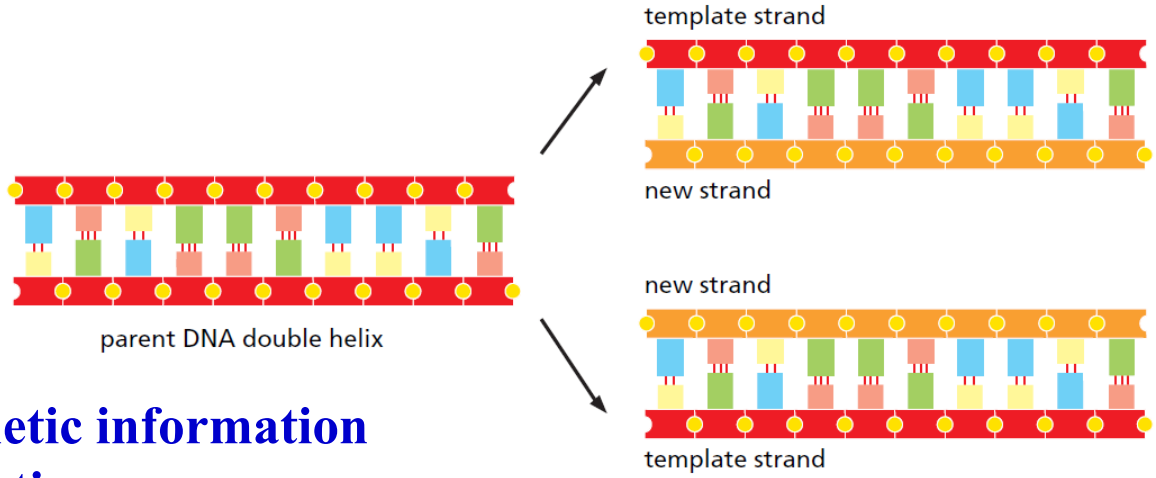
3′ GTAACGGTCA 5′

When read in 5′ to 3′ orientation, the sequence on the reverse strand is:

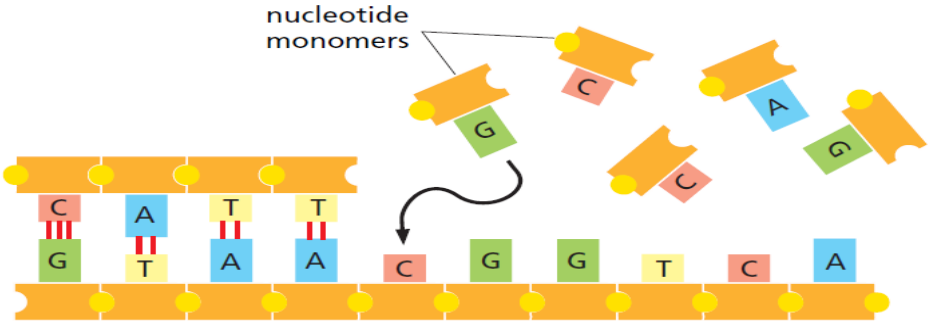
5′ ACTGGCAATG 3′

DNA Replication

In living cells DNA is not synthesized as a free strand in isolation, but on a template formed by a pre-existing DNA strand.

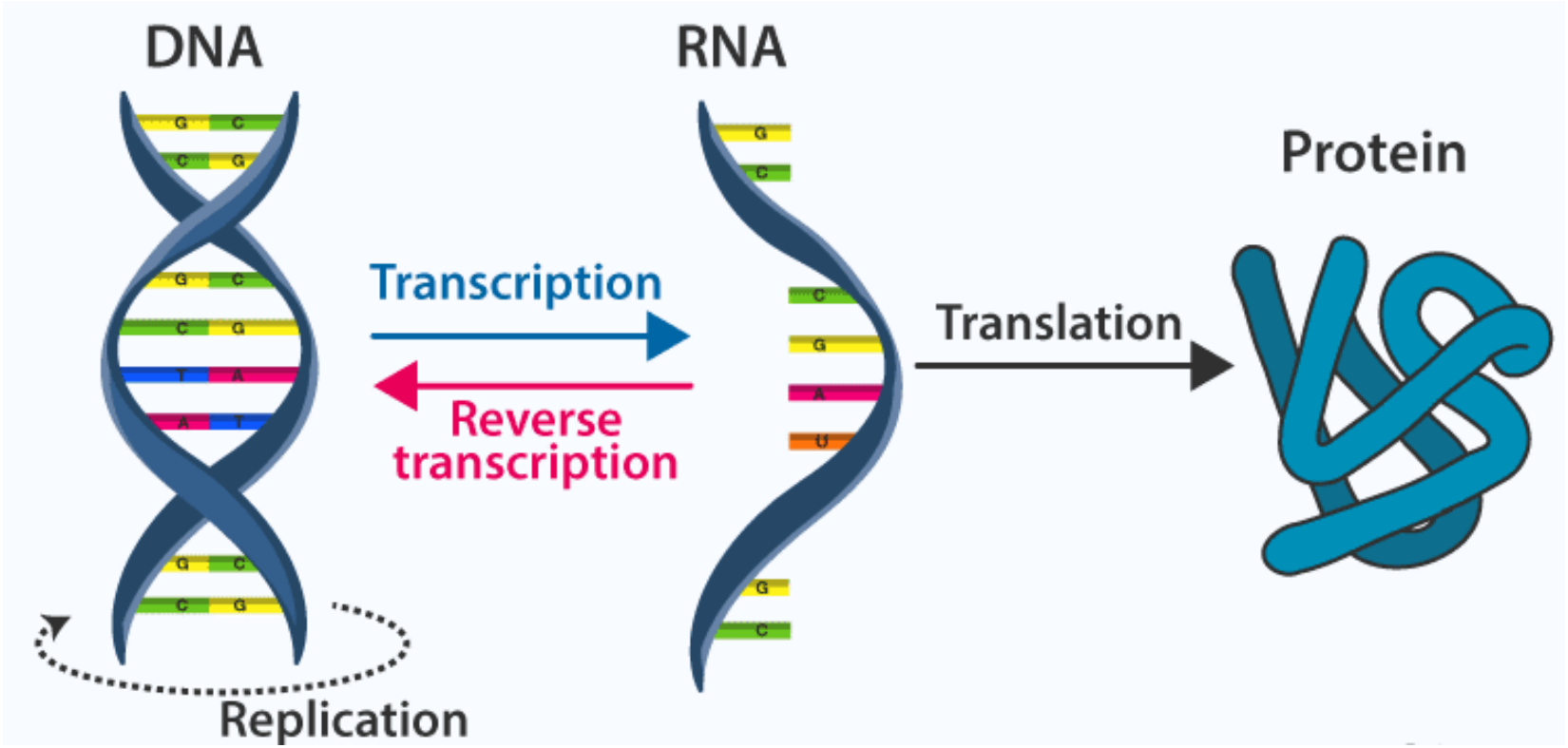


Copying of genetic information by DNA replication



templated polymerization

Central Dogma of Molecular Biology



Ribonucleic Acid (RNA):

It is single-stranded molecule

Composed of four basic units - called nucleotides:

Each nucleotide contains - a sugar (ribose), a phosphate and one of the four bases: Adenine (A), Uracil (U), Guanine (G), Cytosine (C)

RNA polynucleotide strand is built by creating a phosphodiester bond between nucleotides.

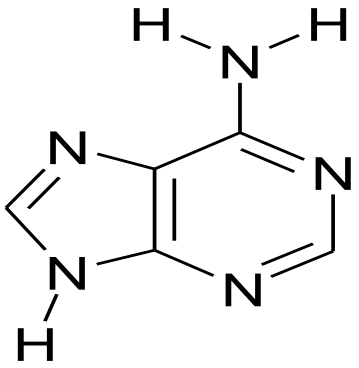
Intra-strand base pairing is a characteristic feature of RNA

Base Pairing – formed by weak H-bonds and follows the following complementarity rule:

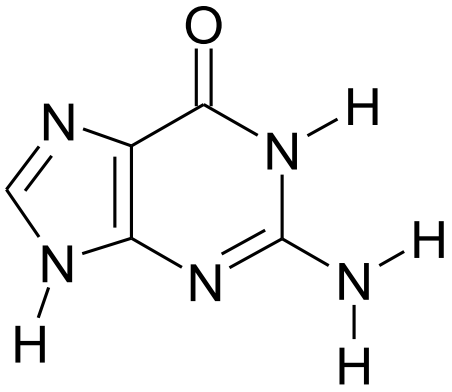
G ⟷ C, A ⟷ U, and G ⟷ U

Ring Structure of Nucleic Acid bases

Adenine (A)

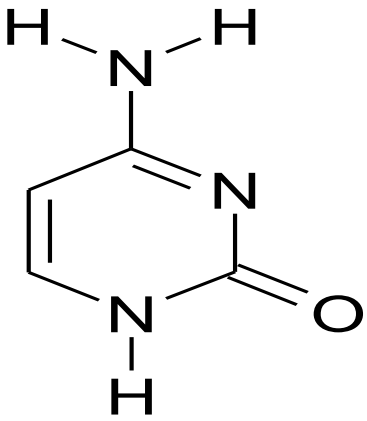


Guanine (G)

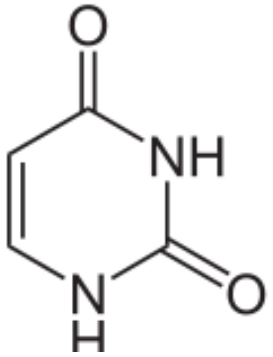


Purines

Cytosine (C)



Uracil (U)

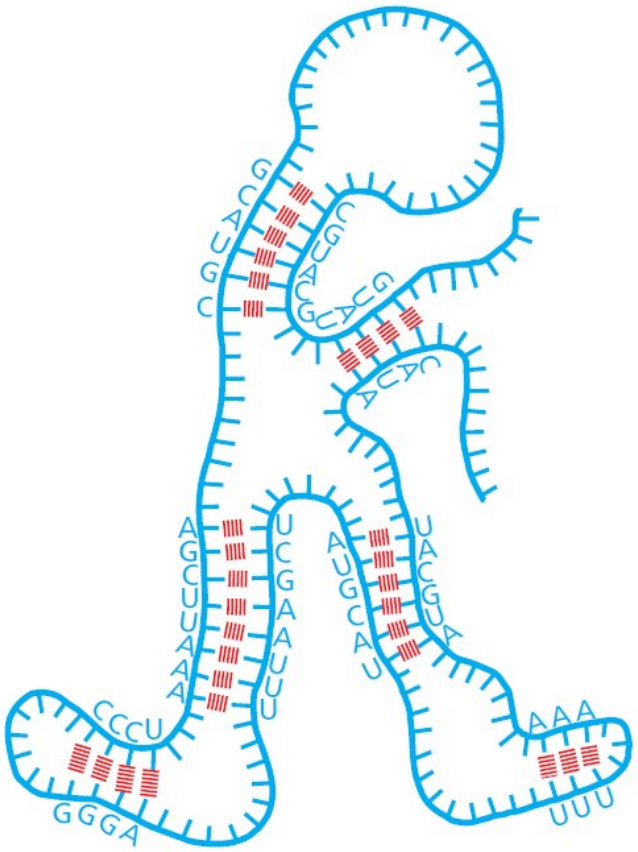


Pyridimines

Note: No CH3 in Uracil as in Thymine

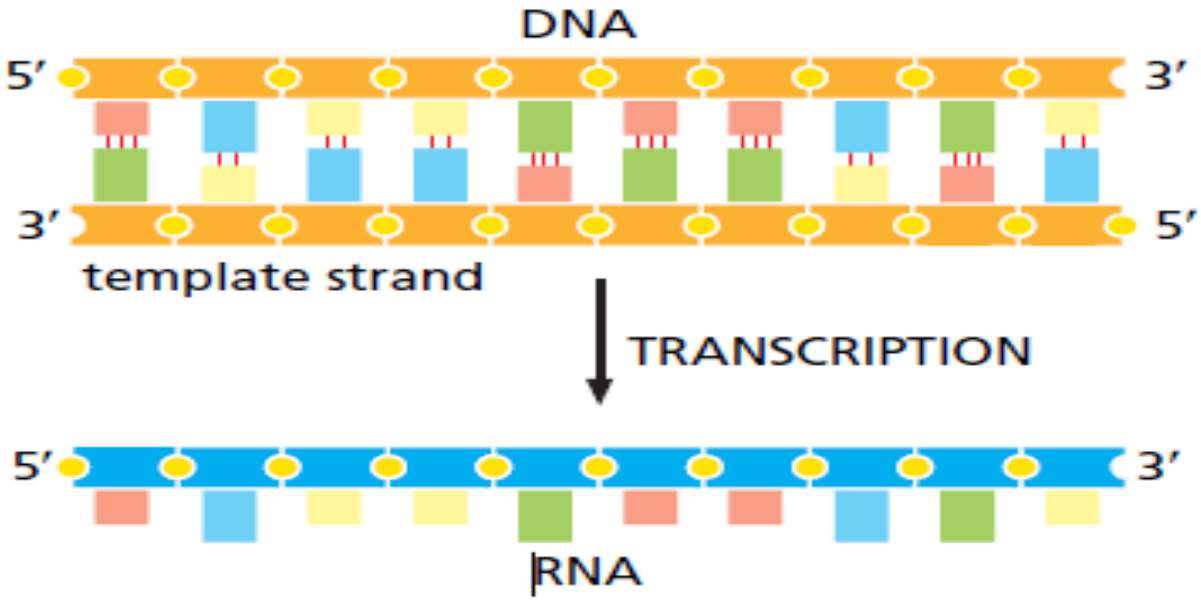
RNA

Nucleotide pairing between different regions of the RNA polymer chain causes the molecule to adopt a distinctive shape - enables it to recognize other molecules by selective binding, or, catalyze chemical changes in the molecules that are bound.



RNA Synthesis:

RNA is also read in the 5′ to 3′ orientation.



RNA molecules that are copied from the genes (which ultimately direct the synthesis of proteins) are called messenger RNA (mRNA) molecules.

RNA Synthesis

1. If the following DNA sequence is the forward strand:

5′ CATTGCCAGT 3′

What will be the sequence of the RNA strand synthesized?

2. If the following DNA sequence is used as template for RNA synthesis:

5′ CATTGCCAGT 3′

Give the sequence of the RNA strand read in 5′ to 3′ orientation.

RNA Synthesis

1. If the DNA sequence in the forward strand is given:

5′ CATTGCCAGT 3′

Template sequence used for RNA synthesis is its complement:

5′ CATTGCCAGT 3′

3′ GTAACGGTCA 5′ template

The synthesized RNA sequence is the reverse complement of the template:

3′ GTAACGGTCA 5′ template

5′ CAUUGCCAGU 3′ RNA

- i.e., synthesized RNA sequence is basically the DNA sequence in the forward strand with T replaced by U

RNA Synthesis

If the following DNA sequence is used as template for RNA synthesis:

5′ CATTGCCAGT 3′

First write its complement:

5′ CATTGCCAGT 3′

3′ GUAACGGUCA 5′ complement

Then the synthesized RNA sequence in 5′ to 3′ orientation is:

5′ ACUGGCAAUG 3′ RNA

- i.e., synthesized RNA sequence is basically the complement of the template DNA sequence with T replaced by U, when read in the 5′ to 3′ orientation

RNA Synthesis:

There are other RNA molecules also obtained from genes. The final product in such cases is RNA.

- these are known as noncoding RNAs because they do not code for protein.

e.g., in yeast *Saccharomyces cerevisiae*, over 1200 genes (~15%) produce RNA as their final product; Humans may produce on the order of 10,000 noncoding RNAs.

These RNAs, like proteins, serve as enzymatic, structural, and regulatory components for a wide variety of processes in the cell.

TABLE 6–1 Principal Types of RNAs Produced in Cells	
Type of RNA	Function
mRNAs	Messenger RNAs, code for proteins
rRNAs	Ribosomal RNAs, form the basic structure of the ribosome and catalyze protein synthesis
tRNAs	Transfer RNAs, central to protein synthesis as adaptors between mRNA and amino acids
snRNAs	Small nuclear RNAs, function in a variety of nuclear processes, including the splicing of pre-mRNA
snoRNAs	Small nucleolar RNAs, help to process and chemically modify rRNAs
miRNAs	MicroRNAs, regulate gene expression by blocking translation of specific mRNAs and cause their degradation
siRNAs	Small interfering RNAs, turn off gene expression by directing the degradation of selective mRNAs and the establishment of compact chromatin structures
piRNAs	Piwi-interacting RNAs, bind to piwi proteins and protect the germ line from transposable elements
lncRNAs	Long noncoding RNAs, many of which serve as scaffolds; they regulate diverse cell processes, including X-chromosome inactivation

Note: rRNA, tRNA and snRNA play an important role in protein synthesis

Protein Synthesis

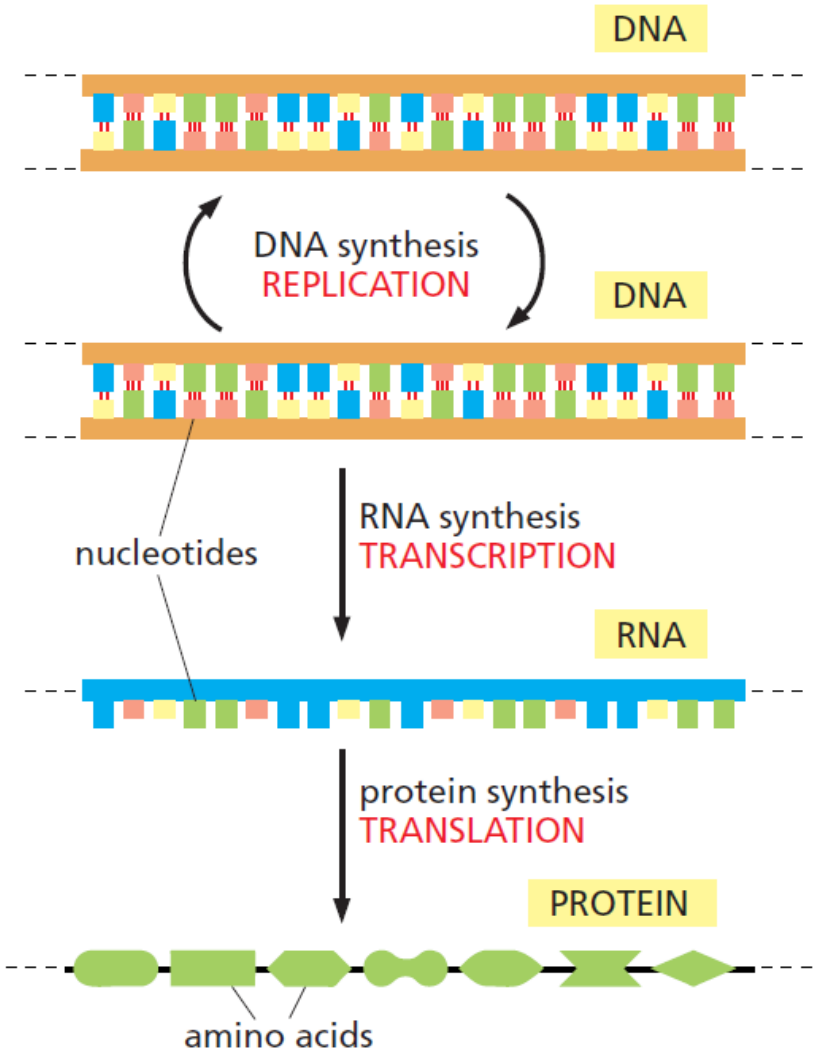
Proteins are synthesized from DNA in a two-step process:

Each chromosome has several genes that code for various traits in the body.

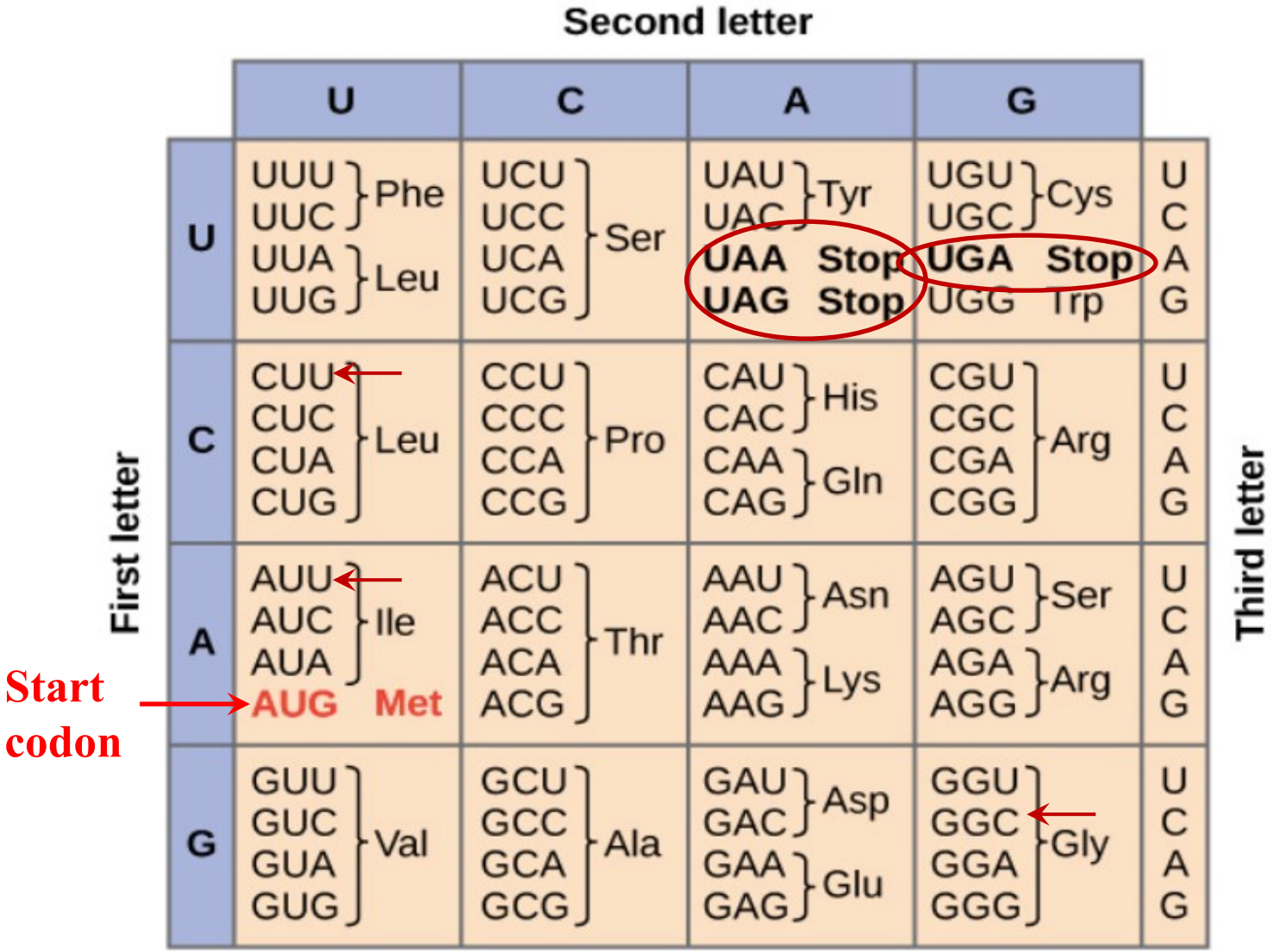
- from enzymes to the color of eye

RNA molecules direct synthesis of proteins in a complex process called translation.

- information in mRNA is read out in groups of three nucleotides, called codons.



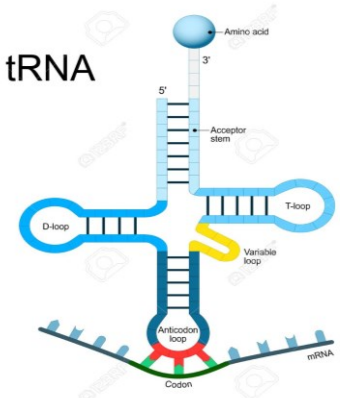
The Genetic Code



The genetic code is degenerate

Protein Synthesis

Using the genetic code, the amino acid sequence synthesized from the following mRNA sequence is:

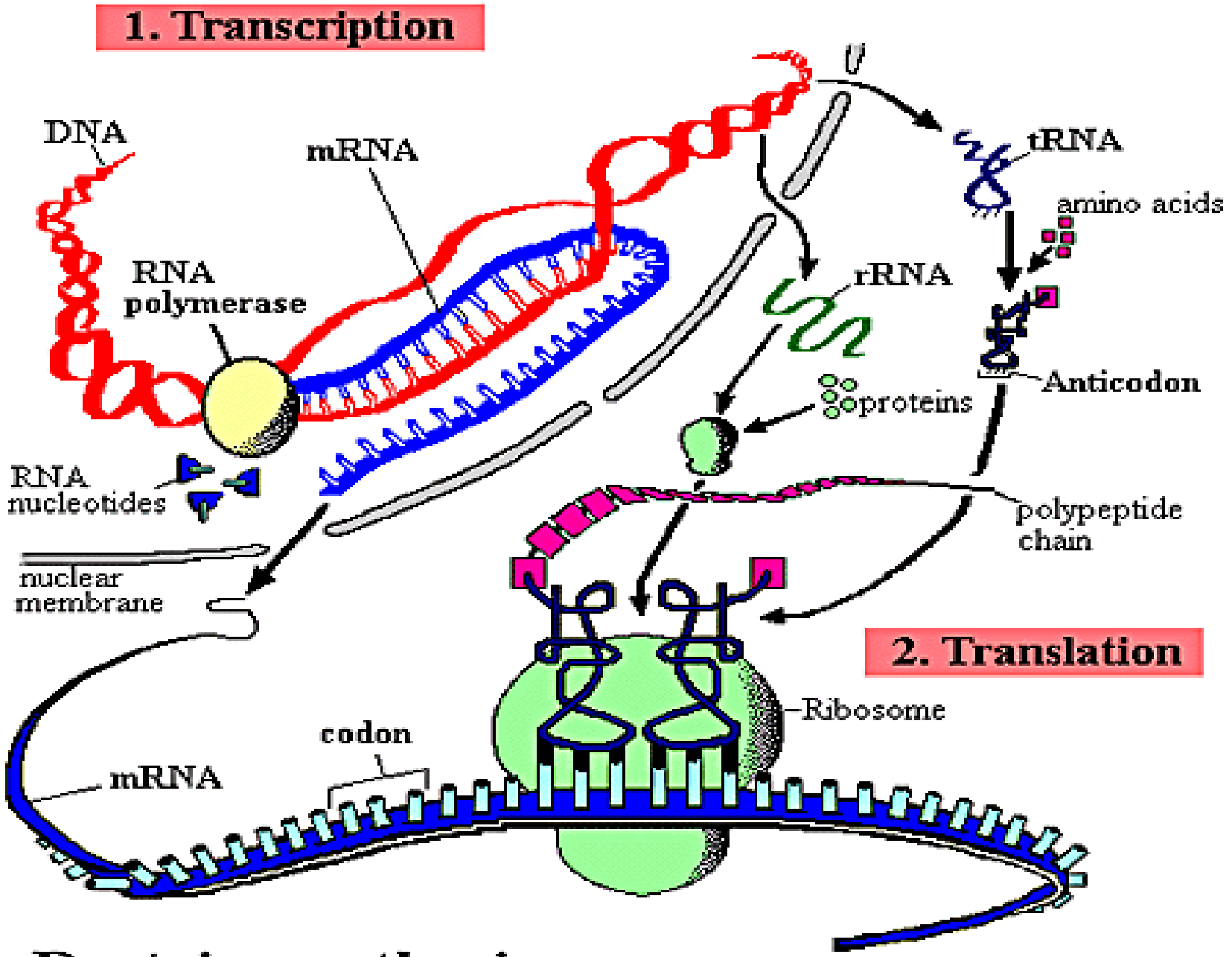


5′ ACU GGC AAU 3′

Thr Gly Asn

This genetic code is read out by a class of small RNA molecules, called transfer RNAs (tRNAs).

- **each type of tRNA attaches at one end a specific amino acid and at its other end has a specific sequence of 3 nucleotides, an anticodon that enables it to recognize, through base-pairing, a particular codon in the mRNA sequence.**
- **This process occurs on ribosome, a large multi-molecular machine composed of both proteins and ribosomal RNA.**



Protein synthesis

Proteins

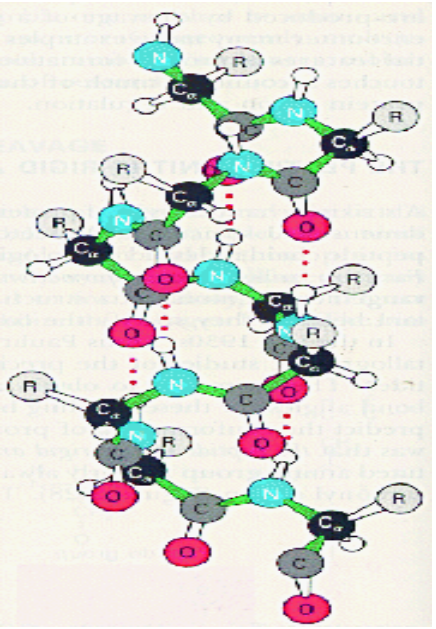
Like DNA and RNA, Proteins carry information in linear sequence on a 20-letter alphabet, called amino acids:

ATRVGTCWPRA

Protein structure is divided in 4 hierarchical levels:

- **Primary structure - represented by AA sequences**
- **Secondary structure - α-helices & β-sheets**
- **Tertiary and Quaternary structures - represented by 3D structures**

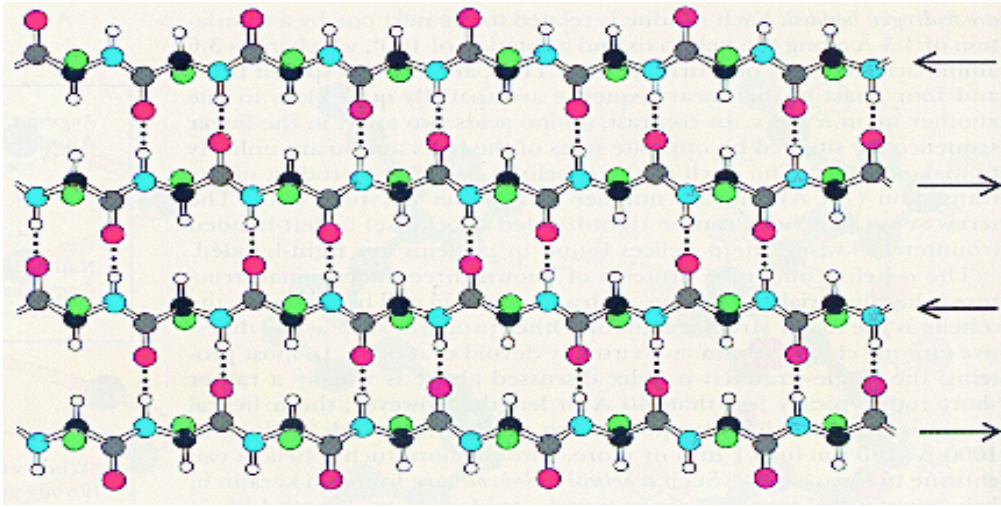
Primary Structure: ATRVGTCWPRA



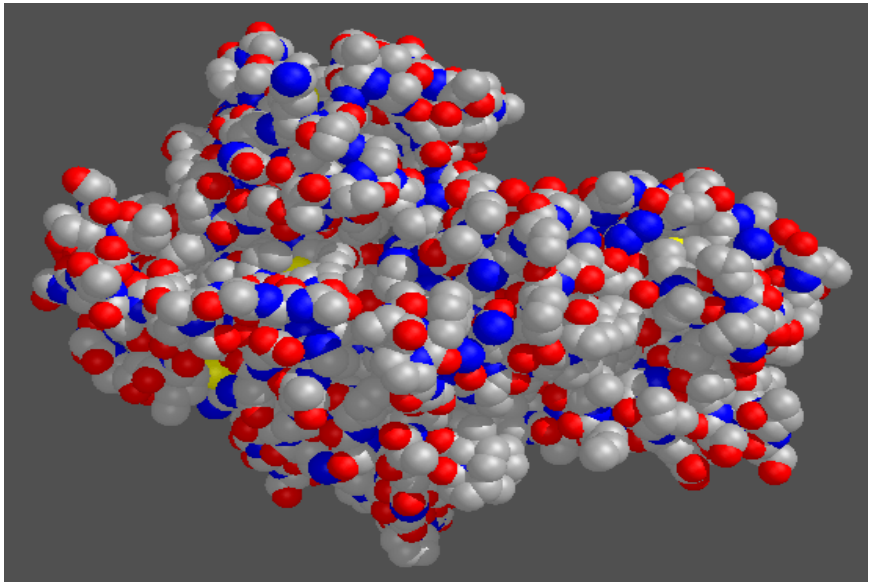
α-helix

Secondary Structures

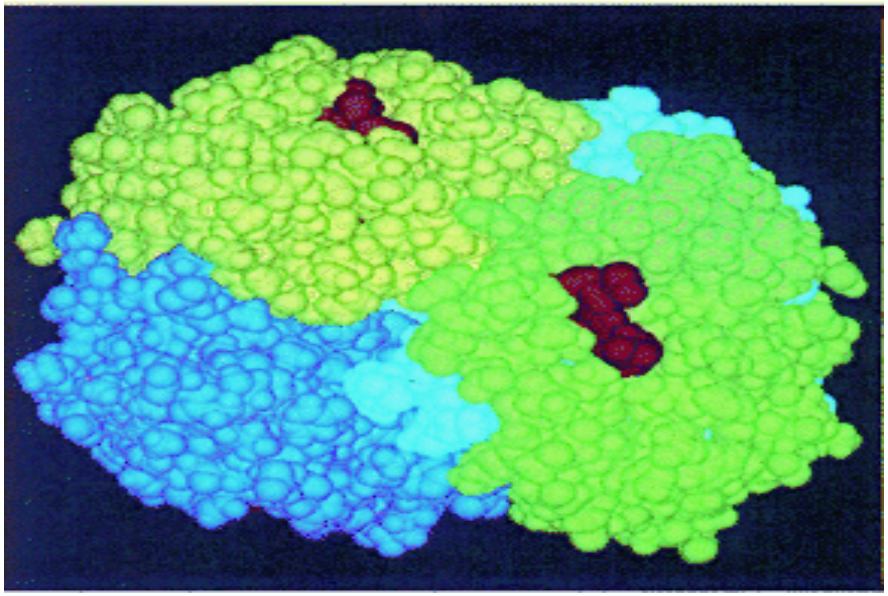
β-sheets



Tertiary Structure



Quaternary Structure



Function of Proteins

- **Proteins make up much of the cellular structure – hair, skin, fingernails, etc.**
- **Enzymes – catalyze chemical reactions within the cell**
- **Transcription factors – regulate the manner in which genes direct production of other proteins**
- **Receptors – proteins on the surface of cells act as receptors for hormones and other signaling molecules**
- **Recognize and bind to Nucleic acids (DNA, RNA) and Proteins – to carry out their functions in the cell**

Genes

Special sequences in the DNA code for genes:

- **Protein-coding genes, for which the final product is a protein.**
  - **same gene may give rise to more than one protein (~ 6 per gene in humans).**
- **Non-coding RNA genes - for which the final product is RNA**

Genotype – An organism's genotype is the set of genes that it carries.

Phenotype – An organism's phenotype is all of its observable characteristics which are influenced both by its genotype and by the environment e.g., height, hair colour, levels of hormones, etc.

Differences in the genotypes can produce different phenotypes

Genes for ear form are different, causing one of the cats to have normal ears and the other to have curled ears

Change in environment can also affect the phenotype. Pinkness is <u>not</u> encoded in the genotype of flamingos - the food they eat makes their phenotype white or pink - a natural pink dye, canthaxanthin, obtained from their diet of brine shrimp and blue-green algae

Genes

The biological function of a gene is to preserve and express the genetic information encoded within it

Genes are normally very stable entities

Genetic stability is not absolute, however.

Genes may occasionally become altered; these changes called mutations create new alleles.

Mutant genes are also stable entities and are inherited in the same way as normal, wild-type genes.

Genes

Normal diploid cells such as somatic cells of humans contain two sets of genes – one set inherited from each parent.

- corresponding genes derived from each parent are called alleles.

Together the two alleles govern the phenotype of an organism.

What is the percentage of genes in a genome?

Genes

Gene-fraction varies from ~70% in prokaryotes to ~ 2 - 3% in humans

- **does this imply prokaryotes have more gene content than eukaryotes?**

- **Size of a prokaryotic genome? Eukaryotic genome?**

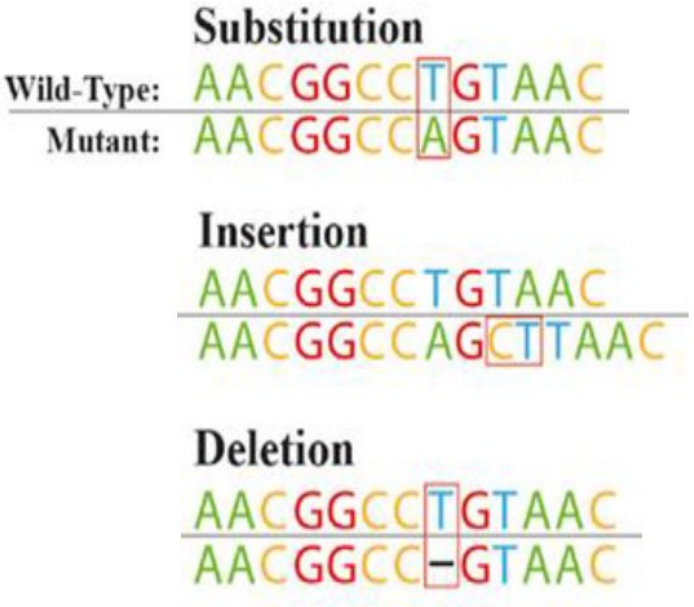
What's the function of remaining ~97-98% of human genome?

The remaining part of the genome consists of noncoding regions, whose functions may include providing chromosomal structural integrity and regulating where, when, and in what quantity proteins are made, and repeats.

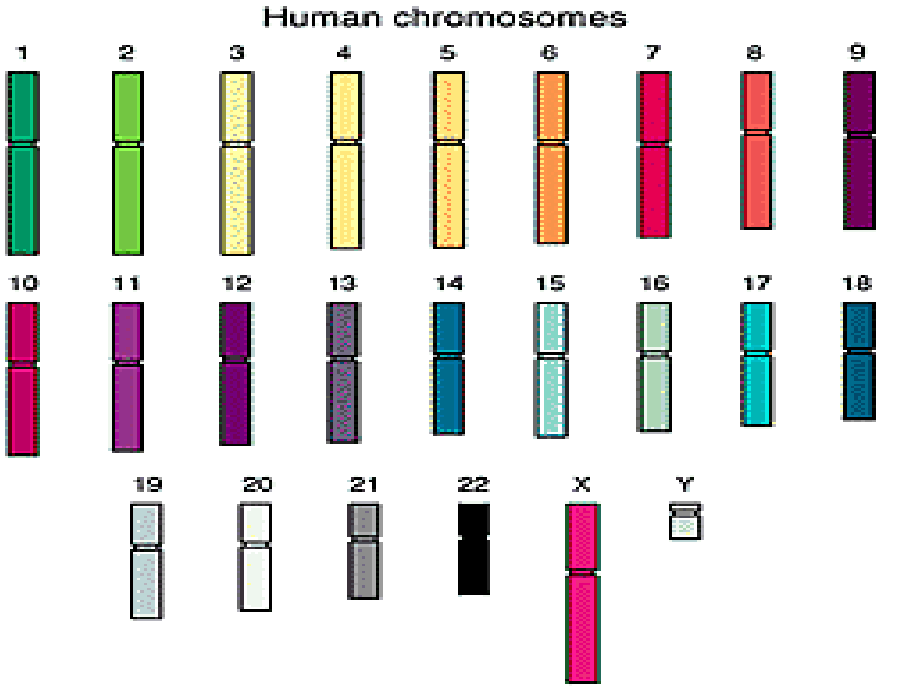
Mutations

Mutations - are local changes in the DNA content, caused by inexact replication and are of various kinds:

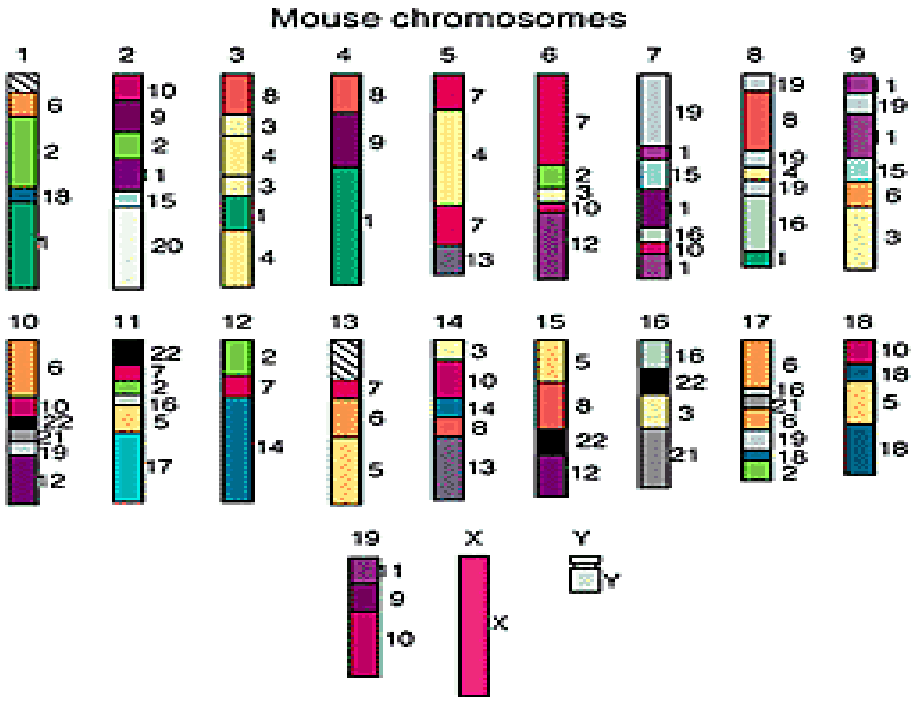
- **Substitution - a base is replaced by another - may or may not alter the protein sequence depending on the place it occurs.**
- **Insertion/Deletion – addition/removal of one or more bases – results in a frame-shift in coding regions.**
- **Rearrangement - a change in the order of complete segments along a chromosome.**



Chromosomal rearrangements occur both within and between chromosomes during evolution



The colors on the mouse chromosomes and the numbers alongside indicate the human chromosomes containing homologous segments.



Mutations

Role of Mutations:

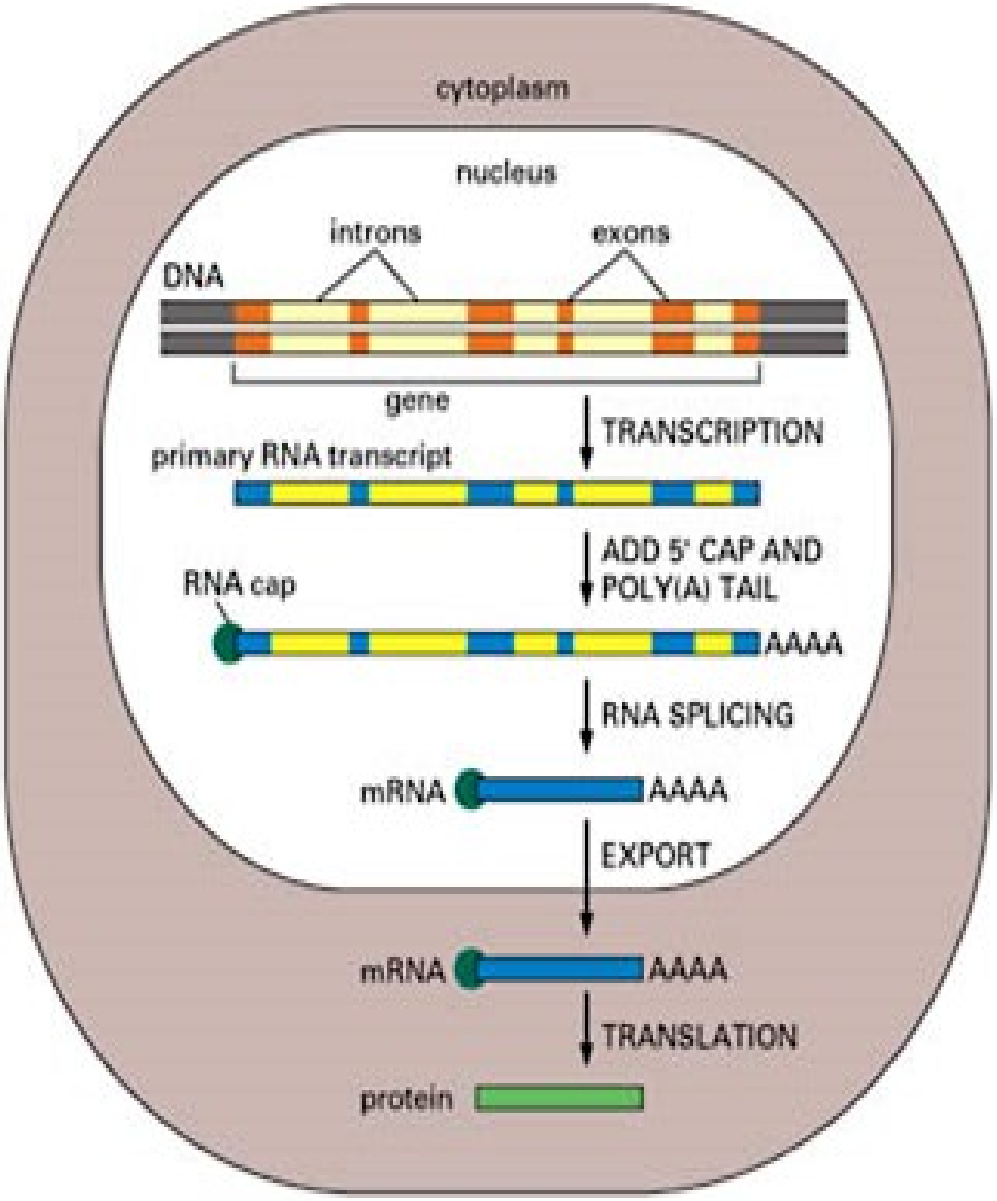
- **Mutations are the source of phenotypic variation on which natural selection acts, creating species & changing them.**

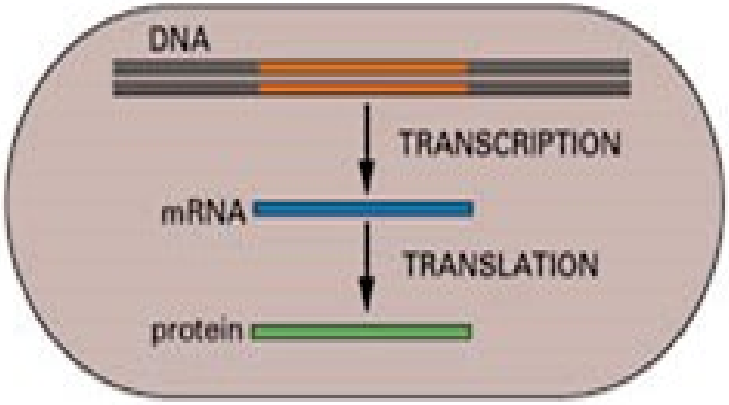
  e.g., the human and mouse genome are very similar – major difference being the internal order of DNA segments.

  Without mutations there wouldn't be any evolution!

- **They are responsible for inherited disorders and diseases, which involve alterations in gene.**

Steps Leading from Gene to Protein

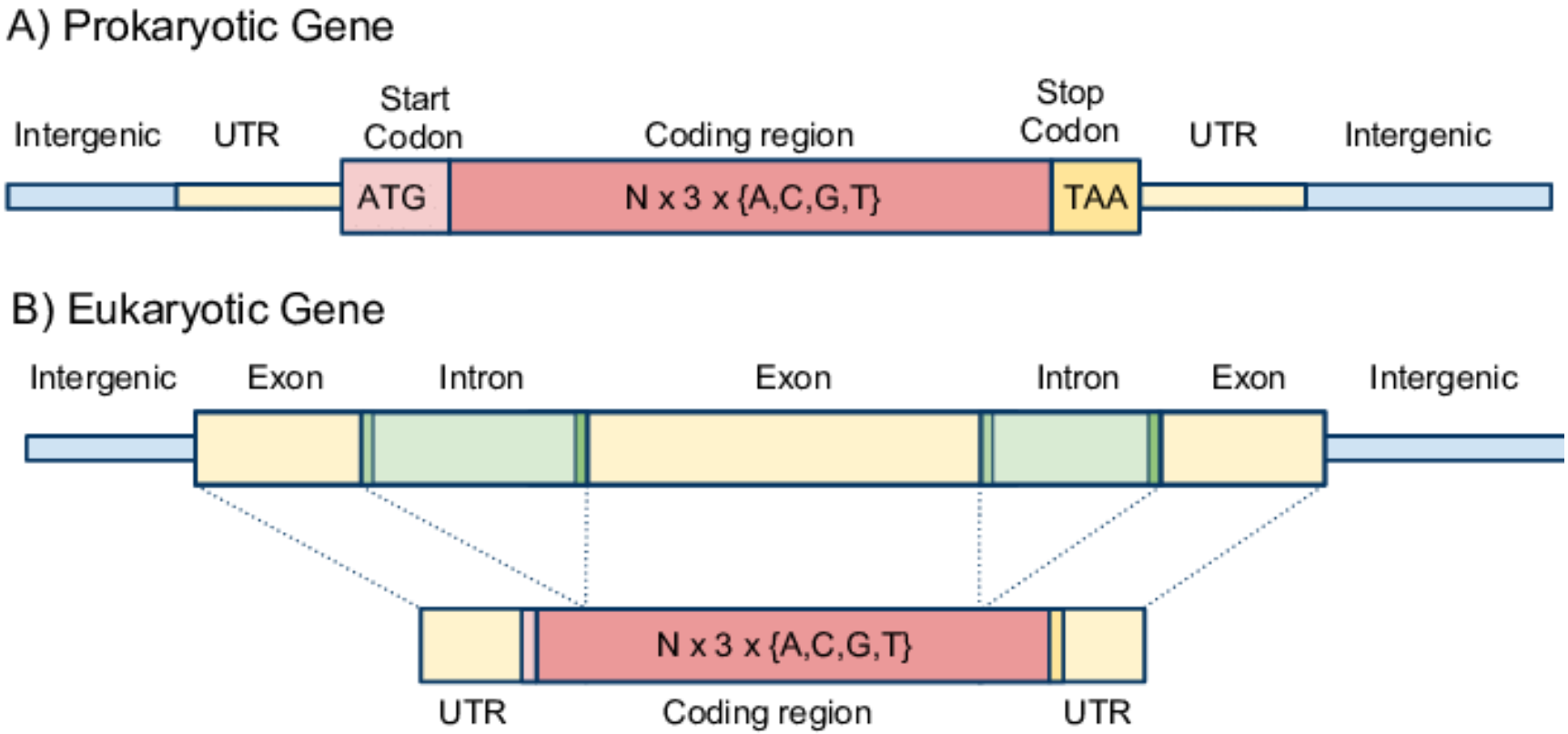




Prokaryotes

Eukaryotes

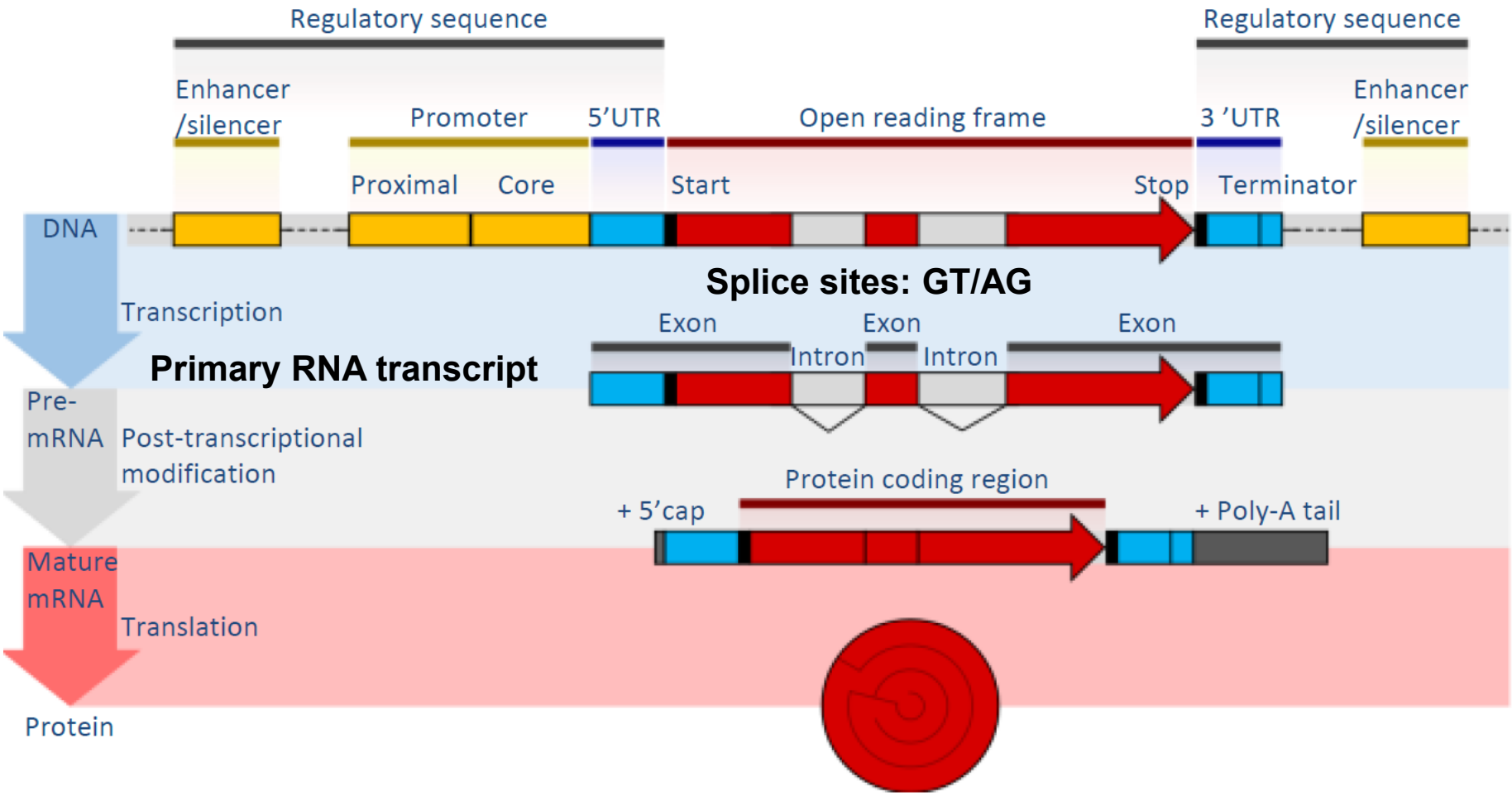
Gene Structure



UTRs – Untranslated Regions – are transcribed, but not translated

Eukaryote Gene Structure

Start codon: ATG **Stop Codon: TAA/TAG/TGA**



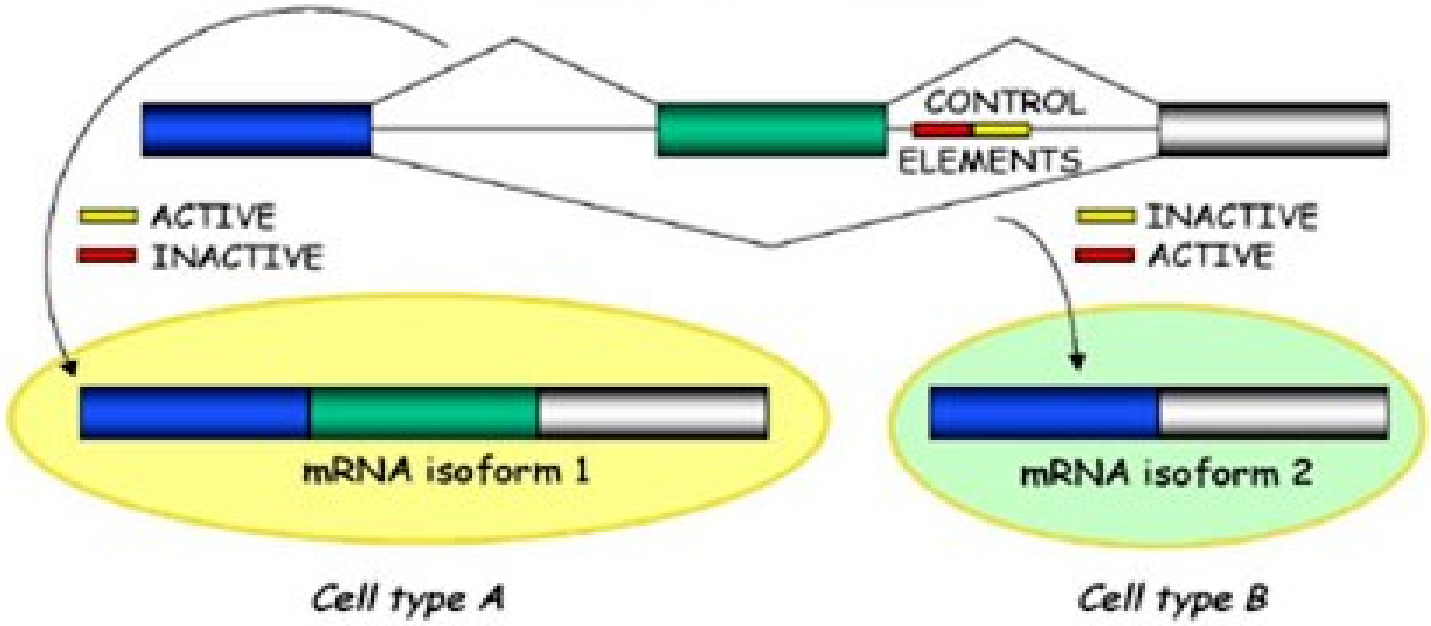
Transcription is initiated only at certain specific positions in the sequence, signaling the beginning of genes, called promoters.

Alternative Splicing

- **In many cases, the pattern of splicing can vary depending on the tissue in which the transcription occurs.**

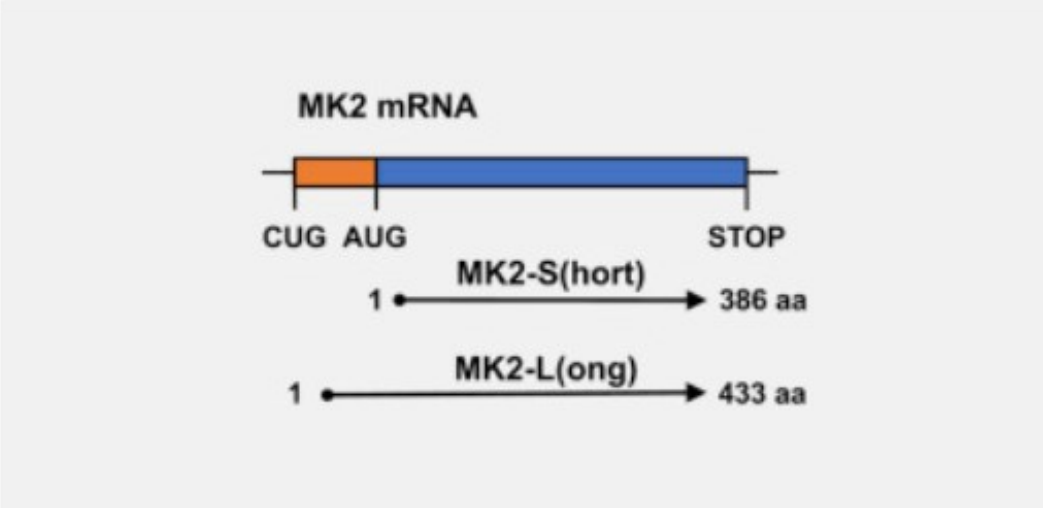
  e.g., an exon maybe spliced in the gene transcribed in liver but retained when transcribed in the brain.

- **This variation called alternative splicing, contributes to the overall protein diversity in the organism**



Alternative Initiation

- Another type of variation that contributes to protein diversity is alternative initiation



- Alternative translation is an important mechanism of post-transcriptional gene regulation leading to the expression of different protein isoforms originating from the same mRNA

Data Representation

DNA - a complex, dynamic, three-dimensional molecule represented as a string of alphabets

- a perfect representation for computer analysis

Aim: to find grammar & syntax rules of DNA language based on this 4-letter alphabet

- similar to English Grammar to form meaningful sentences

Biological Sequence Analysis

Pattern Recognition:

Assumption in biological sequence analysis:

- strings carrying information will be different from random strings

If a hidden pattern can be identified in a string, it must be carrying some functional information

Biological Sequence Analysis

Order of occurrence of bases:

not completely random

- Different regions of the genome exhibit different patterns of the four bases, A, T, G, C

e.g., protein coding regions, regulatory regions, intron/exon boundaries, repeat regions, etc.

Aim: Identify various patterns to infer their functional roles

Example

❄♒♓♦ ♓♦ ♋ ●♏♍♦◆❒♏ □■
✍♓□♓■♐□❒○♋♦♓♍♦

♋♦♐♎ ●&♐♐● ♐♎♐♎ ♦♐♐♦⛝♏
■❖♍❒□♦ ■⌘♍♎♐♒♦□◆

This is a lecture on bioinformatics

asjd lkjfl jdjd sjftye nvcrow nzcdjhspu

Frequency of letters

A.	7.3%	N.	7.8%
B.	0.9%	O.	7.4%
C.	3.0%	P.	2.7%
D.	4.4%	Q.	0.3%
E.	13.0%	R.	7.7%
F.	2.8%	S.	6.3%
G.	1.6%	T.	9.3%
H.	3.5%	U.	2.7%
I.	7.4%	V.	1.3%
J.	0.2%	W.	1.6%
K.	0.3%	X.	0.5%
L.	3.5%	Y.	1.9%
M.	2.5%	Z.	0.1%

Other statistics

Frequencies of the most common first letter of a word, last letter of a word, doublets, triplets, etc.

20 most used words in written English

- **the of to in and a for was is that on at he with by be it an as his**

20 most used words in spoken English

- **the and I to of a you that in it is yes was this but on well he have for**

Parallels in DNA language

ATGGTGGTCATGGCGCCCCGAACCCTCTTCCTGCTG
CTCTCGGGGGCCCTGACCCTGACCGAGACCTGGGCG
GGTGAGTGCGGGGTCAGGAGGGAAACAGCCCCTGC
GCGGAGGAGGGAGGGGCCGGCCCGGCGGG

GTCTCAACCCCTCCTCGCCCCCAGGCTCCCACTCCA
TGAGGTATTTCAGCGCCGCCGTGTCCCGGCCCGGCC
GCGGGGAGCCCCGCTTCATCGCCATGGGCTACGTGG
ACGACACGCAGTTCGTGCGGTTC

Parallels in DNA language

ATG GTG GTC ATG GCG CCC CGA ACC CTC TTC
CTG CTG CTC TCG GGG GCC CTG ACC CTG ACC
GAG ACC TGG GCG GGT GAG TGC GGG GTC AGG
AGG GAA ACA GCC CCT GCG CGG AGG AGG GAG
GGG CCG GCC CGG CGG…

GTC TCA ACC CCT CCT CGC CCC CAG GCT CCC ACT
CCA TGA GGT ATT TCA GCG CCG CCG TGT CCC
GGC CCG GCC GCG GGG AGC CCC GCT TCA TCG
CCA TGG GCT ACG TGG ACG ACA CGC AGT TCG
TGC GGT TC…

1st exon and 1st intron of Human HLA gene

This task needs to be automated because of the large genome sizes:

Smallest genome:

Mycoplasma genitalium 0.5×10^6 bp

Human genome: 3×10^9 bp – not the largest!

~ 10-100 times the Britannica Encyclopedia

Plant genomes are even larger.

DNA Sequence Analysis

- **Evolution has operated on every sequence that we see today - genes and sequences involved in gene regulation are conserved.**
- **these are transferred, like code modules, from one organism to another. Because of evolution, similar sequences have similar functions.**
- **Algorithms for comparing sequences and finding similar regions are at the heart of computational biology.**

Syllabus

Unit 1: Overview – Bioinformatics, Gene & Genome structure

Gene Technology – Restriction Endonucleases, Cloning vectors

DNA sequencing – PCR, cDNA and Whole Genome sequencing, NGS and third generation sequencing technologies

Unit 2: BioDatabases

- **Major Bioinformatics Resources – NCBI, EBI, PubMed,**
- **Primary Nucleotide and Proteins Databases - GenBank, UniProt, PDB,**
- **Genome Browsers – Ensembl, UCSC**
- **k-mer analysis and their significance in biological sequences**

Syllabus

Unit 3: Sequence Alignment:

- **Pairwise Alignment – Types of pairwise alignments – Global, Local and Overlap alignments, Dot Plots, dynamic programming (DP) algorithm,**
- **Scoring matrices for nucleotides and proteins and gap penalties,**
- **Sequence-based Database Search algorithms – BLAST, FASTA,**
- **Multiple Alignment, Algorithms for Global and Local MSA – DP, Progressive based (ClustalX), Iterative methods, Motif search-based methods**

Syllabus

Unit 4: Modeling Molecular Evolution – Phylogeny:

- **Markov models of base substitution,**
- **Computing Phylogenetic Distances,**
- **Phylogenetic Tree Construction Methods, PHYLIP**

Unit 5: Gene Prediction:

Gene Prediction approaches –

- **Open Reading Frames,**
- **Homology search,**
- **Content-based methods,**
- **Markov models**

Gene Technology

For all computational purposes, DNA is represented as a string of 4-letter alphabets - A, T, C, G:

attgctacgttacatcgctgca

How do we get this string representation from a dynamic double-stranded molecule?

DNA Sequencing - determine the precise sequence of nucleotides in a sample of DNA

To carry out this task we need to be able to chop the DNA, store it, make copies of it.

Let's consider the example of detecting if a person is infected by the novel coronavirus SARS-CoV-2

- **uses Real Time RT-PCR Nucleic Acid Detection Kit based on the PCR method which uses a fluorescent probe and a specific primer to detect three specific regions within the SARS-CoV-2 nucleocapsid protein N gene.**

- **How is the SARS-CoV-2 genome sequenced?**
- **How does one identify the coordinates of N gene on it? i.e., how to construct a physical map of a genome?**
- **How does one select which regions in this gene would give specificity for the presence of SARS-CoV-2?***
- **How are the specific probe regions extracted and amplified for detection?**
- **Is it possible to store the DNA sample for re-testing? How?**

To sequence a gene, we need to

- **Identifying the region of interest**
- **Isolate it from the organism – DNA fragmentation**
- **moving it to another easily manageable organism such as a bacterium for obtaining multiple copies – cloning**

Such manipulations are conducted by a toolkit of enzymes:

Restriction endonucleases - used as molecular scissors

DNA ligase - to bond pieces of DNA together

- a variety of additional enzymes that modify DNA are used to facilitate the process.

Restriction endonucleases are enzymes that make **site-specific** cuts in the DNA – **chemical scissors**

Ability to cut DNA into discrete fragments allows to understand

- how genetic material of an organism is **organized**
- how expression of genetic information is **controlled**
- how **alteration** of genetic information can give rise to genetically inherited disorders, etc.
- in **bulk production** of pharmaceutically important proteins

First restriction enzyme was isolated from H. influenzae in 1970 by Daniel Nathans and Kathleen Danna

- awarded the Nobel Prize for Medicine in 1978

Restriction endonucleases are enzymes that make site-specific cuts in the DNA – chemical scissors

First restriction enzyme was isolated from H. influenzae and used to cleave SV40 DNA (a tumor virus):

- 11 distinct DNA bands were visible on polyacrylamide gel electrophoresis, indicating that the enzyme always cut SV40 resulting in the same 11 pieces

Background

How were these restriction endonucleases identified?

Bacteria are under constant attack by bacteriophages – a virus that infect and replicates within a bacterium

To protect themselves, bacteria have developed a method to chop up any foreign DNA - such as that of an attacking phage

These bacteria build an endonuclease - an enzyme that cuts DNA - it circulates in the bacterial cytoplasm, waiting for phage DNA.

These endonucleases are termed "restriction enzymes" because they restrict the infection of bacteriophages.

Why the restriction enzymes do not chew up the genomic DNA of their host?

Background

A bacterium that makes a particular restriction endonuclease, also synthesizes a companion DNA methyltransferase,

- which methylates the DNA target sequence for that restriction enzyme, thereby protecting it from cleavage.

DNA from an attacking bacteriophage will not have these protective methyl groups and will be destroyed.

Methyl groups block the binding of restriction enzymes, but do not block the normal reading and replication of the genomic information stored in the host DNA.

DNA Fragmentation

Different endonucleases present in different bacteria recognize different nucleotide sequences

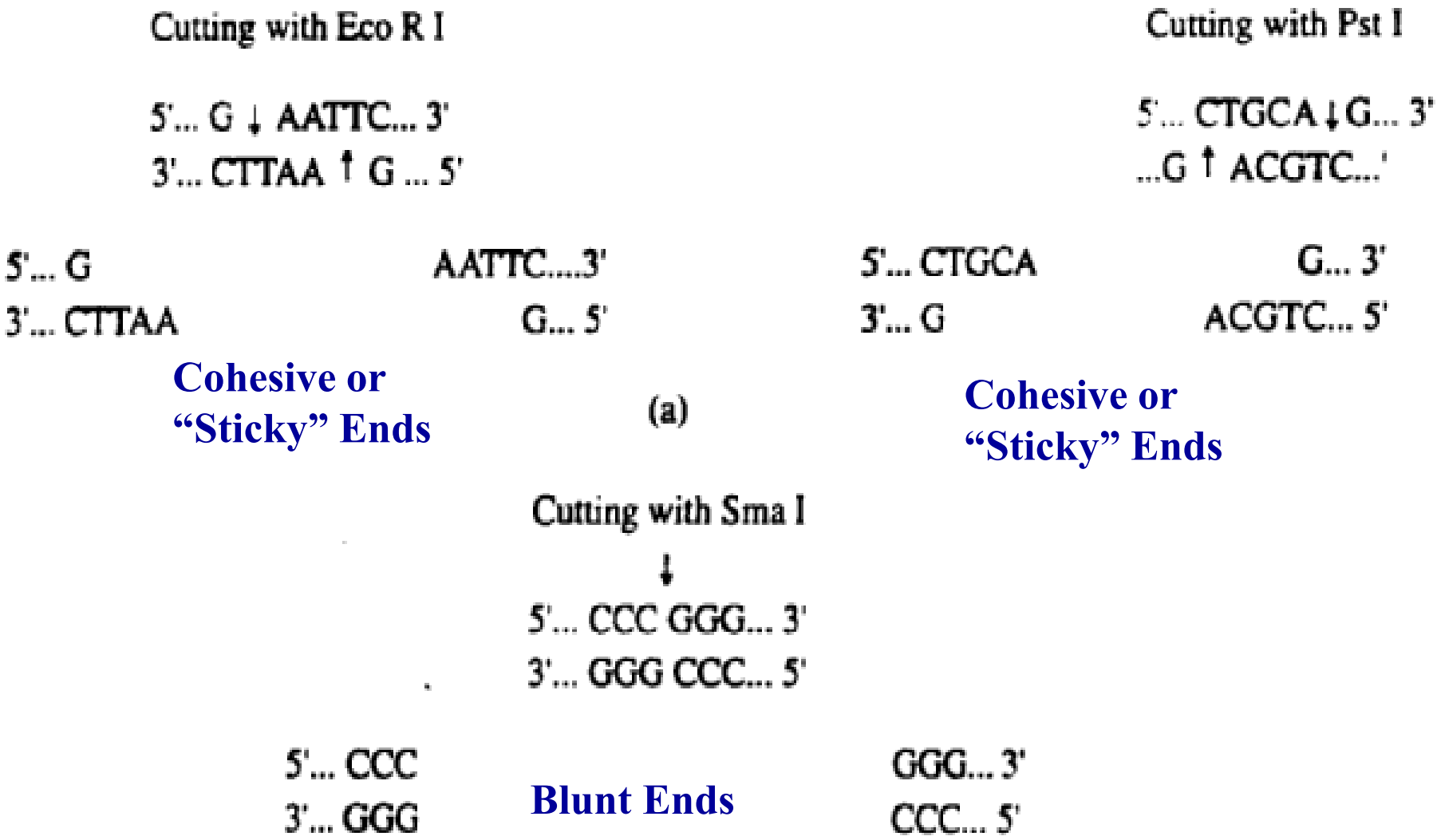
Naming of restriction enzymes - after their host of origin, e.g.,

- **EcoRI -** ***Escherichia coli***
- **Hind II & Hind III -** ***Haemophilus influenzae***
- **XhoI -** ***Xanthomonas holcicola***

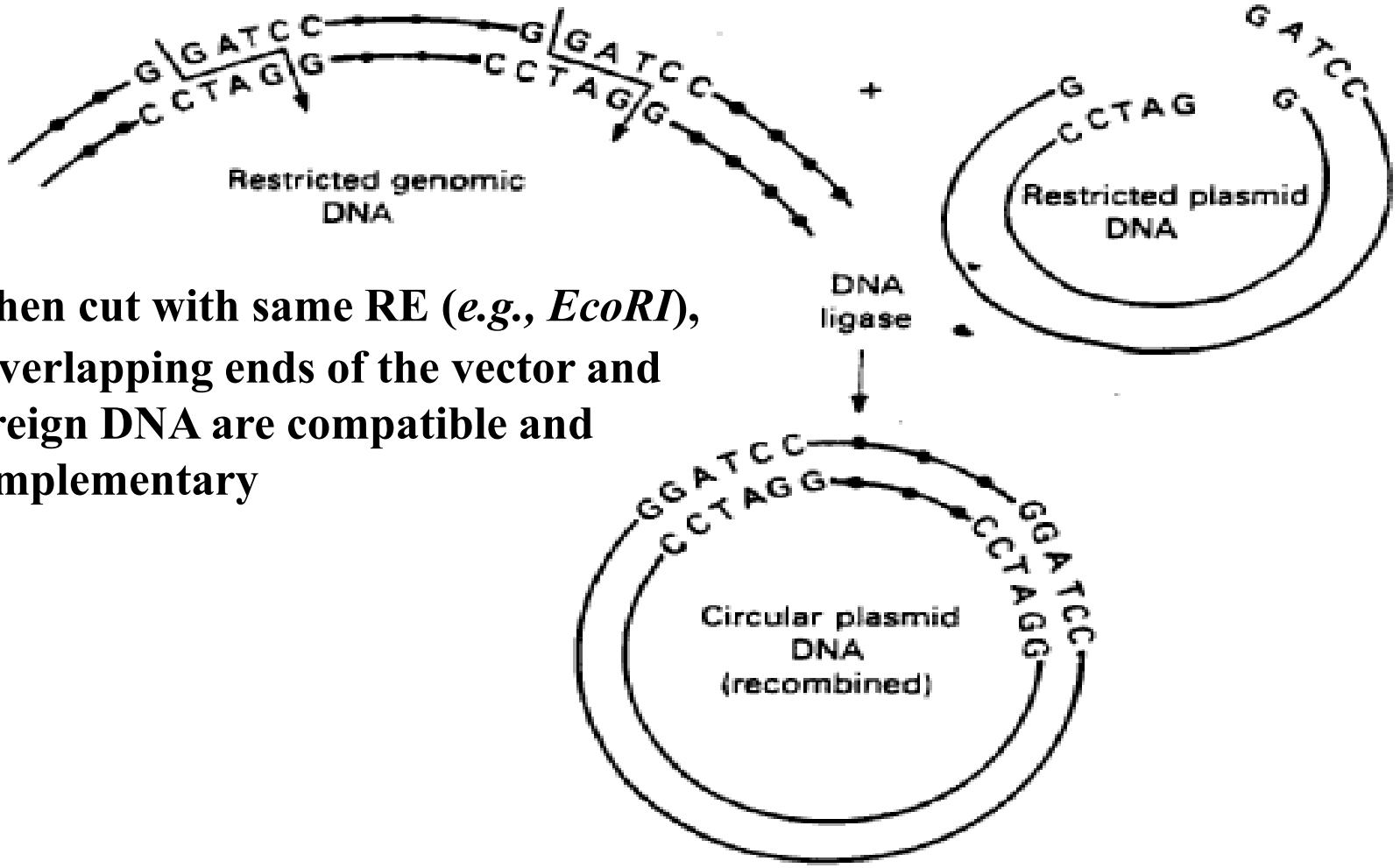
When cut with a restriction enzyme (RE), the ends of the cut DNA fragment can be cohesive or blunt-ended depending on the enzyme.

Enzyme	Recognition Sequence
EcoRI	G↓AATTC
HindIII	A↓AGCTT
BamHI	G↓GATCC
BglI	GCCNNNN↓NGGC
PvuI	CGATC↓G
HaeIII	GG↓CC
MboI	GAT↓C

Generation of Cohesive & Blunt-ended Fragments



Restriction enzyme digestion of genomic DNA and plasmid vector DNA



When cut with same RE (*e.g.*, *EcoRI*),

- overlapping ends of the vector and foreign DNA are compatible and complementary

How does one cut a DNA if it doesn't contain desired RE sites?

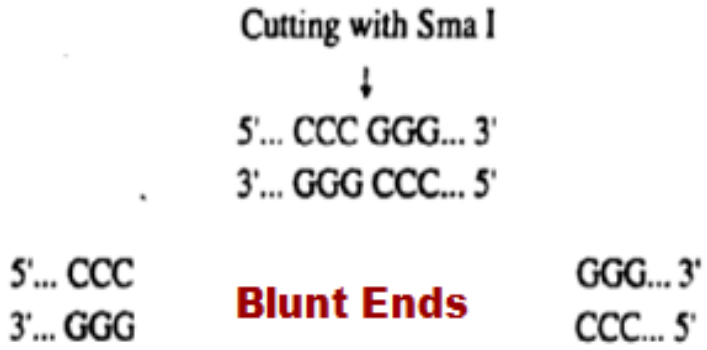
Or

If the RE site is present within the DNA of interest?

(say, within SARS-CoV-2 N gene)

Or

If the RE result in blunt-ended DNA fragments, how do we insert the fragment in a cloning vector?



How to clone a blunt-ended DNA fragment?

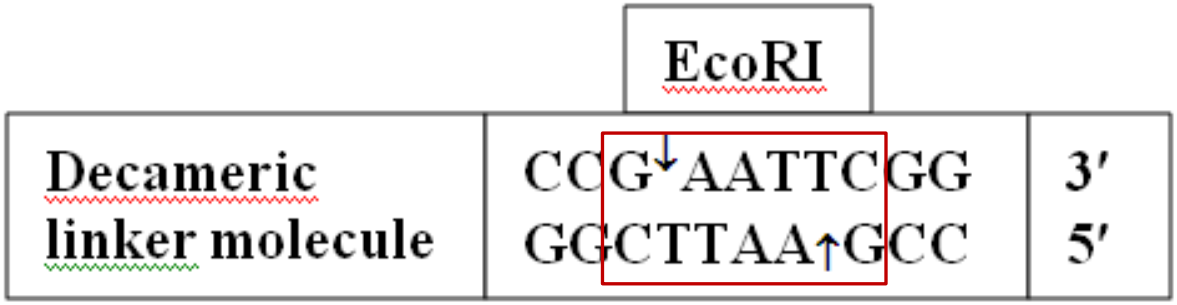
- **a linker molecule can be ligated on either side by DNA ligase, cut with the RE contained in the linker molecule to obtain cohesive ends.**

How does one cut a DNA if it doesn't contain desired RE sites?

- **the DNA maybe be cut with whatever RE sites are available, and then linker or adaptor molecules maybe added to enable ligating it to the vector.**

Linkers & Adaptors

Linkers - short, double-stranded DNA molecules (~ 8-14bp) with one internal site for RE (~ 3-8bp)



Problem?

- **the sites for the enzyme used to generate cohesive ends may be present in the target DNA fragment, limiting its use for cloning.**
- **This problem can be solved using adaptors.**

Linkers & Adaptors

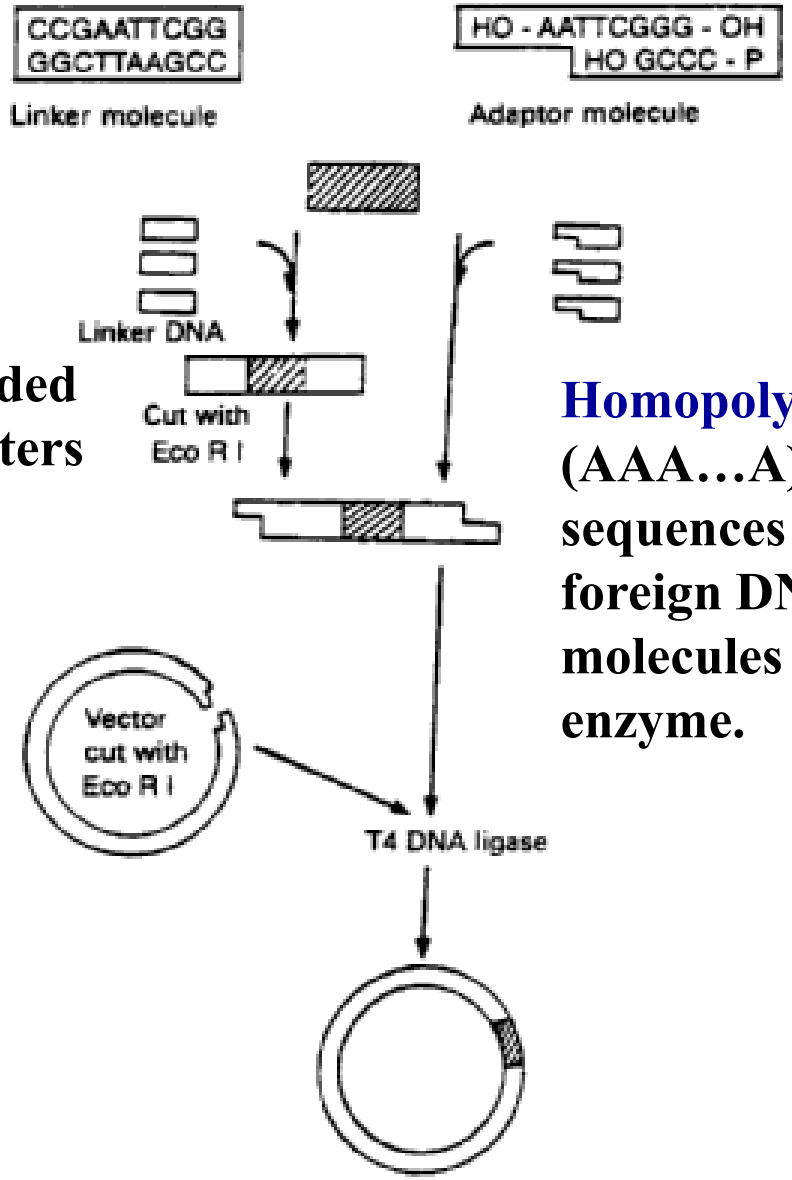
Adaptors - chemically synthesized DNA molecules with pre-formed cohesive ends

- it has one blunt end bearing a 5′ phosphate group and another cohesive end for a specific RE which is not phosphorylated to prevent self ligation.

Adaptor molecule

- reduces the need for restriction digestion following ligation

Use of Linker & Adaptor Molecules in the Formation of Recombinant Plasmids



This step avoided on using adapters

Homopolymer method - oligo A (AAA…A) or oligo T (TTTT…T) sequences added to the ends of foreign DNA & vector DNA molecules by terminal transferase enzyme.

Features of Restriction Enzymes

- **Length of recognition sequence dictates how frequently the enzyme will cut a DNA sequence**

  Which of the recognition sites - of length, 4, 6, or 8, will occur at higher frequency? At what distances will they occur?

- **Different REs can have the same recognition site and are called isoschizomers, e.g., *SacI* & *SstI* : GAGCTC**

- **Restriction recognitions sites can be unambiguous, e.g., *BamH* I recognizes the sequence GGATCC and no other, or ambiguous, e.g., *Hinf* I has a recognition site, GANTC.**

  Recognition sites for Hinf I will occur at what frequency?

Features of Restriction Enzymes

- **Recognition site for one enzyme may contain the restriction site for another, e.g., *BamH* I recognition site (GGATCC) contains the recognition site for *Sau3A* I (GATC).**

  ***Sau3A I* recognizes the sequence GATC and produces the same sticky ends as BamH I upon cutting**

  Will the two REs give the same results? If not, which one will give larger number of fragments?

- **Most recognition sequences are palindromes - they read the same forward and backward**

Can we use the property of palindrome sequence to identify restriction recognition sites?

Applications of Restriction Enzymes

Danna & Nathans showed that it was possible:

- **to prepare a physical map of the SV40 genome**
- **to localize the origin of replication**
- **to position early & late genes of SV40 onto this "restriction map"**
- **that any individual gene could be mapped by testing for biological activity during transformation experiments**
- **informative mutants could be made by deleting one or more of the specific fragments**

Applications of Restriction Enzymes

- **Variations in DNA sequences, *viz.*, mutations in recognition sites, copy number variation of VNTRs, insertions, deletions, inversions and translocations, can be identified by RE analysis**

  - The length variations is known as restriction fragment length polymorphisms (RFLPs).

- **In genetic engineering - using REs DNA may be cut at precise locations & using DNA ligase, reassembled in any desired order, allowing the researchers to assemble customized genomes; create designer bacteria that make insulin, or growth hormones, or add genes for disease resistance to agricultural plants, etc.**

- **in DNA sequencing – first step is to cut the DNA in manageable pieces**

Restriction Map

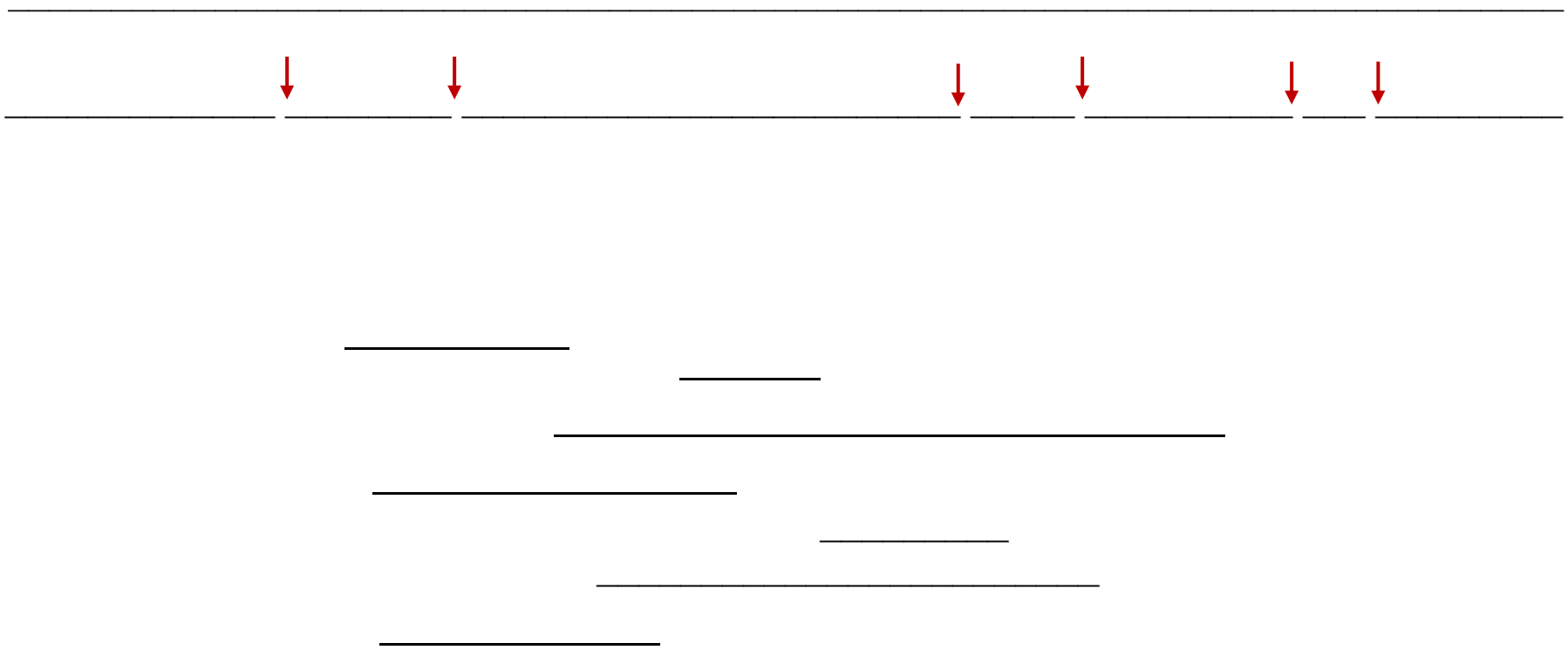
Restriction map is a description of restriction endonuclease cleavage sites within a piece of DNA

- **generating such a map is the first step in characterizing an unknown DNA**

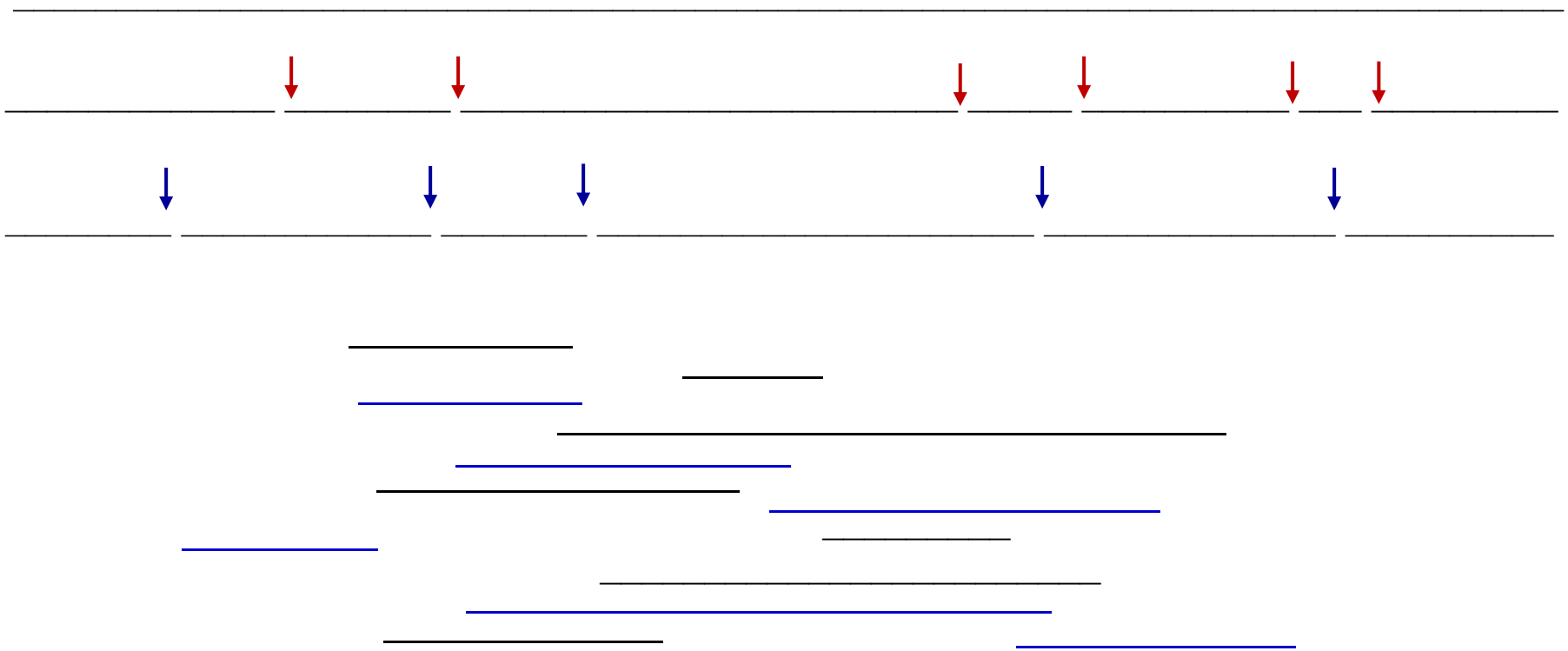
Multiple Complete Digest Mapping – creates a map by digesting DNA with multiple REs

- each recognizing a different specific short DNA sequence and producing a separate fingerprint for each clone

Because of the frequent occurrence of these sites, restriction mapping produces a relatively fine scale of physical map.



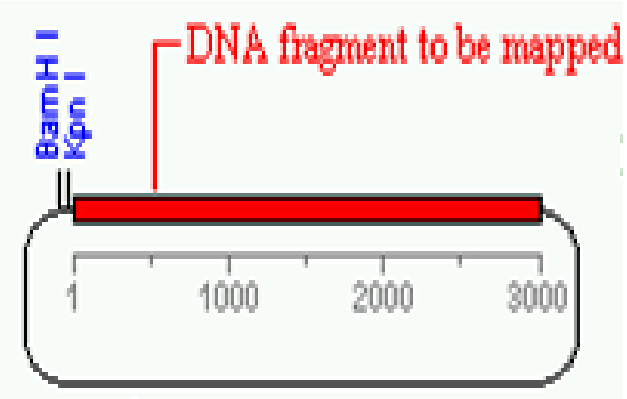
How do you order the fragments in the correct order?



The fragments can be arranged in the correct order by finding the overlapping fragments

Restriction Mapping

Ex: Consider a plasmid that contains a 3000 bp fragment of unknown DNA & unique recognition sites for enzymes Kpn I & BamH I.



Consider first separate digestions with Kpn I & BamH I:

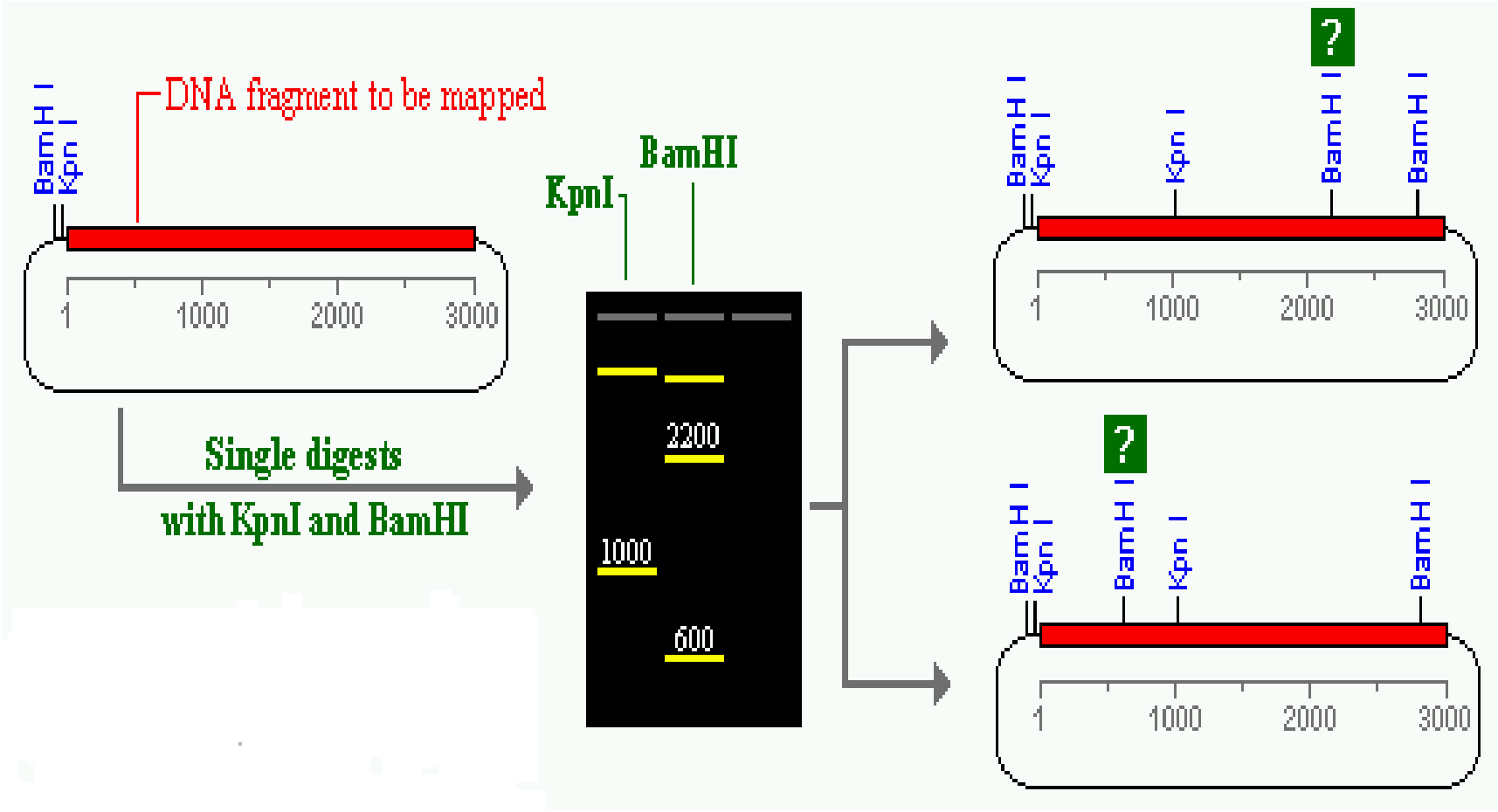
Kpn I yields 2 fragments: 1000bp & "big"

BamH I yields 3 fragments: 600, 2200 & "big"

big – part of unknown DNA sequence + vector

⇒ one Kpn I site & two BamH I sites are present in the unknown DNA sequence, given 1 each on the vector sequence

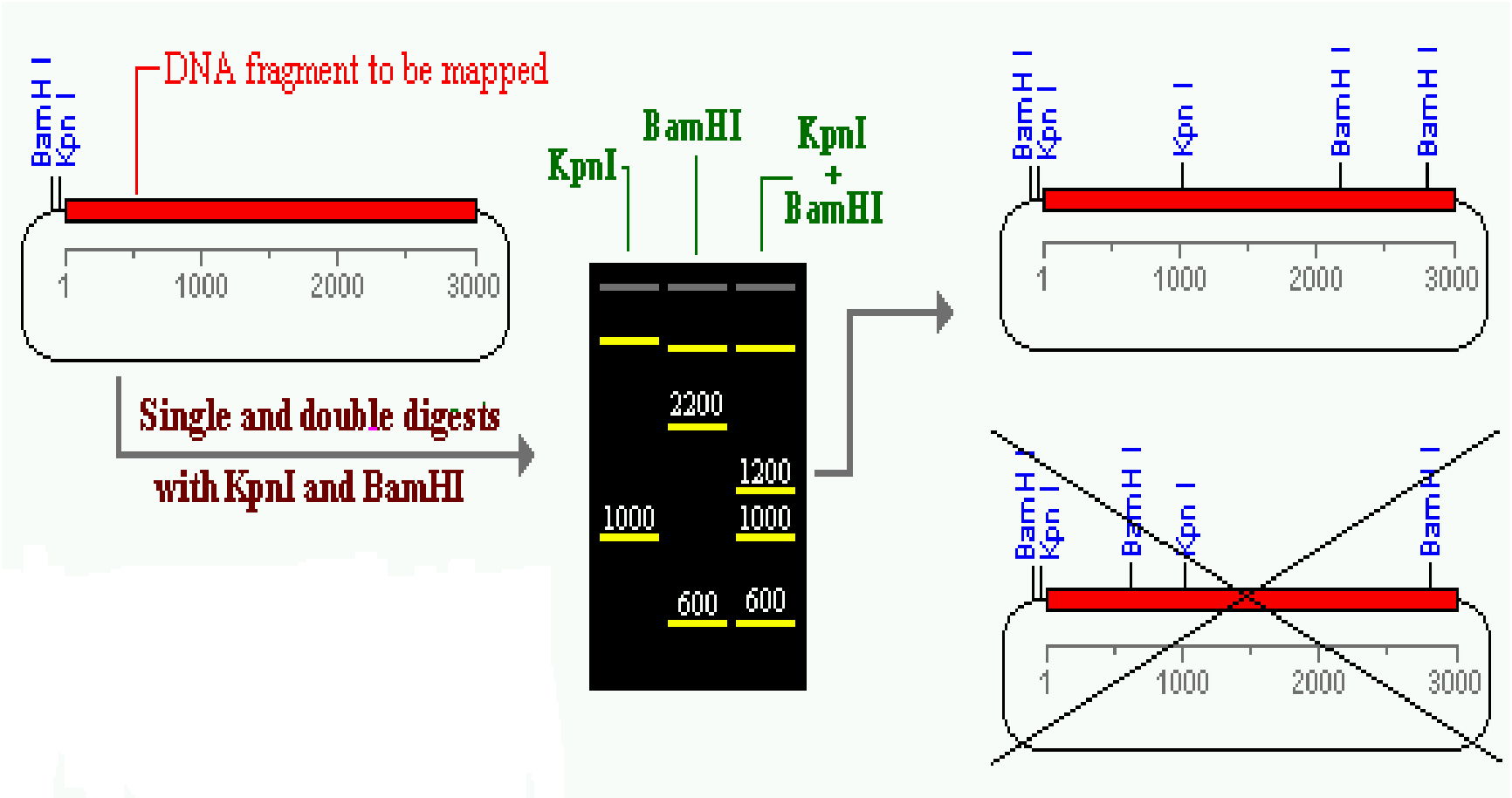
Restriction Mapping



One BamH I site is at 2800 bp. Trick to determine the location of 2nd BamH I site is to digest the plasmid with Kpn I & BamH I together

Restriction Mapping

Double digest yields fragments of 600, 1000 & 1200 bp (plus the "big" fragment).



Restriction Mapping

If the above process is conducted with a larger set of enzymes, a much more complete map would result

single digests - are used to determine which fragments are in the unknown DNA, and

multiple digests - to order and orient the fragments correctly.

For any novel genome, e.g., SARS-CoV-2, can a physical map be constructed computationally?

Restriction Mapping

Using a Computer to Generate Restriction Maps

If the sequence is known, feed it to computer programs, which will search the sequence for various RE recognition sites and build a map.

- **Mapper - available as part of Molecular Toolkit**

  http://arbl.cvmbs.colostate.edu/molkit/mapper/

- **Webcutter**

  http://www.firstmarket.com/cutter/cut2.html

- **RebSite – as part of the REBASE Tools**

  http://tools.neb.com/REBsites/index.php3

REBASE
The Restriction Enzyme dataBASE

A comprehensive database containing information:

- **restriction enzymes, methylases & related proteins involved in restriction-modification processes**
- **recognition and cleavage sites, isoschizomers, neoschizomers, commercial availability, methylation sensitivity, crystal & sequence data.**

All newly sequenced genomes are analyzed for the presence of putative restriction systems and these data included in REBASE

It is updated daily (http://rebase.neb.com/)

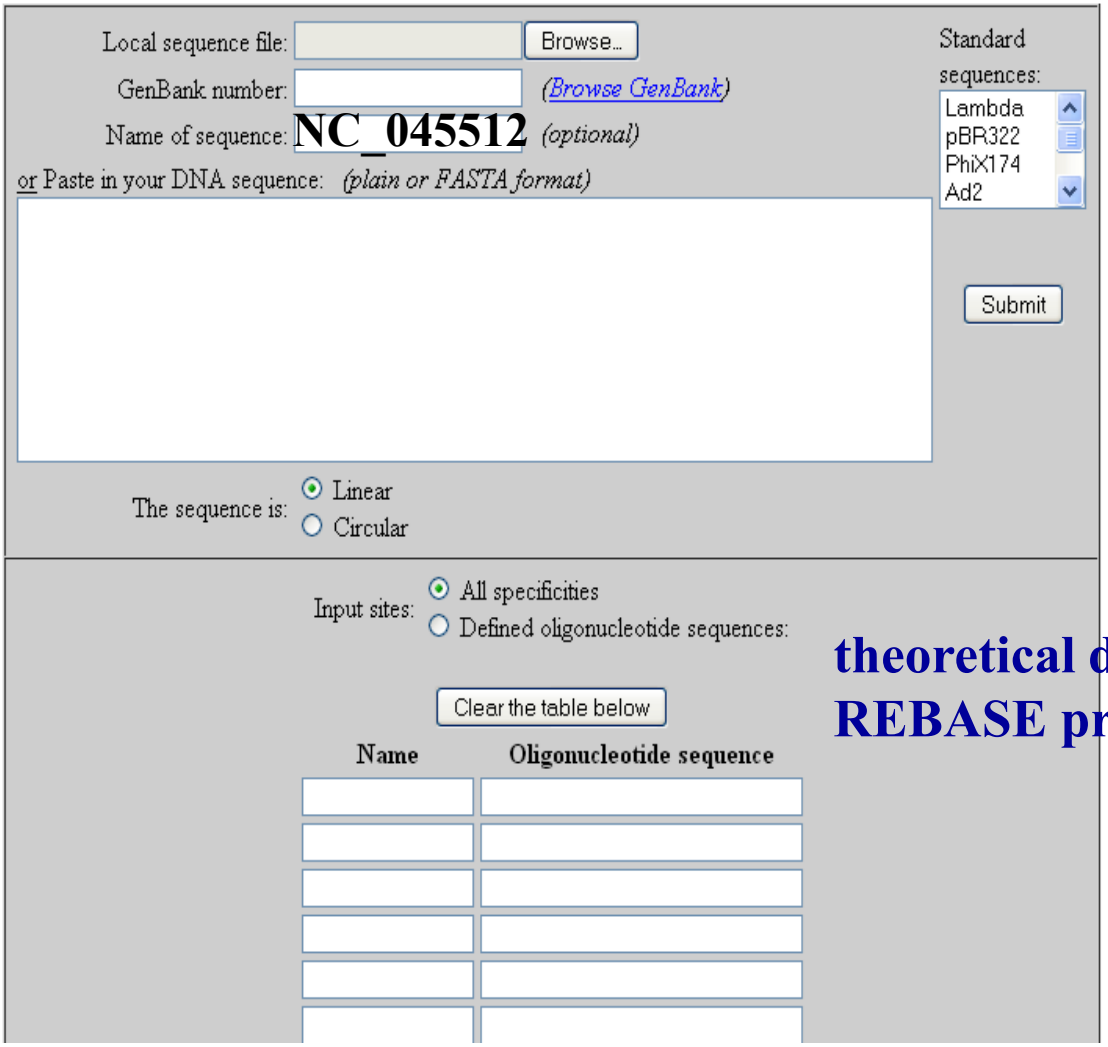
Ref: Robert et al, *Nucl. Acids Res.* 43: D298-D299 (2015)

Back to...

Program Guide

REBsites

This tool will take a DNA sequence and digest it with one example of each of the known Type 2 restriction enzyme specificities.
The maximum size of the input file is 2 MByte, and the maximum sequence length is 200 KBases.



theoretical digest with all REBASE prototypes

[New DNA]

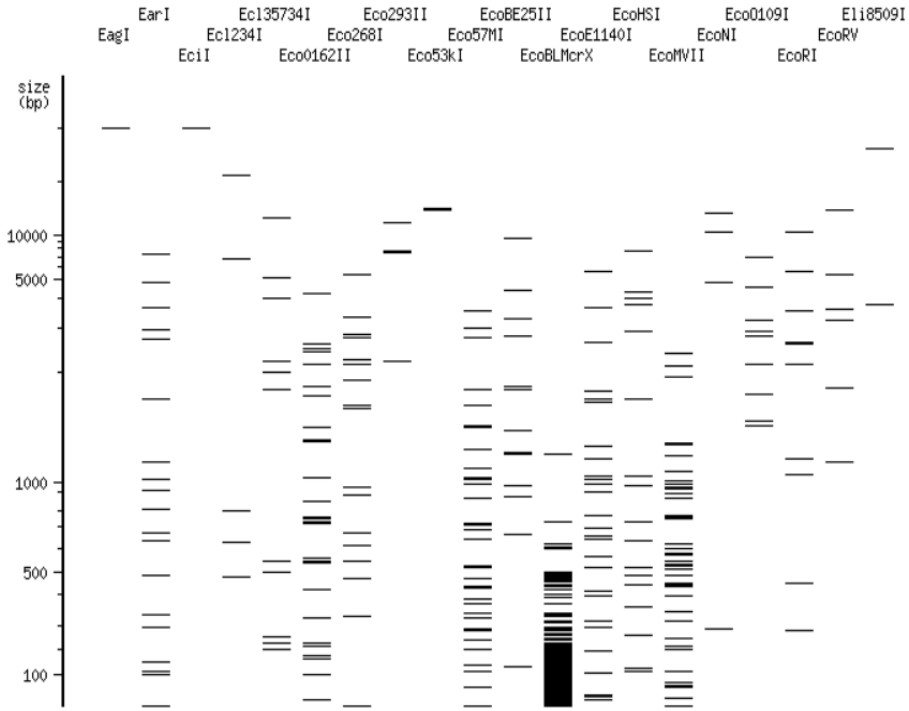
REBsites

NC 045512

Gel: Agarose 0.7%
Order by: enzyme name

[<< Prev] 1 2 3 4 5 6 7 8 9 10 11 12 **13** 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 [Next >>]

[Print]



Click on an enzyme name for a list of fragments/sites.

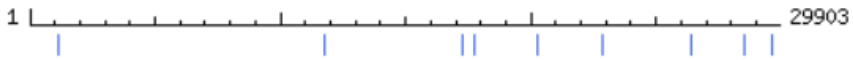
Print

Fragment list

Close

NC 045512 digested with EcoRI

[Sites with flanks]



#	Location	Size [bp]
1	1162-11734	10573
2	11735-17280	5546
3	22871-26439	3569
4	20279-22870	2592
5	17729-20278	2550
6	26440-28551	2112
7	1-1161	1161
8	28552-29620	1069
9	17281-17728	448
10	29621-29903	283

Assignment

- **Write a program to generate a restriction map for Wuhan isolate-1 genome (Acc. Id.: NC_045512) using EcoRI as RE compare your results with REBsites.**
- **Write a program to identify restriction recognition sites in the given DNA sequence.**

Cloning

What is cloning?

The process of cloning involves the production of multiple copies of a DNA fragment of interest by amplification *in vivo*

- depends upon the ability of vectors to continue their life cycles in bacterial or yeast cells in spite of having foreign DNA inserted into them.

Cloning vector - a DNA molecule that carries foreign DNA into a host cell, replicates inside a bacterial (or yeast) cell and produces many copies of itself and the foreign DNA

- a vector containing foreign DNA is termed recombinant vector

Features of Cloning Vectors:

- **sequences that permit the propagation of itself in bacteria (or yeast)**
- **a cloning site to insert foreign DNA; the most versatile vectors contain a site that can be cut by many REs**
- **a method of selecting for bacteria (or yeast) containing a vector with foreign DNA; usually accomplished by selectable markers for drug resistance**

Major requirement of all vectors - an origin of replication for a given host cell in order that they may replicate autonomously (i.e., independently of the host's chromosome)

Types of Vectors

Vector	Insert size (kb)
Plasmids	<10 kb
Bacteriophage	9 - 20 kb
Cosmids	33 - 47 kb
Bacterial artificial chromosomes (BACs)	75 - 125 kb
Yeast artificial chromosomes (YACS)	100-1000 kb

Types of Vectors

Plasmids - an extra-chromosomal double-stranded circular DNA molecules that replicates autonomously inside the bacterial cell

Plasmids are important as one can:

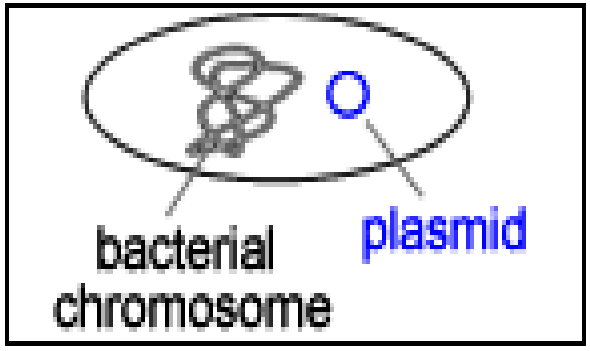
(i) isolate them in large quantities,

(ii) cut & splice them, add DNA of choice,

(iii) put them back into bacteria, where they replicate along with the bacteria's own DNA,

(iv) isolate them again to get billions of copies of inserted DNA

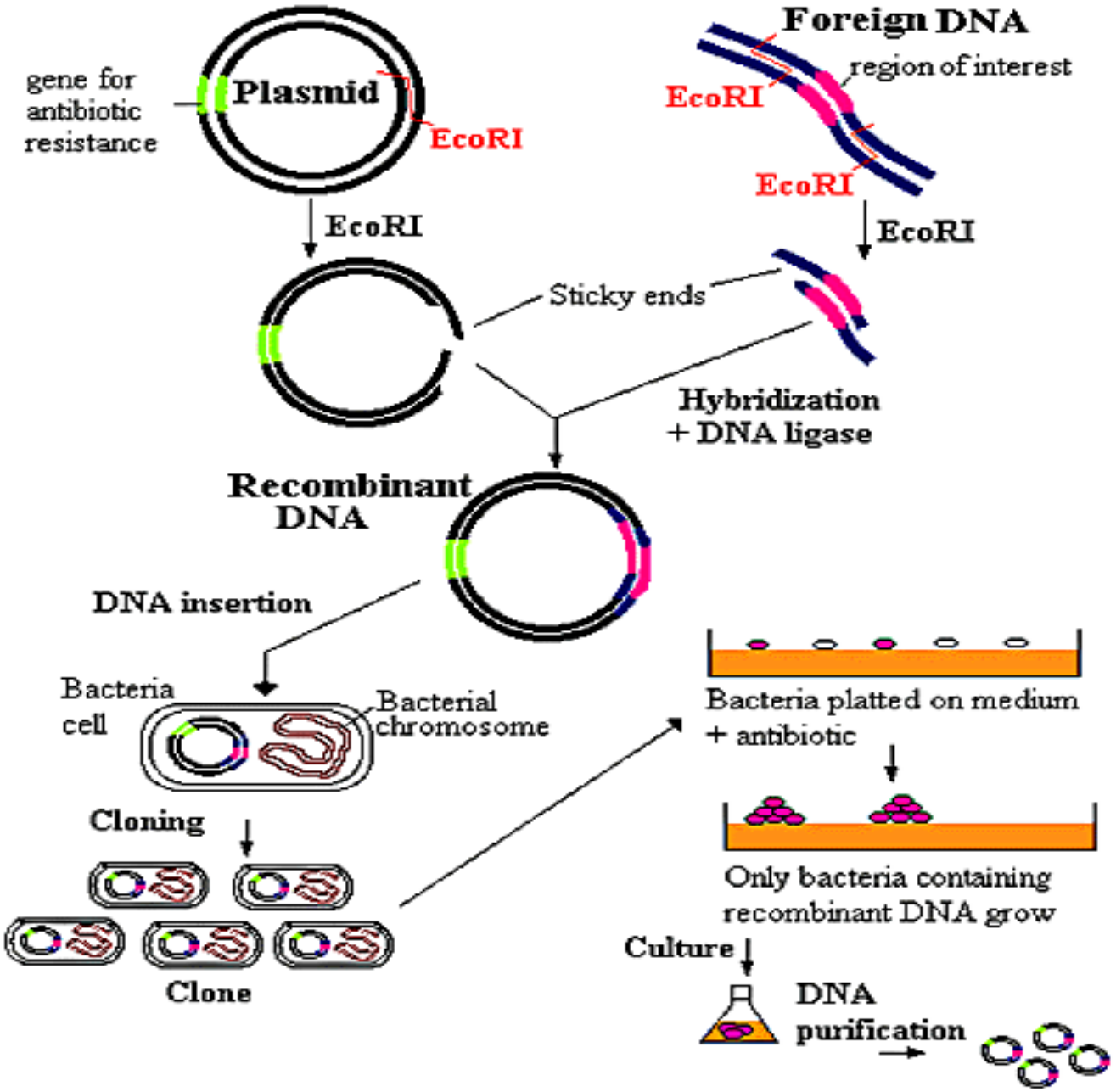
Limitation: size of DNA that can be introduced into the cell by transformation (~ 2 - 10kb)



Plasmid vectors are derived from naturally occurring plasmids of *E. coli* such as ColE1 or from related plasmid pMB1

pBR322 – most widely used cloning vectors of *E. coli*, is a hybrid between ColE1 & genes coding for resistance to antibiotics tetracycline & ampicillin

What's the advantage of inserting genes coding for resistance to antibiotics into a vector?



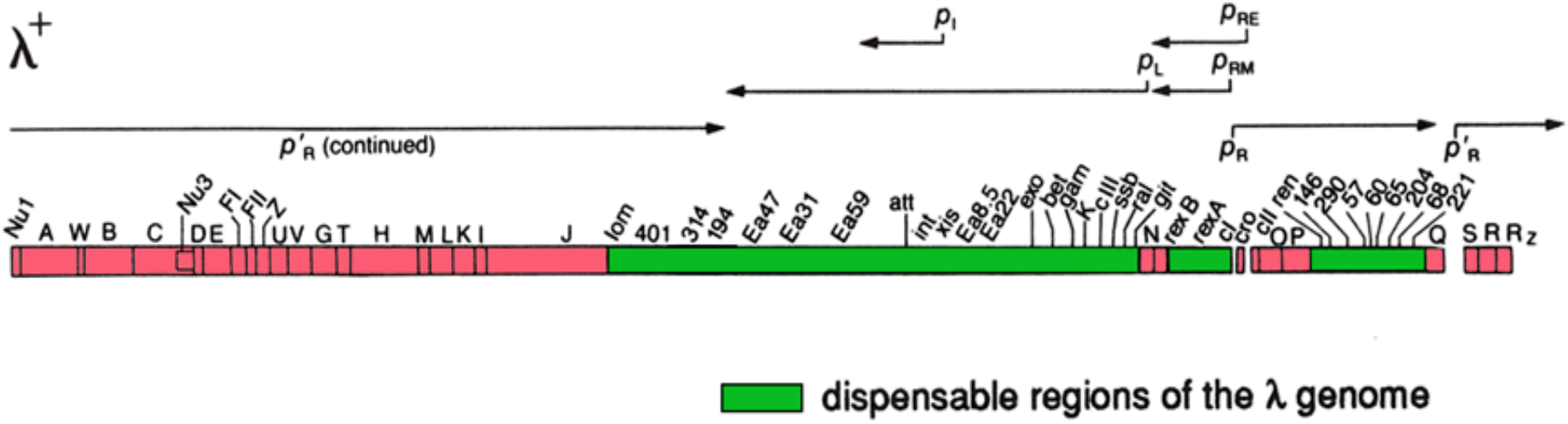
Cloning into a plasmid

Types of Vectors

Bacteriophage Vectors

– a double-stranded linear molecule of size 49.5Kbp

Cloning limit: 9 - 20 kb



Enterobacteria phage λ is a bacterial virus, or bacteriophage, that infects the bacterial species *E. coli*.

Artificially Constructed Vectors

Cosmids - an extra-chromosomal circular DNA molecule that combines features of plasmids and cos gene of phage lambda

Cloning limit: 35 - 50 kb

BAC - Bacterial Artificial Chromosome

- based on naturally occurring F-factor plasmid found in the bacterium E. coli.

Cloning limit: 100-300 kb

YAC - Yeast Artificial Chromosomes

- it is a vector constructed from yeast DNA, used to clone large DNA fragments

Cloning limit: 100-1000 kb

Useful for cloning long segments of eukaryotic DNA

YAC - a functional self-replicating artificial chromosome. It includes three specific DNA sequences that enable it to propagate from one cell to its offspring:

- **TEL: The telomere which is located at each chromosome end, protects the linear DNA from degradation by nucleases**
- **CEN: The centromere which is the attachment site for mitotic spindle fibers, "pulls" one copy of each duplicated chromosome into each new daughter cell.**
- **ORI: Replication origin sequences, specific DNA sequences that allow the DNA replication machinery to assemble on the DNA and move at the replication forks**

It also contains few other specific sequences like:

- **A and B: selectable markers that allow easy isolation of yeast cells that have taken up the artificial chromosome.**
- **Recognition site for two REs: EcoRI & BamHI**

Why is it important to be able to clone large sequences?

To map the entire human genome ($3x10^9$ bps) would require more than 1000,000 plasmid clones (~10Kb limit).

In principle, the human genome could be represented in about 10,000 YAC clones (~1Mb limit)

What determines the choice vector?

- **insert size**
- **vector size**
- **restriction sites**
- **copy number**
- **cloning efficiency**
- **ability to screen for inserts**

DNA Sequencing

DNA Sequencing - determine the precise sequence of nucleotides in a sample of DNA – the order of A, T, G, C

Various types of sequencing:

- **Sequencing a region of interest, e.g., gene.**
- **Whole Genome/Exome Sequencing**
- **cDNA Sequencing – sequencing cDNA libraries of the expressed genes**
- **High-throughput sequencing – next-generation, 3rd & 4th generation sequencing - whole Genome/Exome/targeted**
- **Metagenome sequencing - sequencing of environmental samples**

- depending on the nature of analysis, type of sample, or type of sequencer used

Sequencing a Region of Interest

First requirement in sequencing a region of DNA is

- to have enough starting template for sequencing.

This is achieved by PCR - Polymerase Chain Reaction

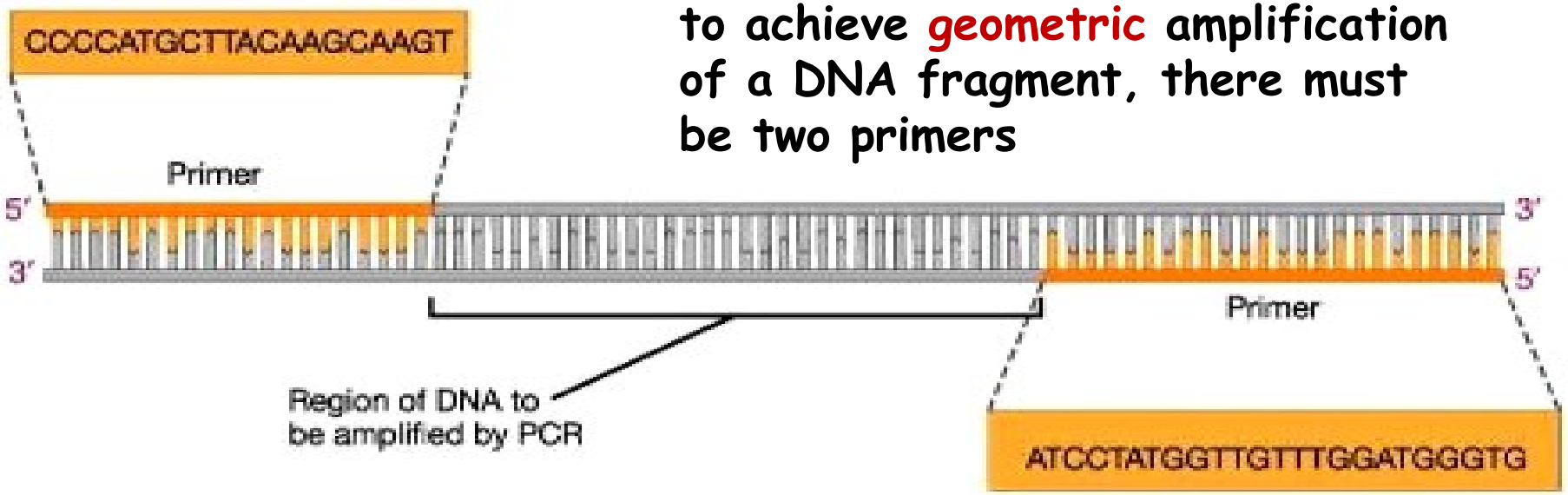
- **carried out in an automated cycler for 30 - 40 cycles.**

Essential requirements for a PCR:

- **a mixture of 4 deoxy-nucleotides in ample quantities: dATP, dGTP, dCTP, dTTP**
- **Taq DNA polymerase**
- **Primers ?**
- **Genomic DNA of interest**

What is the advantage of using PCR over traditional gene cloning?

Region of DNA to be amplified by PCR



to achieve geometric amplification of a DNA fragment, there must be two primers

Primers - short single-stranded oligonucleotides which anneal to the DNA template and serve as a starting point for DNA synthesis

Why are primers required?

The Cycling Reactions

Step-1: Denaturation at 94°C

- opens up double stranded DNA, all enzymatic reactions stop.

Step-2: Annealing at 54°C

- Primers jiggling around because of Brownian motion, binds to single stranded template once an exact match is found; the polymerase then attaches and start copying the template.

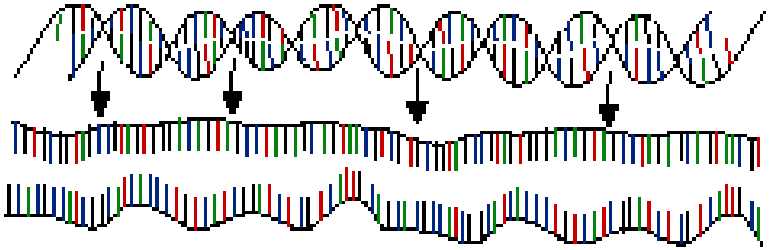
Step-3: Extension at 72°C

- ideal working temperature for the polymerase. Bases complementary to the template are coupled to the primer on 3' side (reading the template from 3' to 5' side)

Different Steps in PCR

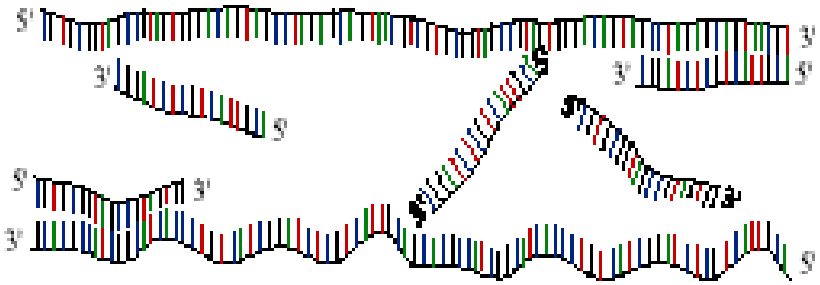
PCR : Polymerase Chain Reaction

30 - 40 cycles of 3 steps :



Step 1 : denaturation

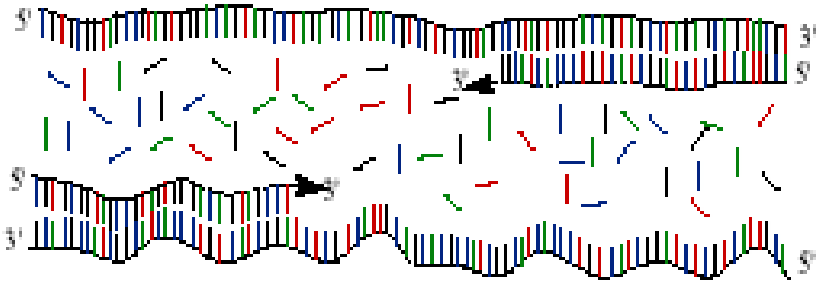
1 minut 94 °C



Step 2 : annealing

45 seconds 54 °C

forward and reverse primers !!!

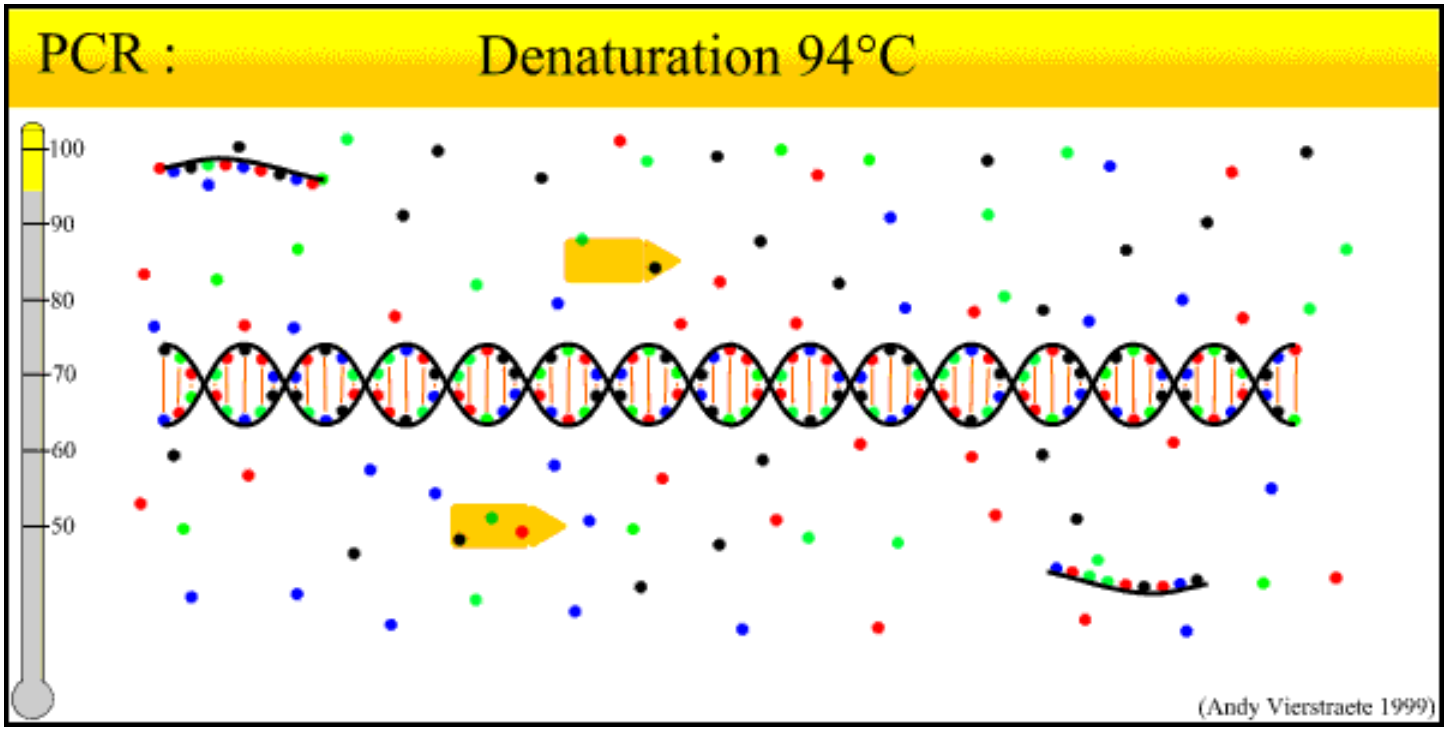


Step 3 : extension

2 minutes 72 °C

only dNTP's

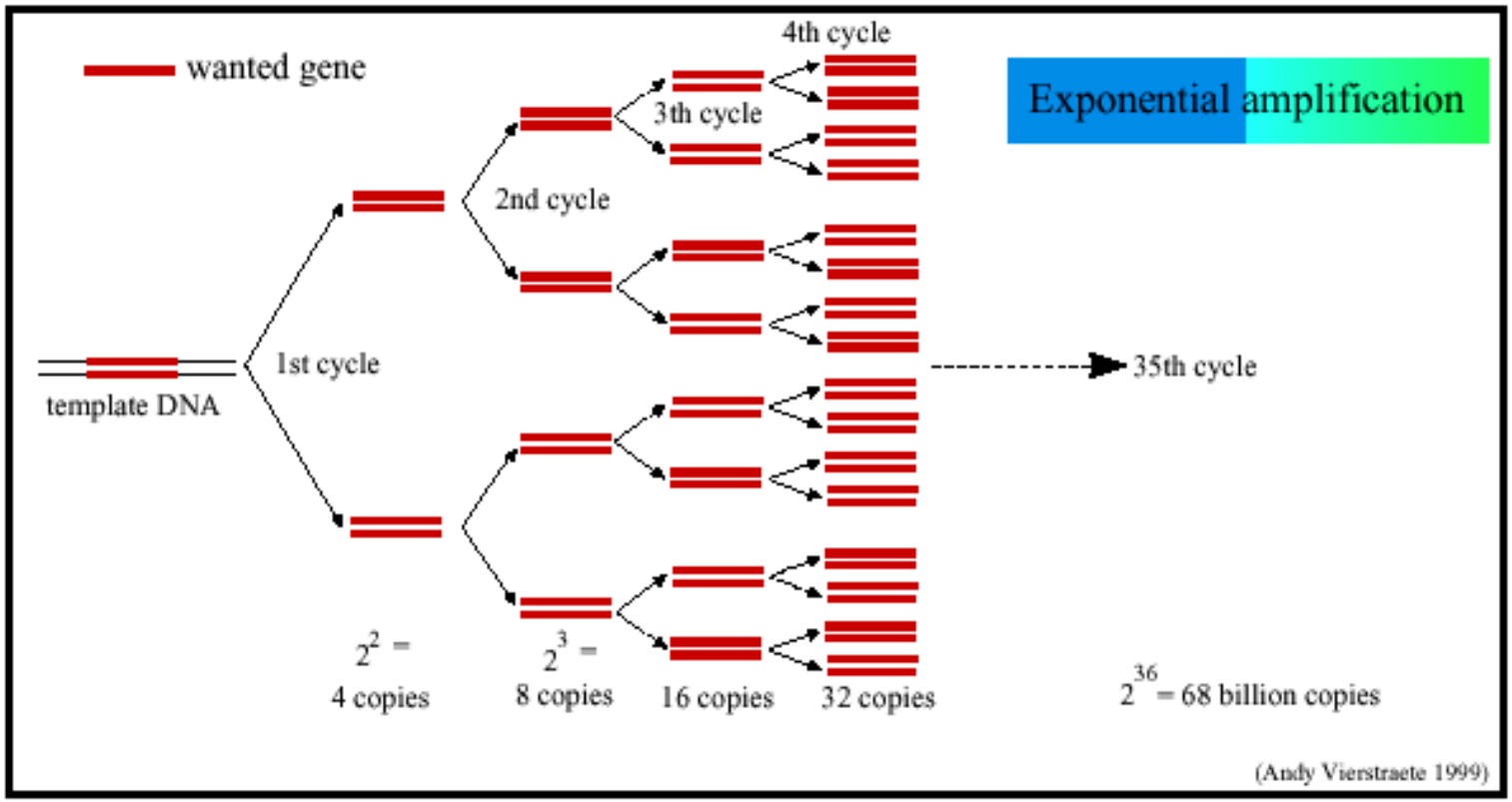
Different Steps in PCR



Exponential amplification of region of interest

<u>Both strands</u> are copied during PCR

- leading to an exponential increase of the number of copies of the region of interest.



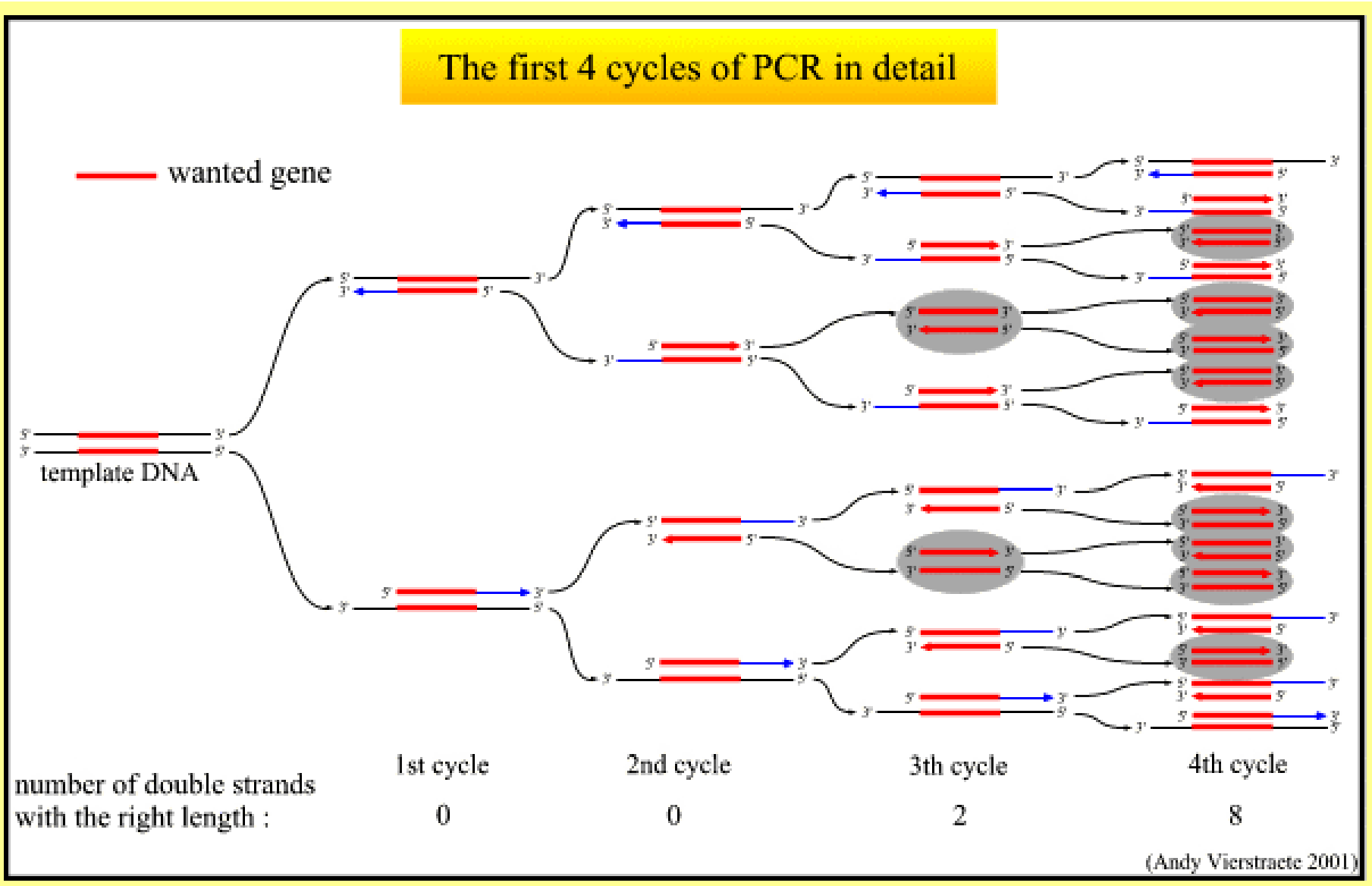
Verification of PCR Product

Is the template copied during PCR and is it the right size?

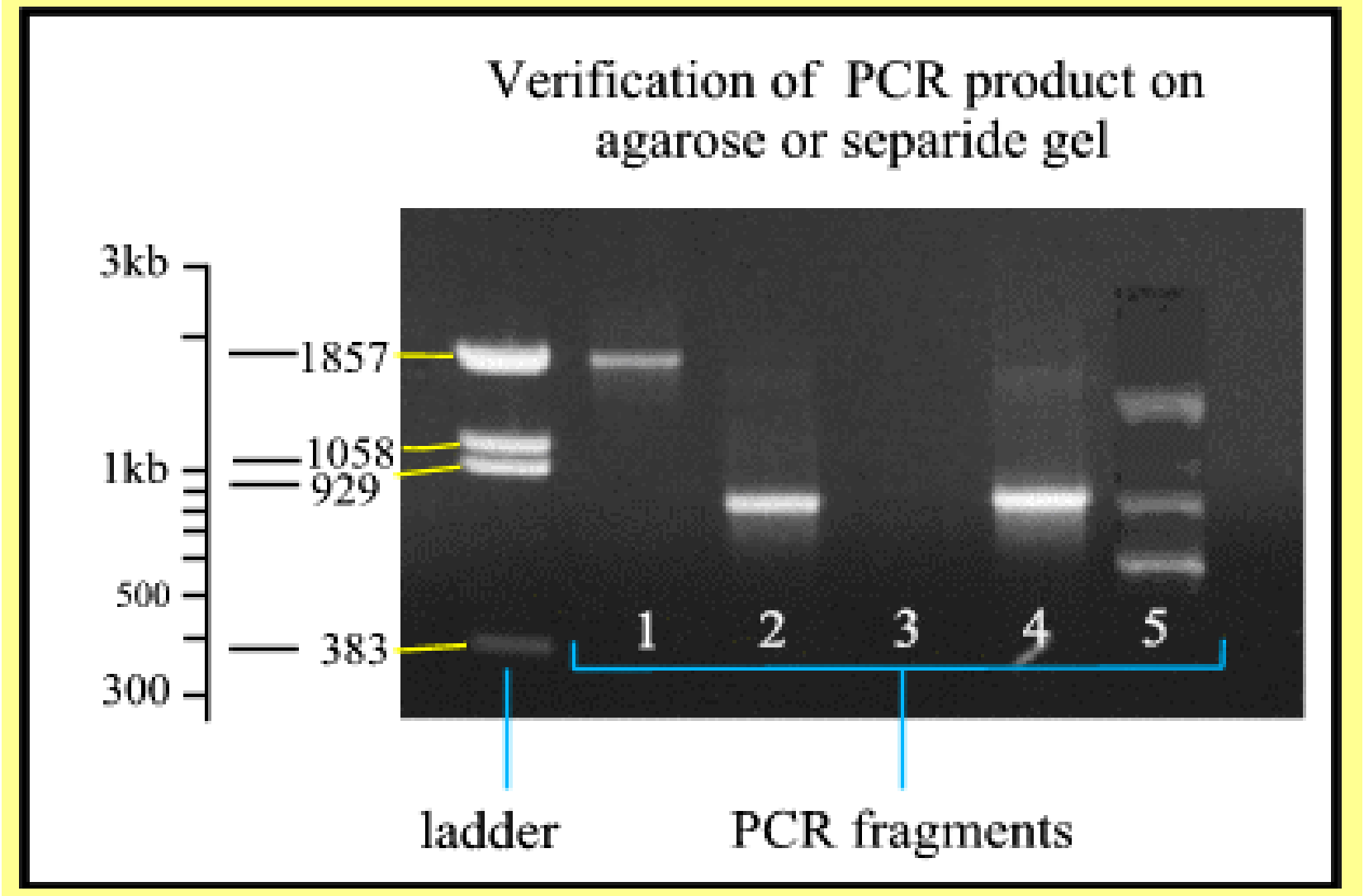
Before the PCR product is used in further applications, it has to be checked if:

1. **A product is formed**
2. **The product is of the right size**
3. **Only one band is formed**

First 4 cycles of a PCR reaction



Verification of the PCR product

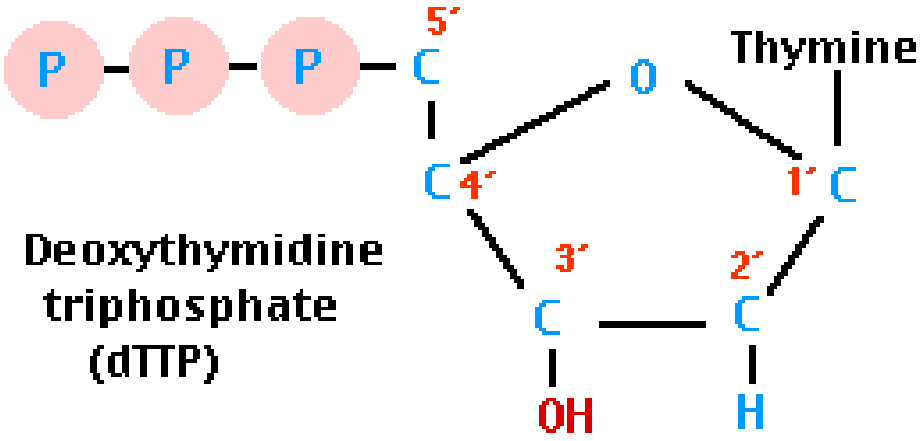


PCR Sequencing

For sequencing, we don't start from gDNA (like in PCR) but mostly from PCR fragments or cloned genes.

Amplified PCR product is supplied with

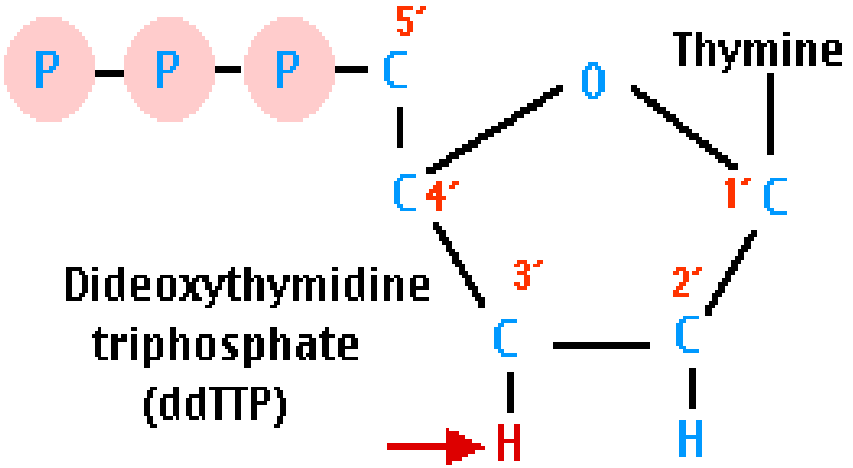
- **a mixture of all four <u>normal</u> (deoxy) nucleotides in ample quantities**
  - **dATP**
  - **dGTP**
  - **dCTP**
  - **dTTP**



- ***Taq* DNA polymerase**

PCR Sequencing

- a mixture of all four dideoxynucleotides, each present in limiting quantities and each labeled with a "tag" that fluoresces a different color:
  - ddATP
  - ddGTP
  - ddCTP
  - ddTTP



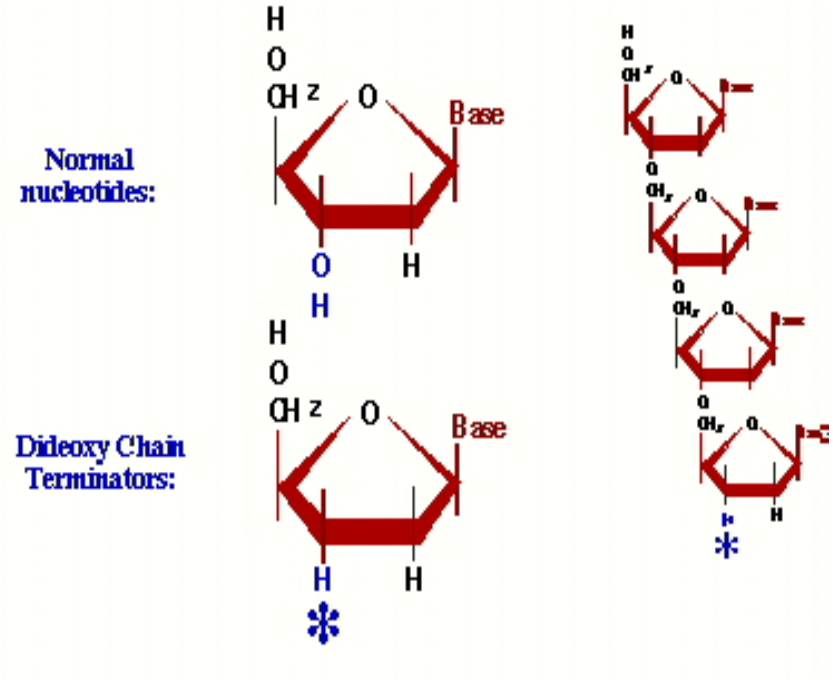
This method of DNA sequencing is called dideoxy method, or chain termination method, or Sanger's method.

PCR Sequencing

Dideoxy method: DNA is synthesized from four deoxynucleotide triphosphates.

Each new nucleotide is added to 3′ -OH group of the last nucleotide added.

When a dideoxynucleotide, ddNTP is added to the growing DNA strand, chain elongation stops because there is no 3′-OH for the next nucleotide to be attached to.



Steps in PCR Sequencing

I The sequencing reaction

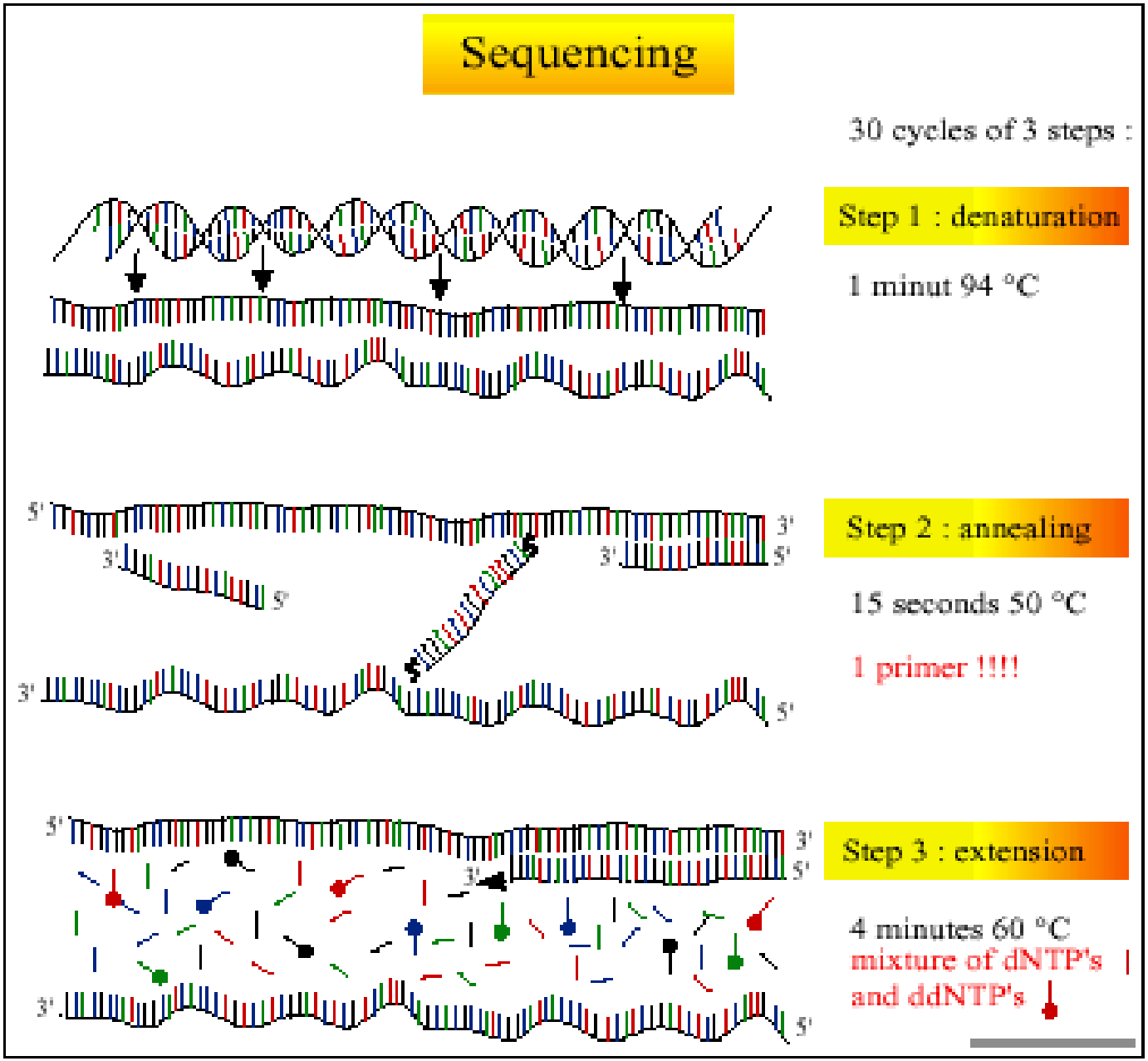
- **Denaturation at 94°C**
- **Annealing at 50°C**
- **Extension at 60°C** ← **instead of 72°C**

II Separation of the fragments

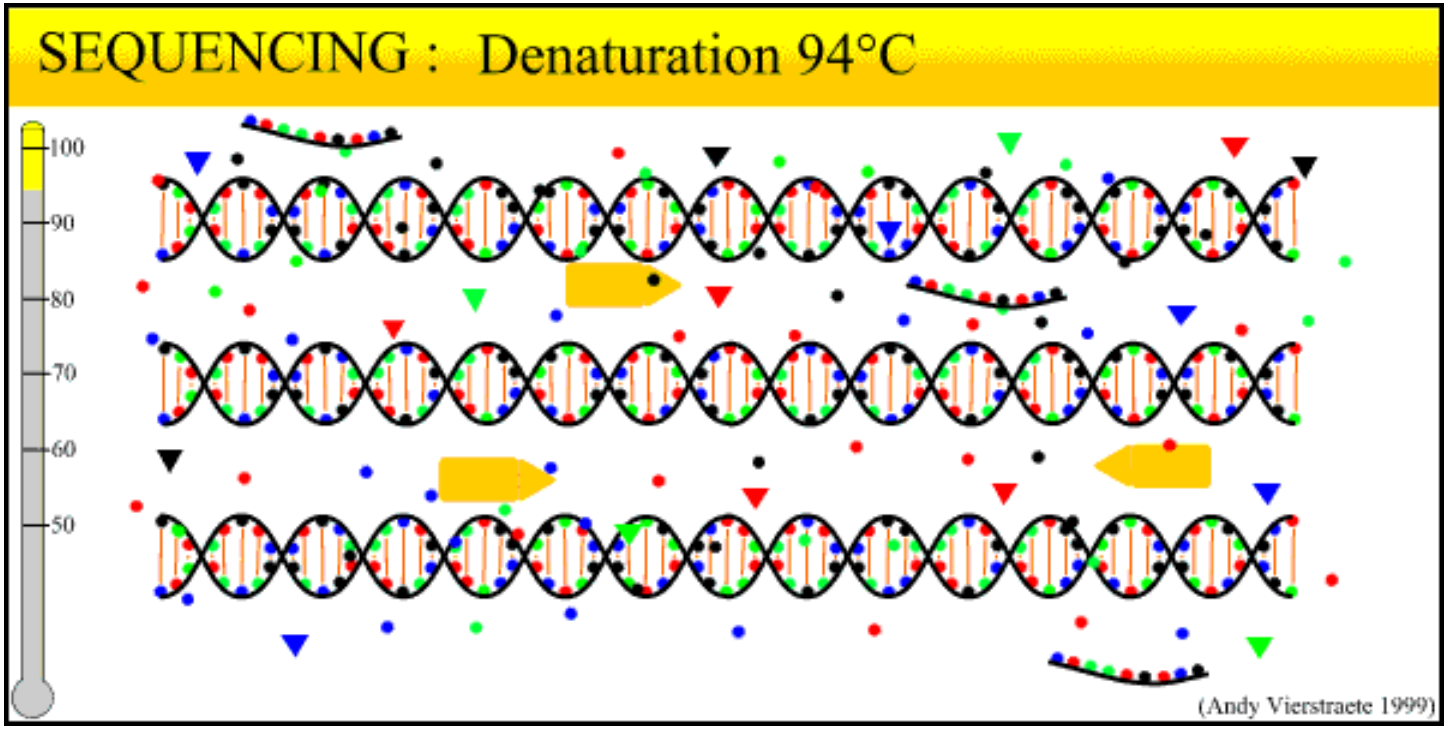
III Detection on an automated sequencer

IV Assembling the sequenced parts

Different steps in Sequencing

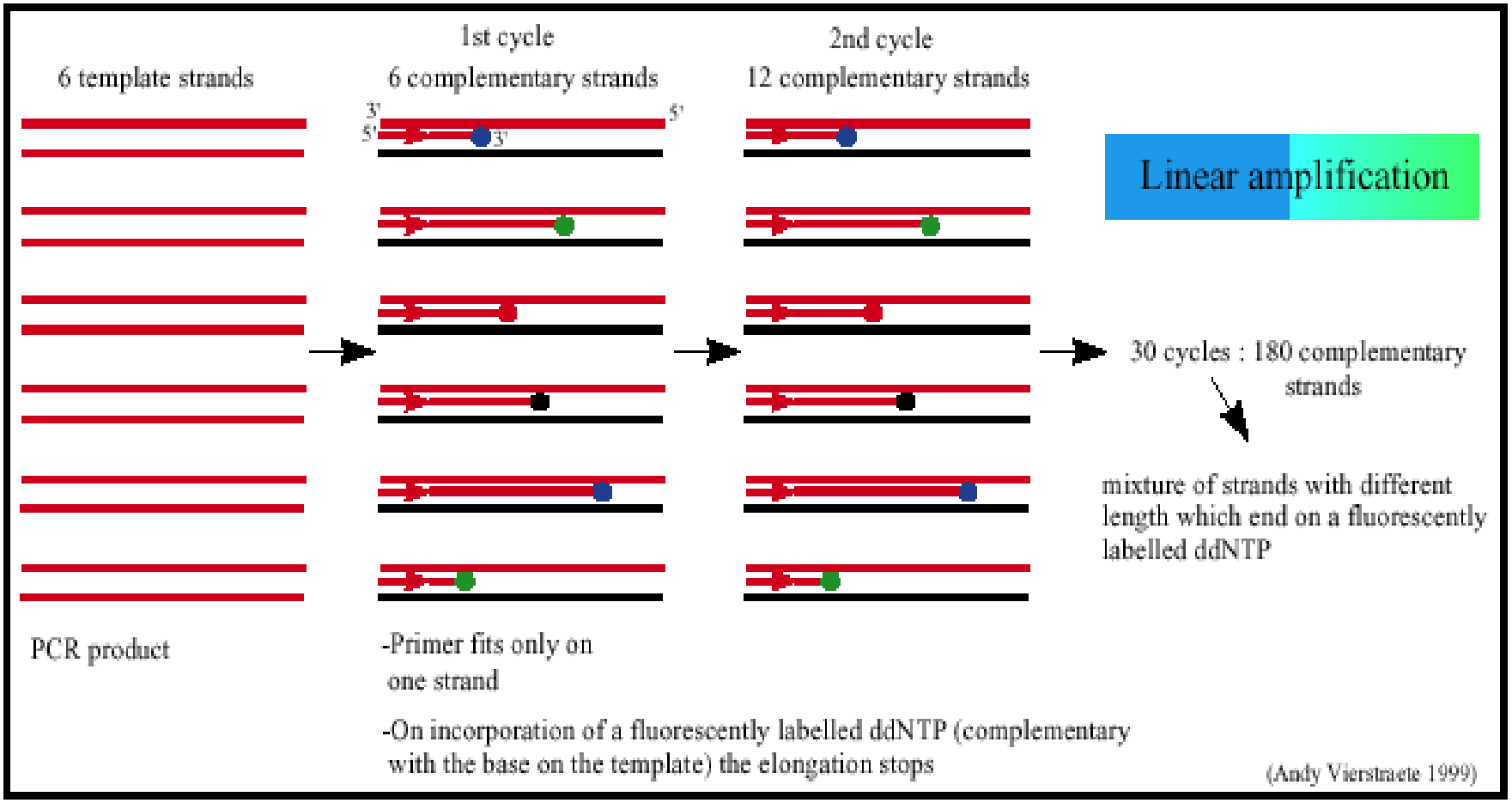


Different steps in Sequencing



PCR Sequencing

Since only one primer is used, only one strand is copied during sequencing – resulting in a linear increase of the number of copies of one strand of the gene. Hence, a large amount of DNA in the starting mixture for sequencing is required.



PCR Sequencing

II Separation of the molecules:

After the sequencing reactions, the mixture of strands of different lengths, all ending on a fluorescently labeled ddNTP, need to be separated

- done by loading the mix on an acrylamide gel - gel electrophoresis.

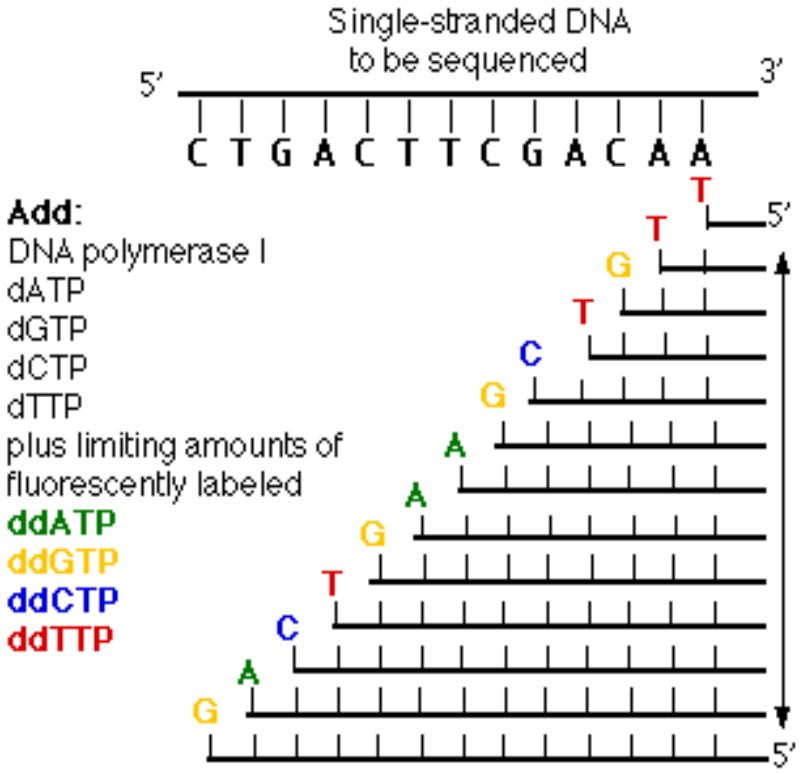
During electrophoresis, a voltage is created across the gel making one end positive and the other negative. DNA being –vely charged, migrates to the positive side.

DNA strands of different length migrate at different rates and thus can be separated based on their size - the smallest strand travels the fastest.

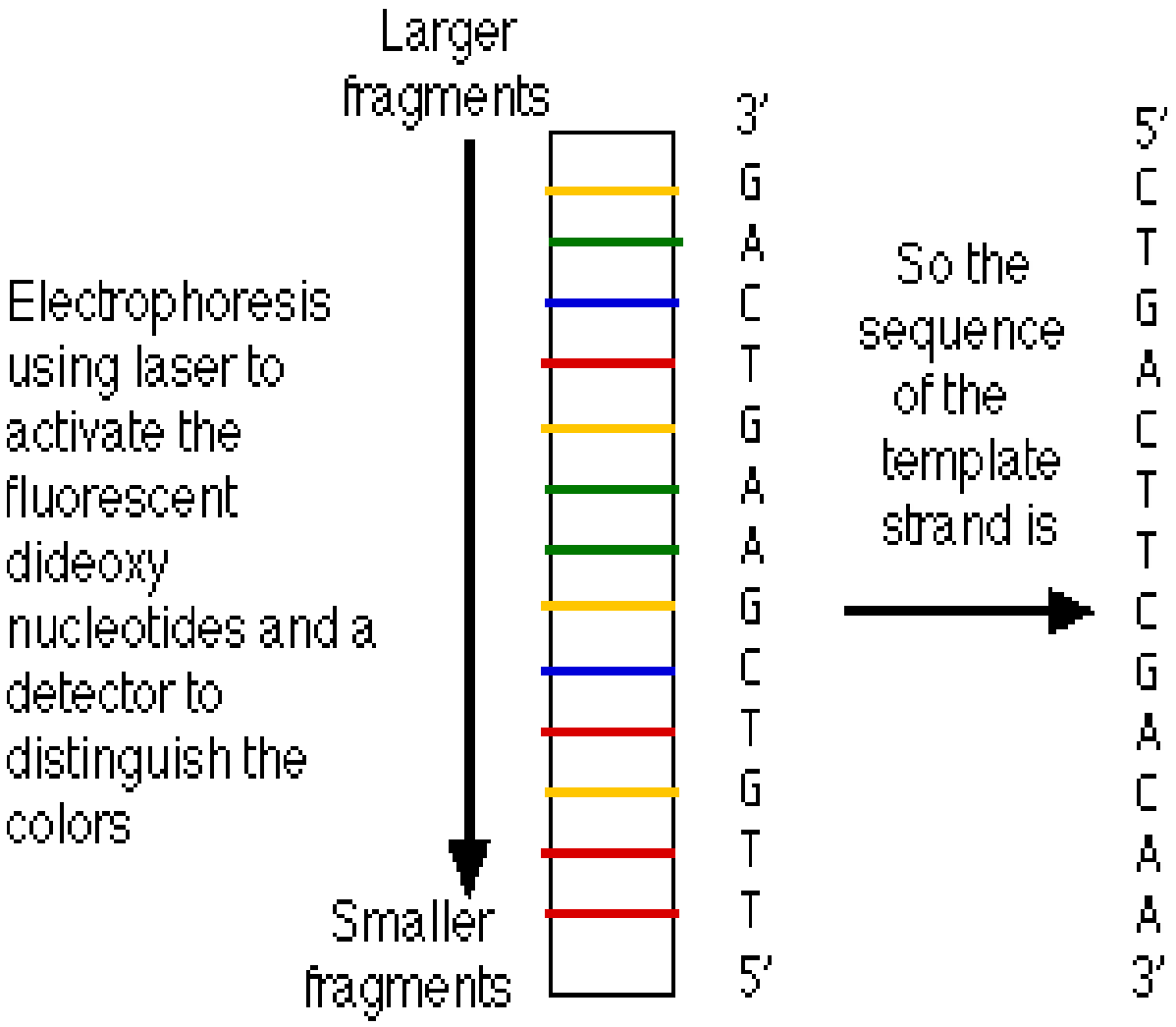
Separation of molecules with electrophoresis

Very good resolution - a difference of even one nucleotide is enough to separate a strand from the next shorter or longer strand.

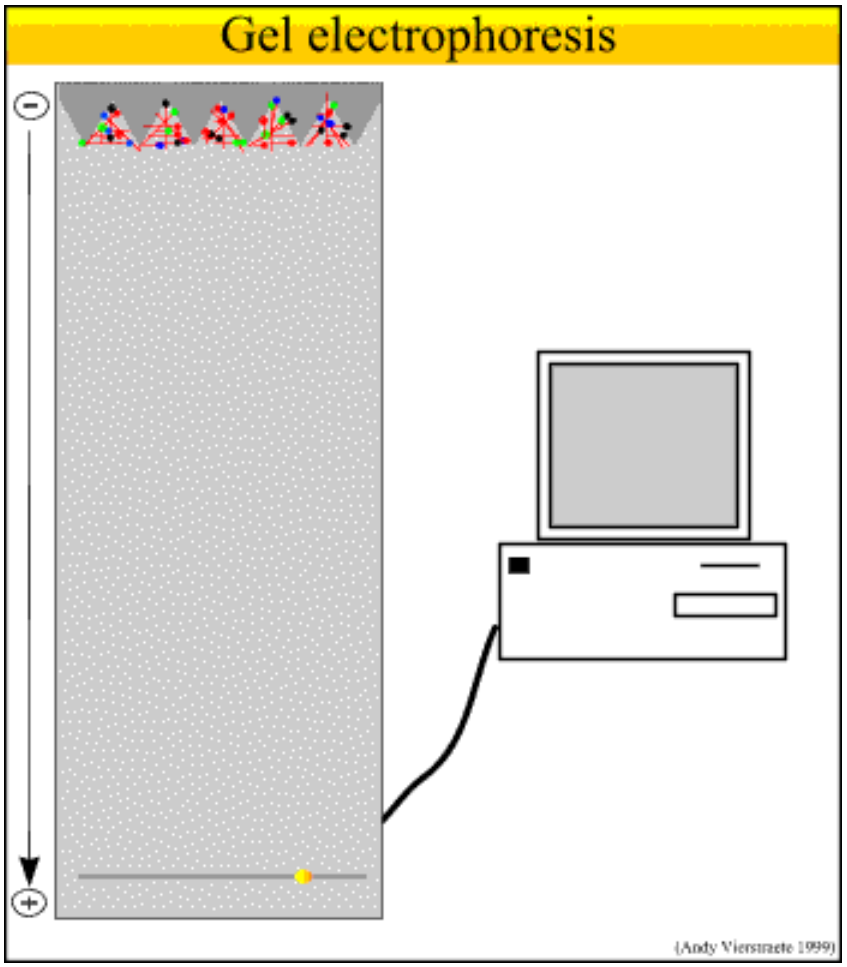
Four dideoxynucleotides fluoresces a different color when illuminated by a laser beam and an automatic scanner provides a printout of the sequence.



Separation of Molecules with Electrophoresis



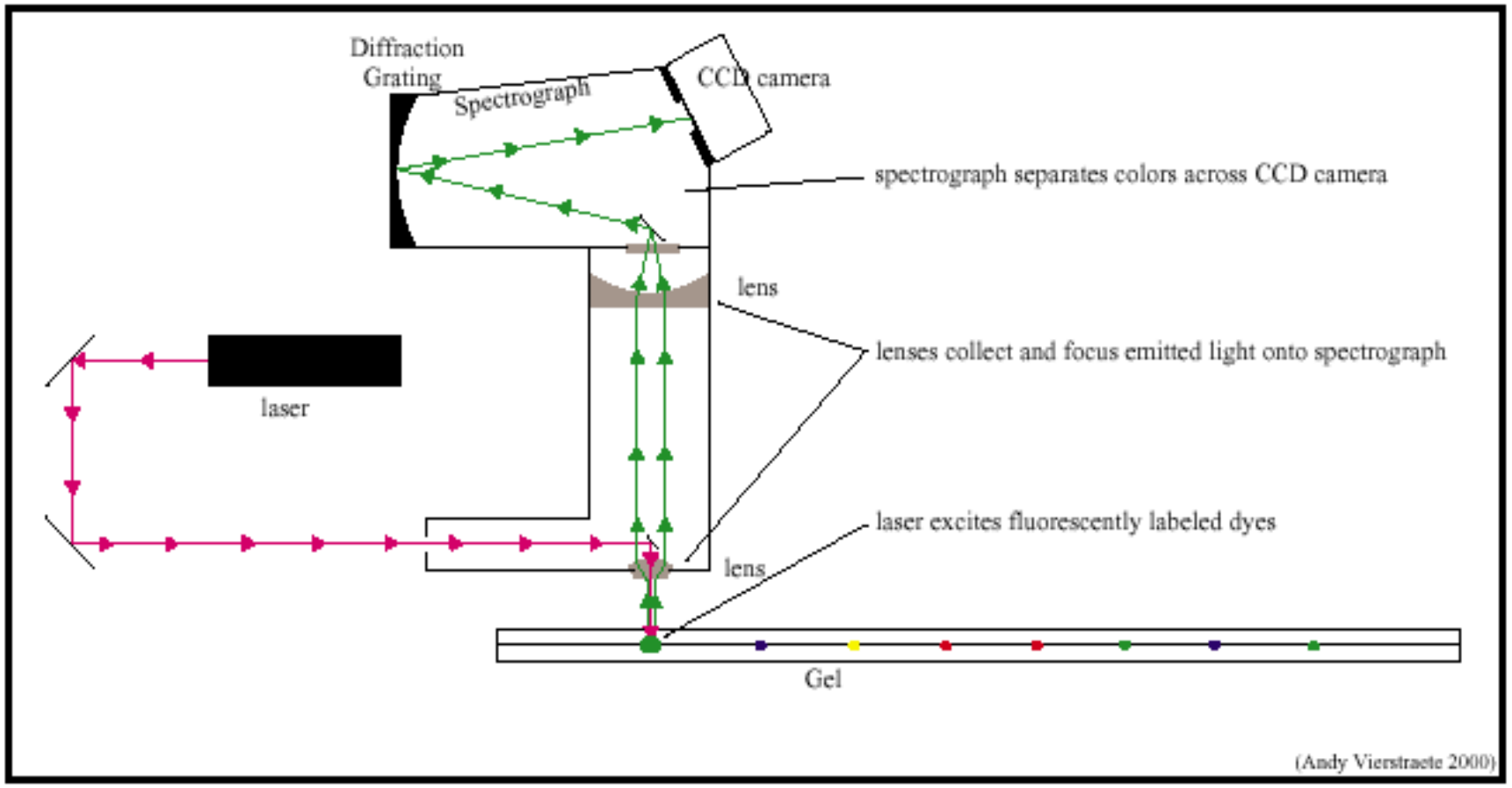
Separation of the Molecules with Electrophoresis



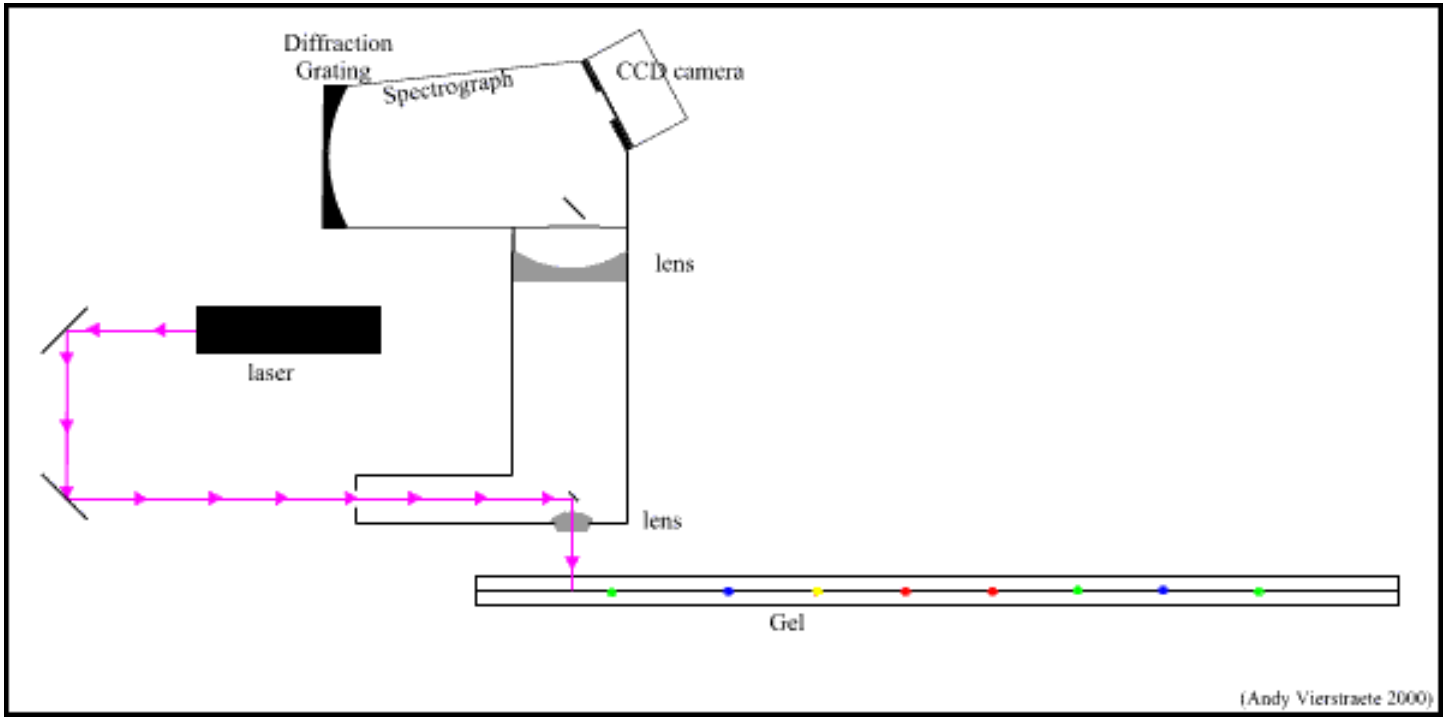
PCR Sequencing

III Detection on an automated sequencer:

Fluorescently labeled fragments that migrate through the gel pass a laser beam at the bottom of the gel.

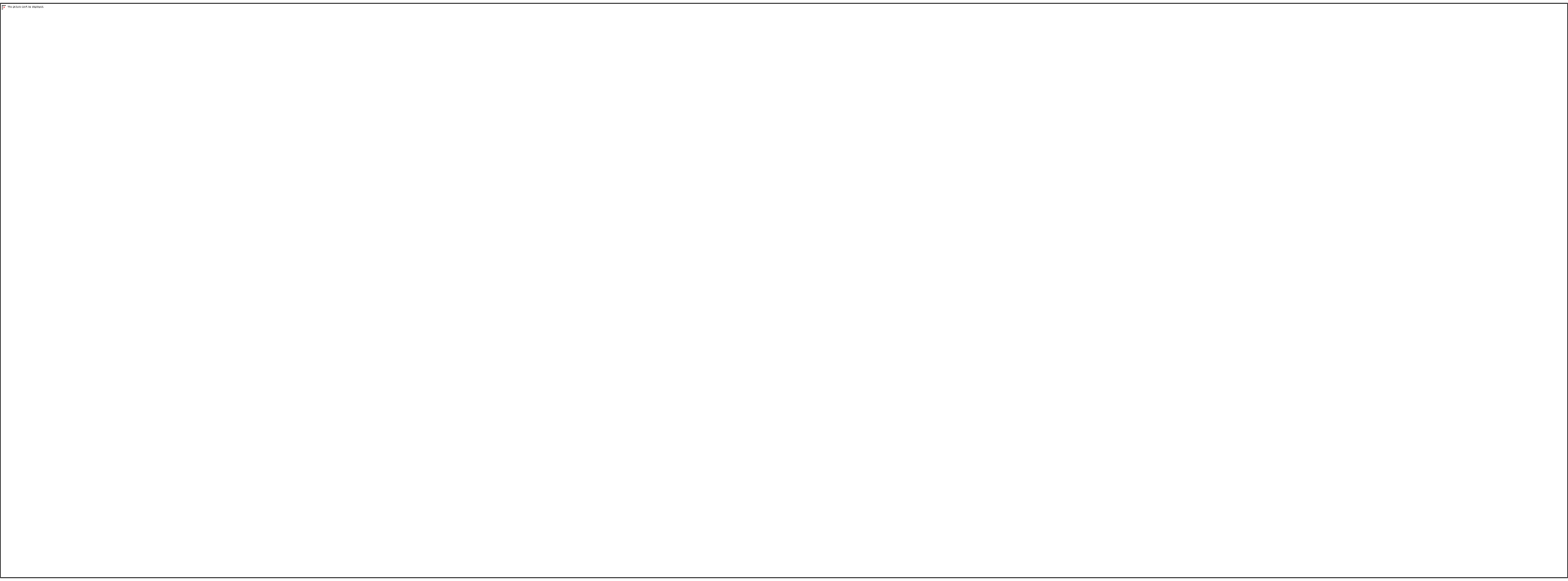


Scanning & Detection System on a Sequencer



PCR Sequencing

Plot of the colors detected in a 'lane' of the gel (one sample), scanned from smallest fragments to largest.



The computer interprets the colors by printing the nucleotide sequence across the top of the plot.

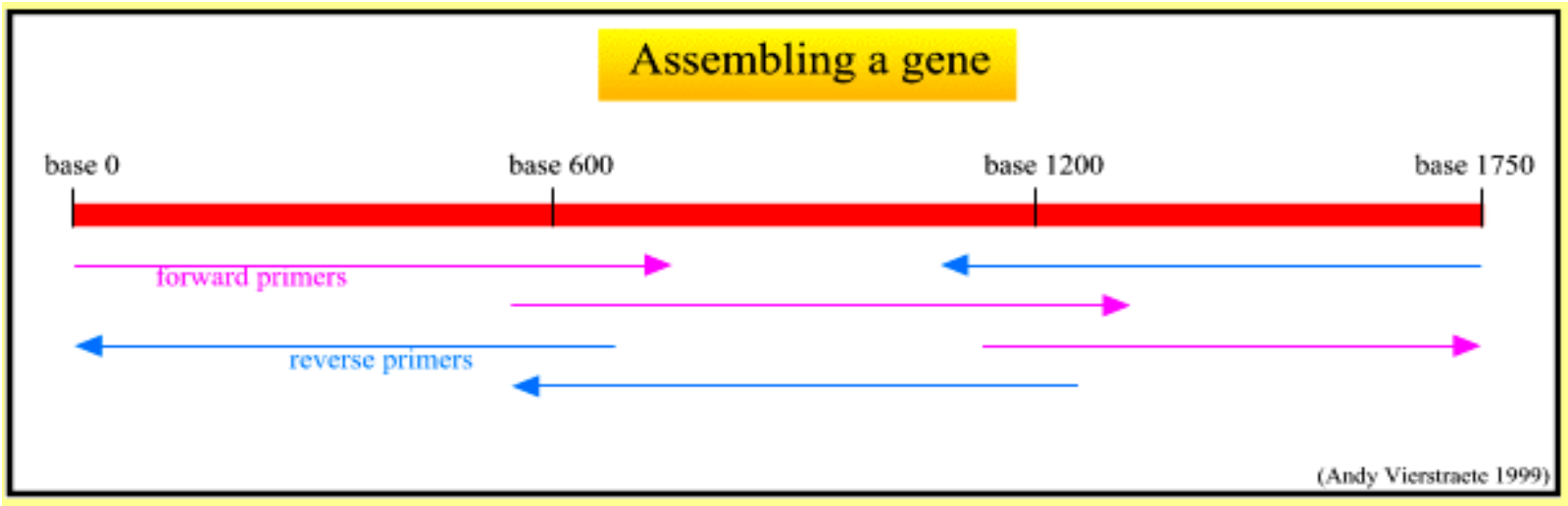
PCR Sequencing

IV Assembling the sequenced parts of a gene:

For publication, a gene sequence has to be confirmed in both directions using forward & reverse primers

Since it is only possible to sequence ~ 700-800 bases in one run, a gene of, say, 1800 bases, is sequenced with internal primers.

- the sequenced fragments are assembled using a computer program to obtain complete gene sequence.



Genome Sequencing

Genome Sequencing

By Sanger's method, we can sequence a fragment of DNA ~ 1000bp long.

But what about longer pieces?

Human genome is 3 billion bases long, arranged on 23 pairs of chromosomes.

Sequencing machine reads just a drop in the ocean!

Genome Sequencing

Solution: Break the entire genome into manageable pieces and sequence them.

Two approaches were used for sequencing Human genome:

- **Publicly funded Human Genome Project (HGP) – clone-by-clone or hierarchical shotgun sequencing method**
- **Privately Funded Sequencing Project - Celera Genomics – whole genome shotgun sequencing method**

Genome Sequencing

Hierarchical shotgun sequencing approach:

- **genomic DNA is cut into pieces of about 150 Mb**
- **inserted into BAC vectors,**
- **transformed into *E. coli* where they are replicated and stored.**

BAC inserts are isolated & mapped to determine the order of each cloned 150 Mb fragment - referred to as the Golden Tiling Path

Begun formally in 1990, Human Genome Project was a 13-yr effort coordinated by the U.S. DAE and NIH.
- completed in 2003

Genome Sequencing

Each BAC fragment in the Golden Path is

- fragmented randomly into smaller pieces,

- each piece is cloned into a plasmid and sequenced on both strands.

These sequences are aligned so that identical regions overlap.

Contiguous pieces are then assembled into finished sequence once each strand had been sequenced about 5 times to produce 10× coverage of high-quality data.

Genome Sequencing

Whole genome shotgun sequencing (WGS)

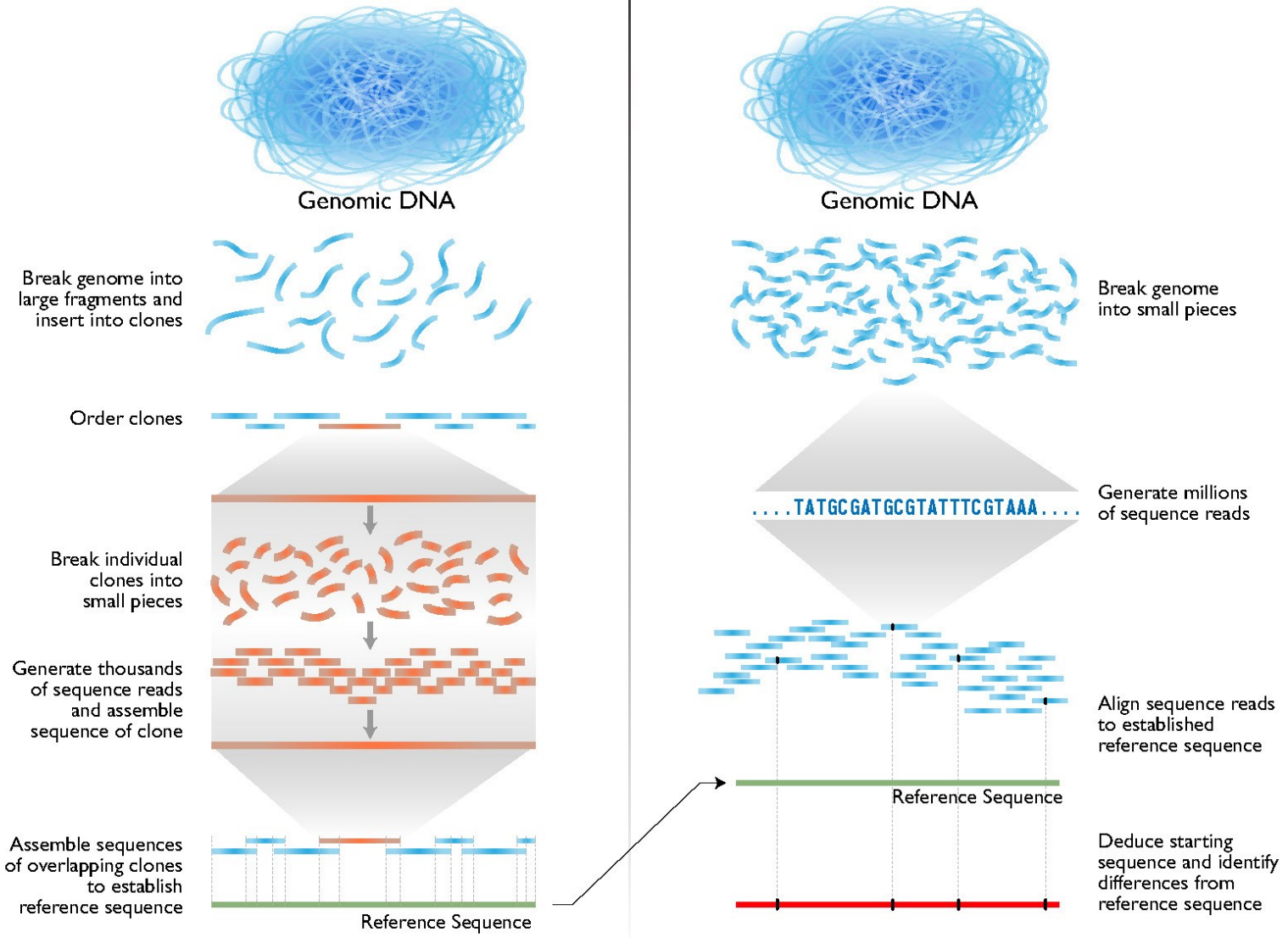
- method developed and preferred by Celera Genomics

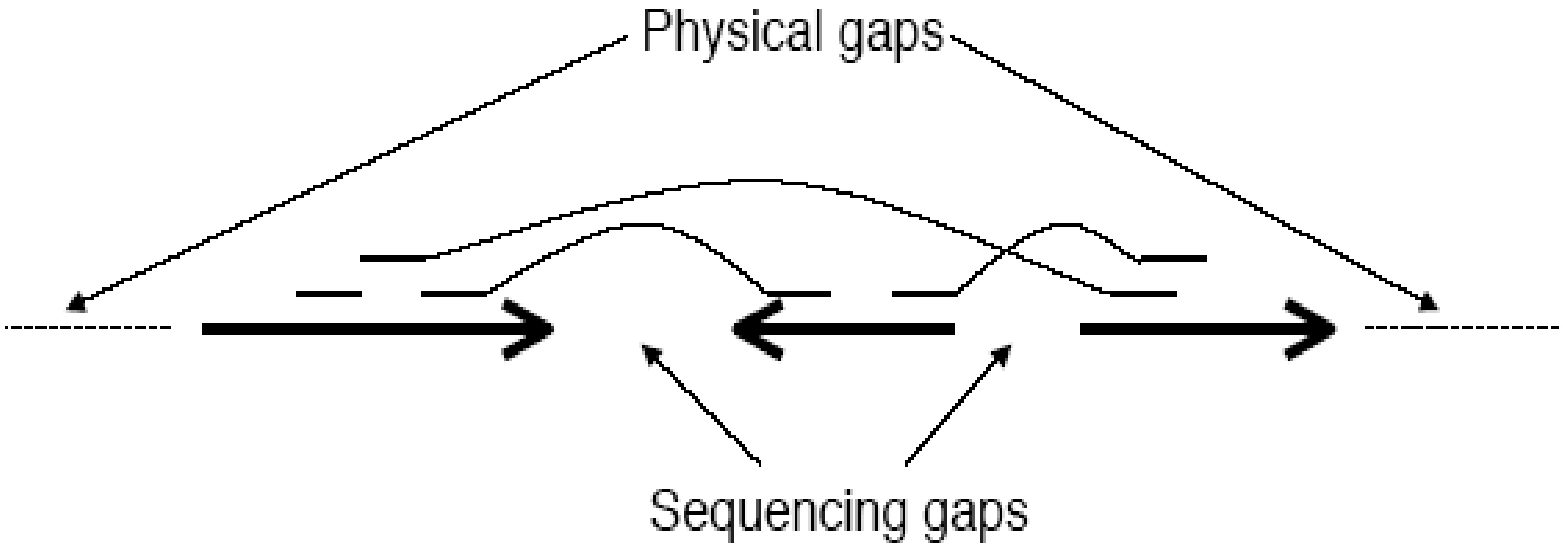
- skips the entire step of making libraries of BAC clones

Blast apart entire human genome into fragments of 2 - 10 kb and sequence them.

Challenge is then to assemble these fragments into the whole genome sequence.

Human Genome Sequencing





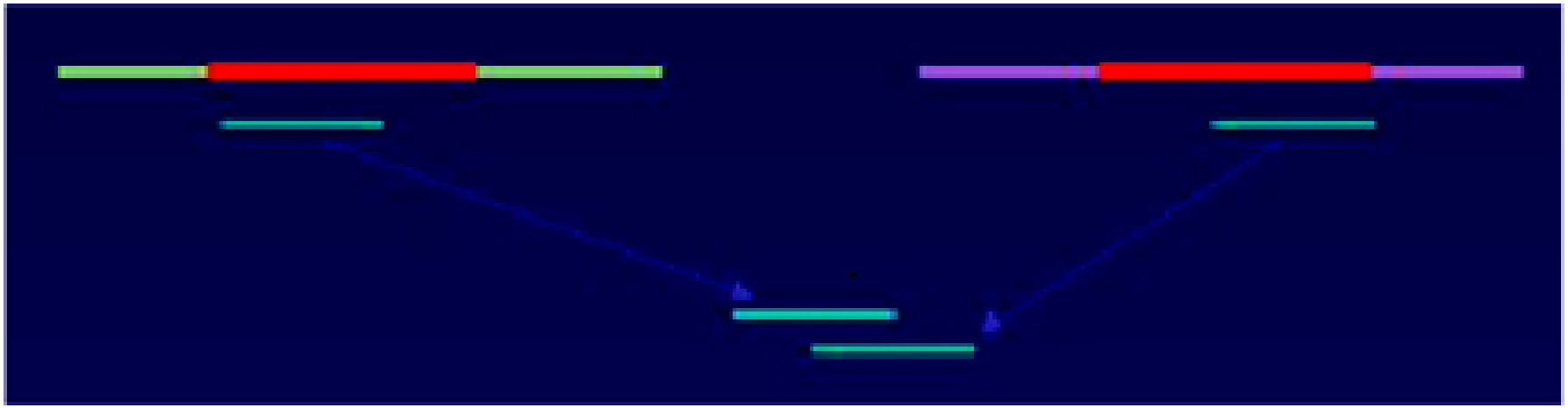
sequencing gap - we know the order and orientation of the contigs and have at least one clone spanning the gap

physical gap - no information known about the adjacent contigs, nor about the DNA spanning the gap

Whole Genome Shotgun Method

What makes the task of assembling the genome fragments especially challenging

- **repeats in the genome (~ 50% in human genome).**



Because of the various ways a fragment could align with a repeat, and the different areas adjacent to the repeats in the original genome, assemblers need to be designed so as not to incorrectly join fragments

Whole Genome Shotgun Method

Adding to the challenge is the sheer computational complexity of the task.

e.g., human genome is 3 billion base pairs long and if the length of one read is 500 bps and the desired coverage is 10x, then $6 * 10^7$ reads would be required:

GenomeLength * DesiredCoverage / ReadLength = RequiredReads

With 60 million reads to assemble, we need algorithms that run in near linear time ($O(nlogn)$)

Whole Genome Shotgun Method

Which method is better?

Depends on the size and complexity of the genome

Note: Celera had access to the HGP data but the HGP did not have access to Celera data.

Which method is preferable for sequencing the genome of a novel coronavirus – SAR-CoV-2? Why?

cDNA Sequencing

Sequencing cDNA Libraries of Expressed Genes

Two common goals in sequence analysis are

- **to identify sequences that encode proteins, which determine all cellular metabolism, and**
- **to discover sequences that regulate the expression of genes or other cellular processes.**

Genomic sequencing meets both the goals.

However, only a small percentage of the genomic sequence actually encodes proteins

cDNA Sequencing

Computational methods for analyzing genomic sequences and finding protein-encoding regions are not completely reliable

cDNA libraries are prepared that have the sequences of the mRNA molecules expressed in the cells, or else cDNA copies are sequenced directly by RT-PCR

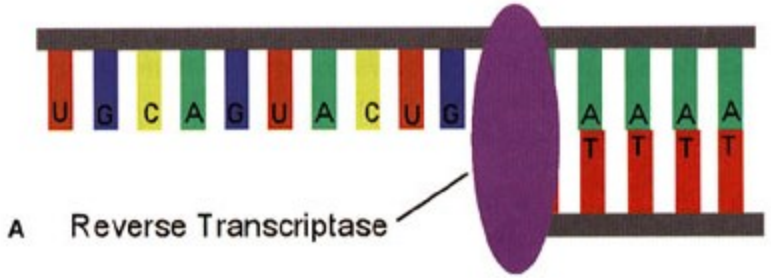
Reverse transcription polymerase chain reaction (RT-PCR) - is used to qualitatively detect gene expression through creation of complementary DNA (cDNA) transcripts from RNA.

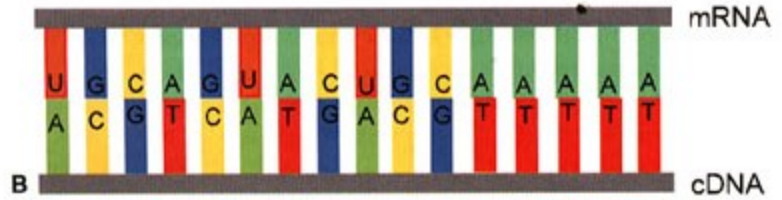
RT-PCR

RNA Template

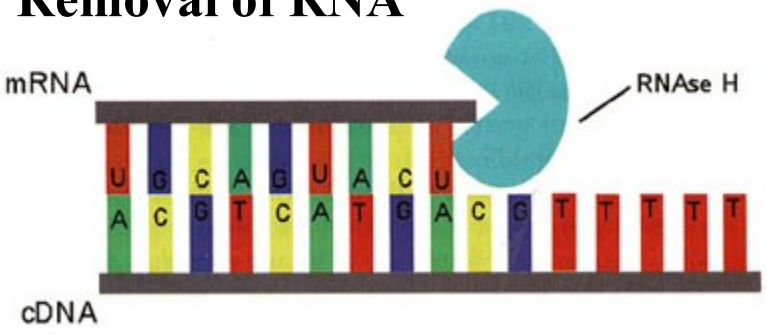
Priming for Reverse Transcription

First Strand Synthesis

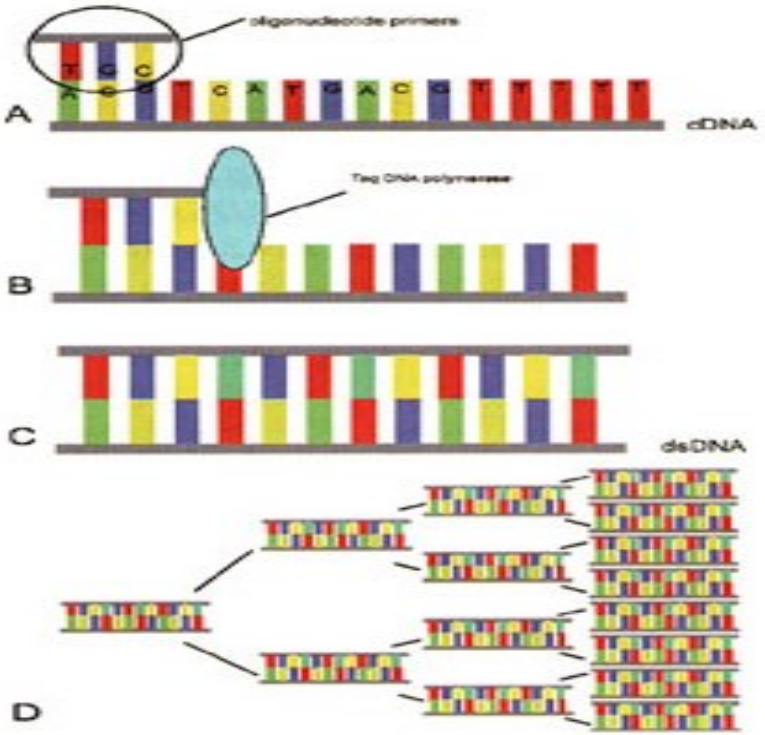




Removal of RNA



The PCR Reaction



COVID-19 Testing

This is how the coronavirus is detected in real time RT-PCR diagnostic test, by amplifying certain regions in the viral genome:

- **Targets include the RdRp (RNA-dependent RNA polymerase) gene ORF1ab and N (nucleoprotein).**

cDNA Sequencing

Can all protein-coding genes of an organism be identified by cDNA sequencing?

cDNA Sequencing

Can all protein-coding genes of an organism be identified by cDNA sequencing?

Difficulty with this approach - a gene of interest may be developmentally expressed or regulated in such a way that the mRNA is not present

This problem is circumvented by pooling mRNA from a variety of tissues & developing organs, or subjecting the organism to several environmental influences

Current gold standard for protein-coding gene annotation is EST or full-length cDNA sequencing followed by alignment to a reference genome.

EST – expressed sequence tag

EST Sequencing

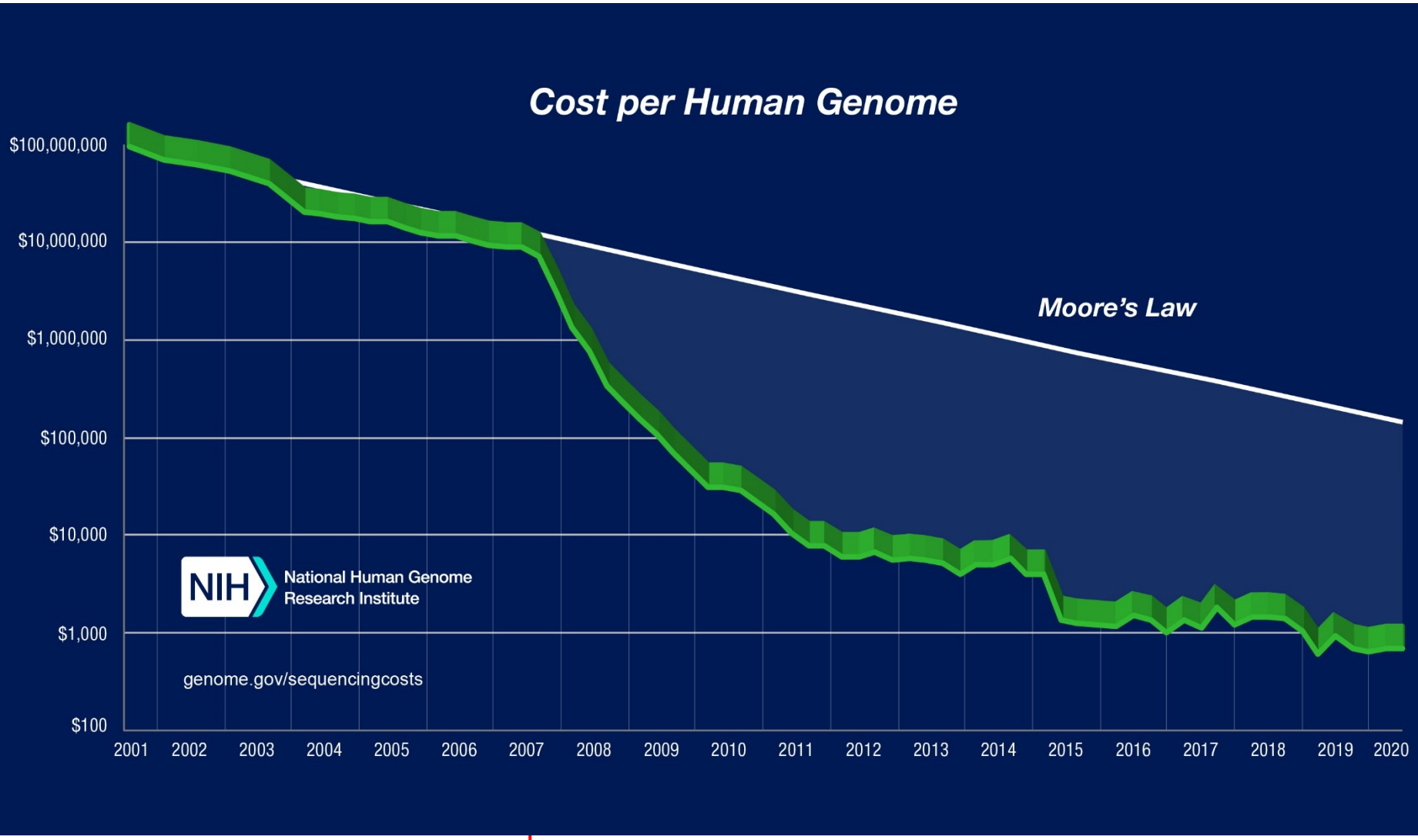
An important development in computational approaches was by Craig Venter - to prepare databases of partial sequences of expressed genes, called expressed sequence tags or ESTs.

- which are long enough to give a pretty good idea of the protein sequence.

To identify the function of the cloned gene, translated EST sequence can be compared to a database of protein sequences - to find its homologs with known function.

Corresponding cDNA clone of the gene of interest can then be obtained and the gene completely sequenced.

High-throughput / Next-Generation Sequencing



13yrs, $3 billion **S.S.** **8days, $10,000** **15min, <$1,000**

DNA sequencing beating Moore's law

HTS/NGS Sequencing

High-throughput sequencing (HTS) technologies have revolutionized the way biologists acquire and analyze genomic data.

- massively parallel sequencing

HTS instruments such as

- **454 from Roche Diagnostics,**
- **Illumina Genomic Analyzer,**
- **Applied Biosystems SOLiD System,**
- **Helico's Single-molecule sequencing platform**
- **MinION, Oxford Nanopore Technologies**

- can generate tens of gigabases per week, at a cost 200-fold less than previous methods, potentially enabling the routine sequencing of human and other genomes.

Sequencing Machines: Overview

	Roche GS FLX+	Illumina HiSeq 2000	SOLiD™ 4	Ion Torrent PGM
Bases per run	700Mb	600 Gb	100 GB	1 Gb
Time per run	23h	~11 days	~14 days	4.5 h
Reads per run	1 Million	6 Billion (paired-end) 3 Billion (single)	1.4 Billion	Millions
Read length	~700 bp	2 x 100 bases	2 x 50 bases	35–400 bases

Single-molecule sequencing technology - ≥1500bp

MinION – 10-100Kb read lengths, high error rates (~10-15%)

Sequencing Machines: Overview

Roche GS-FLX

1. Pyrosequencing

Life Technologies SOLiD

3. Sequence by ligation

Illumina HiSeq

2. Sequence by Synthesis

Life Technologies Ion Torrent

4. Proton Detection

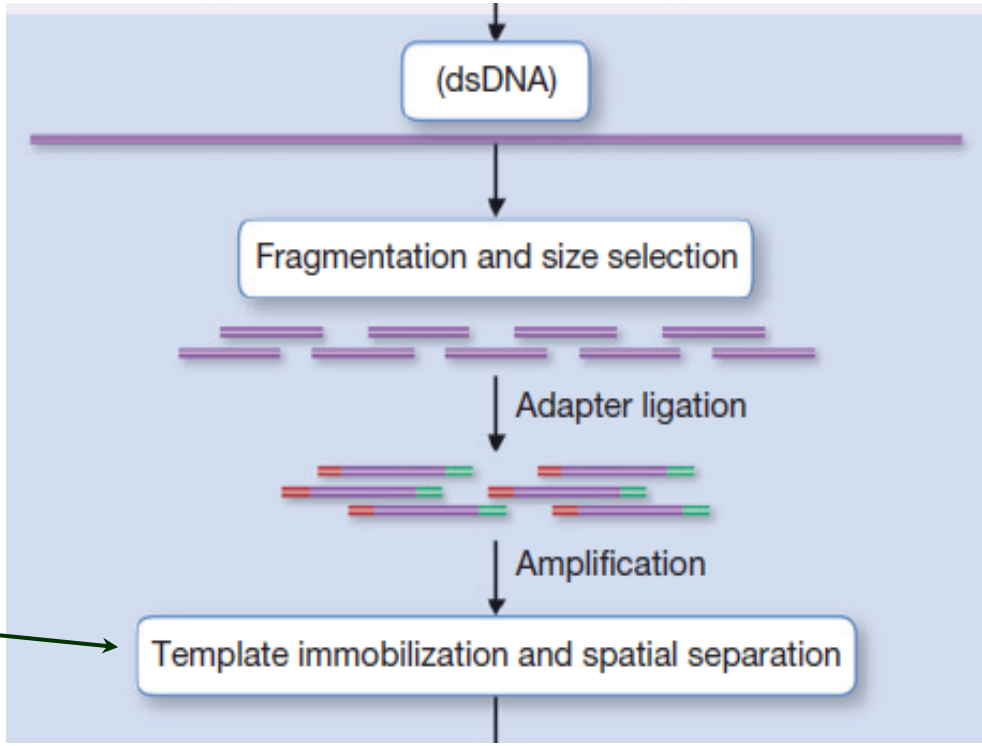
5. Nanopore sequencing

Basic workflow: Template Generation

Sequence library – convert starting material into a library of sequencing reaction templates.

Require common steps:

- **Fragmentation**
- **Size selection**
- **Adapter ligation**



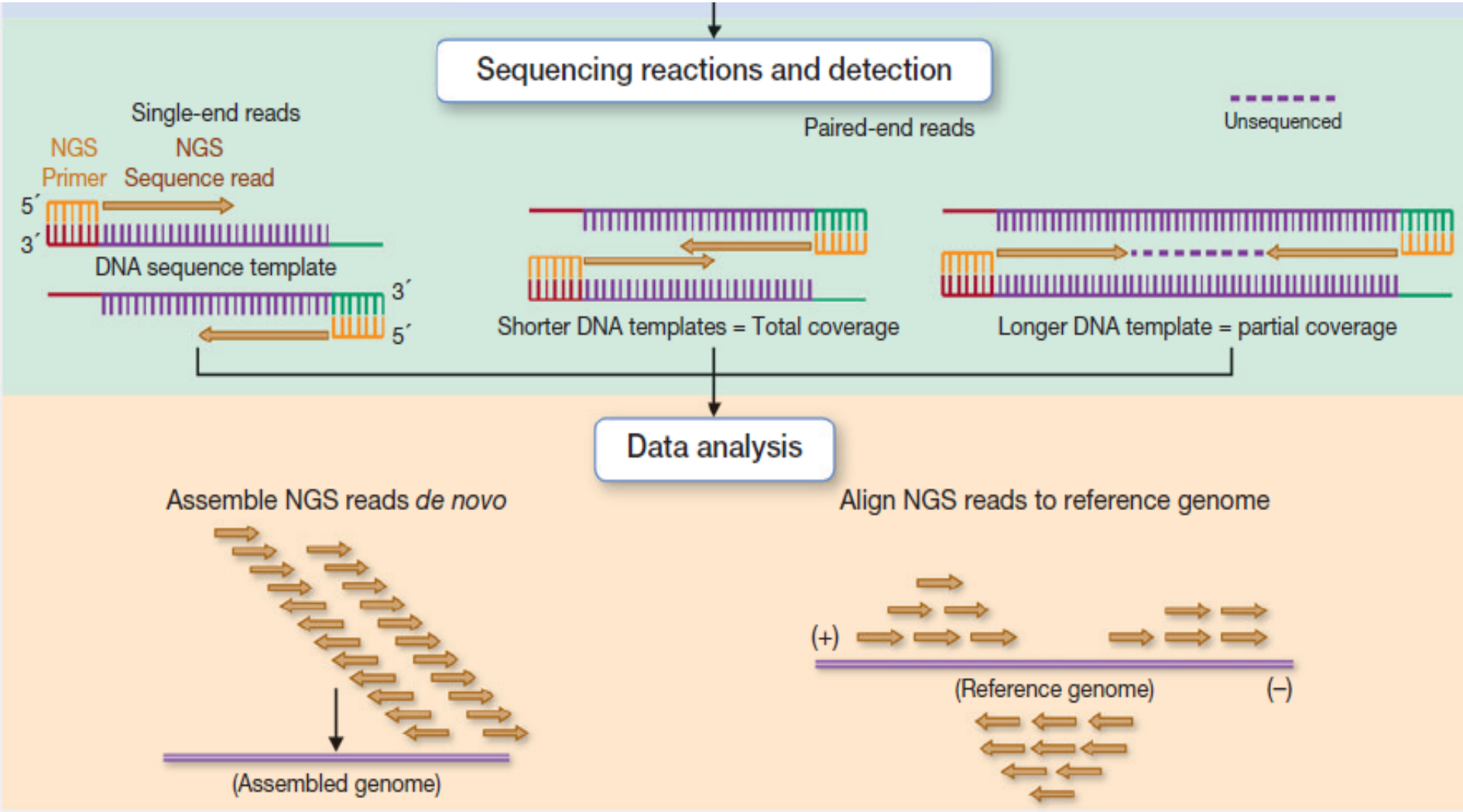
by attachment to solid surfaces or beads

Amplification-based - "second-generation" sequencing technology

Single-molecule - "third-generation" sequencing technology

A library is either sequenced directly - Single-molecule templates, or amplified then sequenced - Clonally amplified templates

Basic workflow: Detection & Data Analysis



Data Analysis

The scale and nature of data produced by all NGS platforms place substantial demands on IT at all stages of sequencing, including data tracking, storage, and quality control.

- read lengths: 50 – 100bp, No. of reads: ~ GBs

Initial analysis or base calling - by proprietary software on the sequencing platform.

After base calling, sequencing data are aligned to a reference genome if available or a *de novo* assembly is conducted.

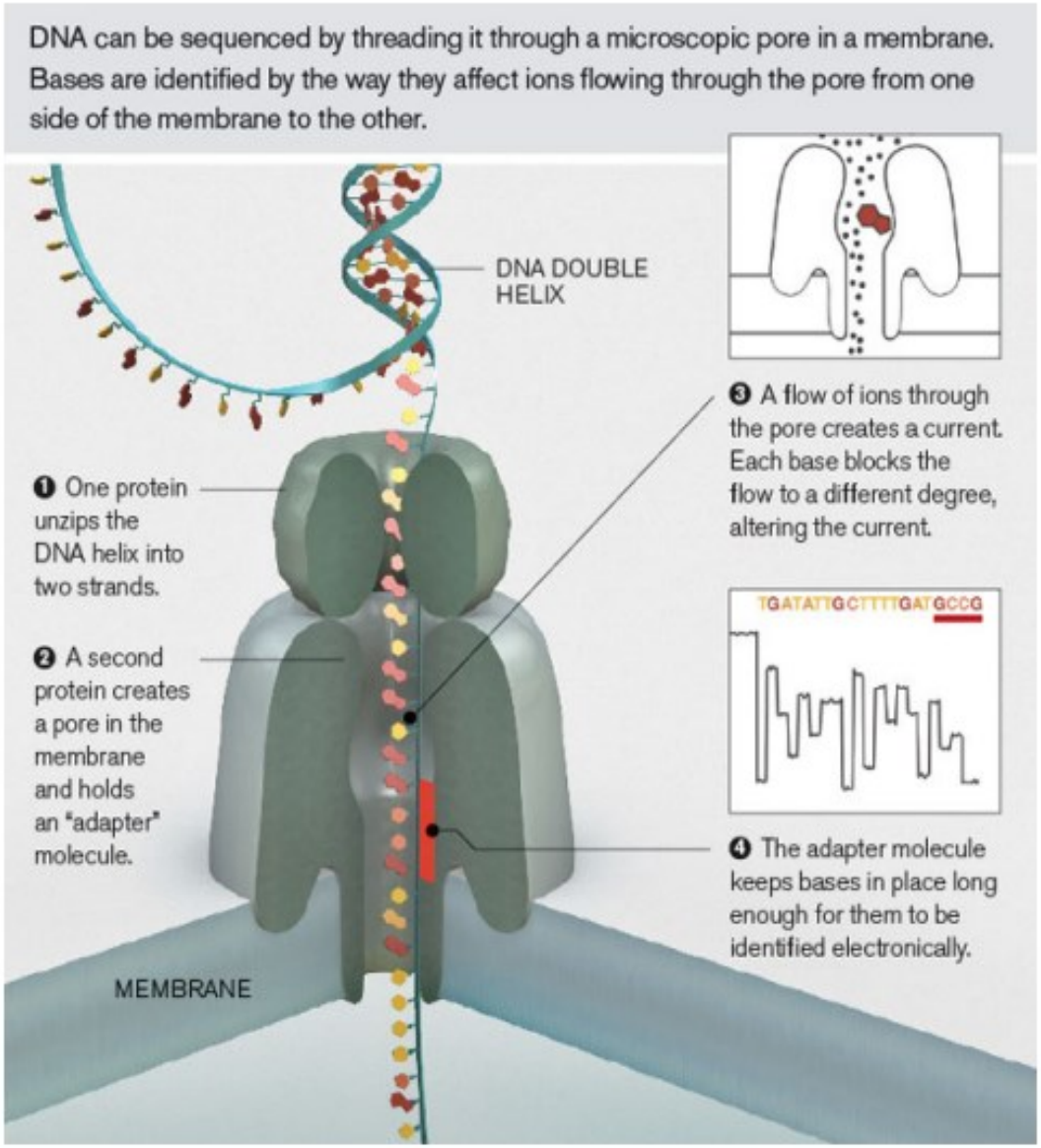
Once the sequence is aligned to a reference genome, the data needs to be analyzed in an experiment-specific fashion.

Sequence alignment & assembly is an active area of computational research

Third Generation Sequencing (TGS)

- **'Long read sequencing' – read length: ~ 10 – 60Kb**
- **Single molecule sequencing**
- **No PCR step involved**
- **Faster and portable**
- **Under active development**
- **e.g., PacBio Single molecule real time sequencing (SMRT) and Oxford Nanopore**

Oxford Nanopore - MinION



HTS Applications

genome	*de novo* sequencing: the initial generation of large eukaryotic genomes	Velasco et al., 2007 Diguistini et al., 2009 Huang et al., 2009 Li et al., 2010
	De novo, whole-genome and targeted sequencing	
	whole-genome resequencing: comprehensive SNP, indels, copy number and structual variations in individual human genomes	Bentley, 2006 Ossowski et al., 2008 Denver et al., 2009 Xia et al., 2009
	targeted resequencing: targeted polymorphism and mutation discovery	Hodges et al., 2007 Porreca et al., 2007 Harismendy et al., 2009
trancriptome	quantification of gene expression and alternative splicing; transcript annotation; discovery of transcribed SNPs or somatic mutations	Axtell et al., 2006 Sultan et al., 2008 Sugarbaker et al., 2008 Jacquier, 2009
	Deep sequencing of RNA transcripts	
	small RNA profiling	Berezikov et al., 2006 Houwing et al., 2007
epigenome	transcription factor with its direct targets	Johnson et al., 2007 Robertson et al., 2007
	Deep sequencing of DNA fragments pulled down by Chip-Seq	
	genomic profiles of histone modifications	Impey et al., 2004 Mikkelsen et al., 2007
	DNA methylation	Cokus et al., 2008 Costello et al., 2009
	Deep sequencing of bisulfite-treated DNA	
	genomic profiles of nucleosome positions	Fierer et. al., 2006 Johnson et al., 2006
metagenome	environmental	Edwards et al., 2007 Hubert et al., 2007
	Species classification by metagenomics & pangenomics	
	human microbiome	Turnbaugh et al., 2007 Qin et al., 2010

HTS Applications

One of the most prominent applications of NGS is re-sequencing:

- **whole genome resequencing**
- **target-region resequencing**
- **exome resequencing**

Any human individual's genome available in NCBI?

- genome-wide analysis of single nucleotide variations and other structural variations in multiple individuals, or strains, cancer sequencing, population-based sampling of a species, migration patterns of a virus, e.g., SARS-CoV-2, etc.

HTS Applications

RNA sequencing – has several applications, including RNA expression, *de novo* transcriptome sequencing for non-model organisms and novel transcript discovery

***viz.*, mRNAs, noncoding RNAs, small RNAs, miRNA**

For RNA and microRNA expression profiling, NGS has significant advantages compared to microarray methods in better quantification of common & rare transcripts.

Transcriptome - the complete set of transcripts in a cell, and their quantity, for a specific developmental stage or physiological condition.

NGS Applications

Epigenomic Analysis – NGS technologies have been applied in several epigenomic areas, *viz.*,

- **characterization of DNA methylation patterns,**
- **posttranslational modifications of histones,**
- **interaction between transcription factors and their direct targets, and**
- **nucleosome positioning on a genome-wide scale.**

Epigenetics is the study of heritable gene regulation that does not involve the DNA sequence itself but its modifications and higher-order structures.

HTS Applications

Metagenome Sequencing – sequencing the bacterial 16S rRNA gene across a number of species, for studying phylogeny and taxonomy, particularly in diverse metagenomic samples

e.g., cataloging human gut microbial genes by metagenomic sequencing (Qin *et al*, 2010).

~ 570Gb of sequence data from 124 individuals was generated, assembled and characterized 3.3 million non-redundant microbial genes.

This helped scientists, for the first time, to define the minimal human gut metagenome.

Metagenomics involves genomic analysis of microorganisms by direct extraction of DNA from uncultured ensemble of microbial communities

PCR Sequencing

How would you go about sequencing SARS-CoV-2 genome, 29903 bases long?

What technique is used for diagnostic testing of COVID-19?

While sequencing a novel genome for the first time, how are primers identified?

Can we now answer these Qs:

- **How is the SARS-CoV-2 genome sequenced?**
- **How does one identify the coordinates of N gene on it? i.e., how to construct a physical map of a genome?**
- **How does one select which regions in this gene would give specificity for the presence of SARS-CoV-2?***
- **How is the specific probe regions extracted and amplified for detection?**
- **Is it possible to store the DNA sample for re-testing? How?**

References:

1. **Concepts in Biotechnology, ed. D. Balasubramanyam**
2. **Restriction Endonucleases and DNA Modifying Enzymes http://arbl.cvmbs.colostate.edu/hbooks/genetics/biotech/enzymes/index.html**
3. **REBASE: restriction enzymes and methyltransferases, Nucleic Acids Research, Vol. 31 (1), 418–420 (2003)**

Genes

Special sequences in the DNA code for genes:

- **Protein-coding genes, for which the final product is a protein.**
  - **same gene may give rise to more than one protein (~ 6 per gene in humans).**
- **Non-coding RNA genes - for which the final product is RNA**

Genotype – An organism's genotype is the set of genes that it carries.

Phenotype – An organism's phenotype is all of its observable characteristics which are influenced both by its genotype and by the environment e.g., height, hair colour, levels of hormones, etc.

Differences in the genotypes can produce different phenotypes

Genes for ear form are different, causing one of the cats to have normal ears and the other to have curled ears

Change in environment can also affect the phenotype. Pinkness is <u>not</u> encoded in the genotype of flamingos - the food they eat makes their phenotype white or pink - a natural pink dye, canthaxanthin, obtained from their diet of brine shrimp and blue-green algae

Genes

The biological function of a gene is to preserve and express the genetic information encoded within it

Genes are normally very stable entities

Genetic stability is not absolute, however.

Genes may occasionally become altered; these changes called mutations create new alleles.

Mutant genes are also stable entities and are inherited in the same way as normal, wild-type genes.

Genes

Normal diploid cells such as somatic cells of humans contain two sets of genes – one set inherited from each parent.

- corresponding genes derived from each parent are called alleles.

Together the two alleles govern the phenotype of an organism.

What is the percentage of genes in a genome?

Genes

Gene-fraction varies from ~70% in prokaryotes to ~ 2 - 3% in humans

- **does this imply prokaryotes have more gene content than eukaryotes?**

- **Size of a prokaryotic genome? Eukaryotic genome?**

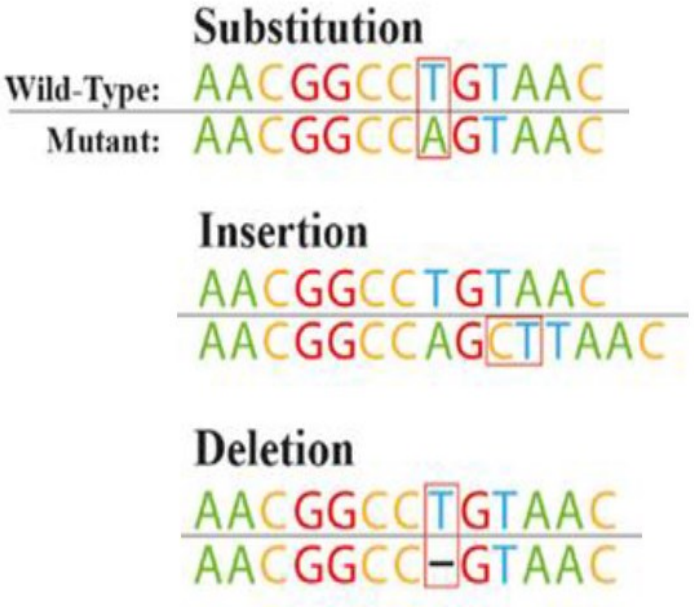
What's the function of remaining ~97-98% of human genome?

The remaining part of the genome consists of noncoding regions, whose functions may include providing chromosomal structural integrity and regulating where, when, and in what quantity proteins are made.

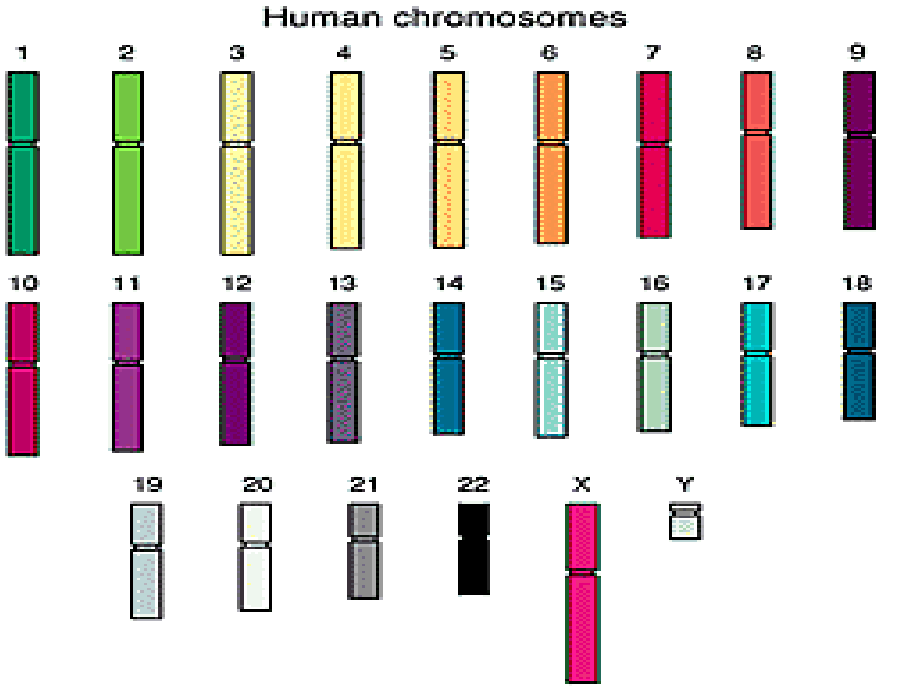
Mutations

Mutations - are local changes in the DNA content, caused by inexact replication and are of various kinds:

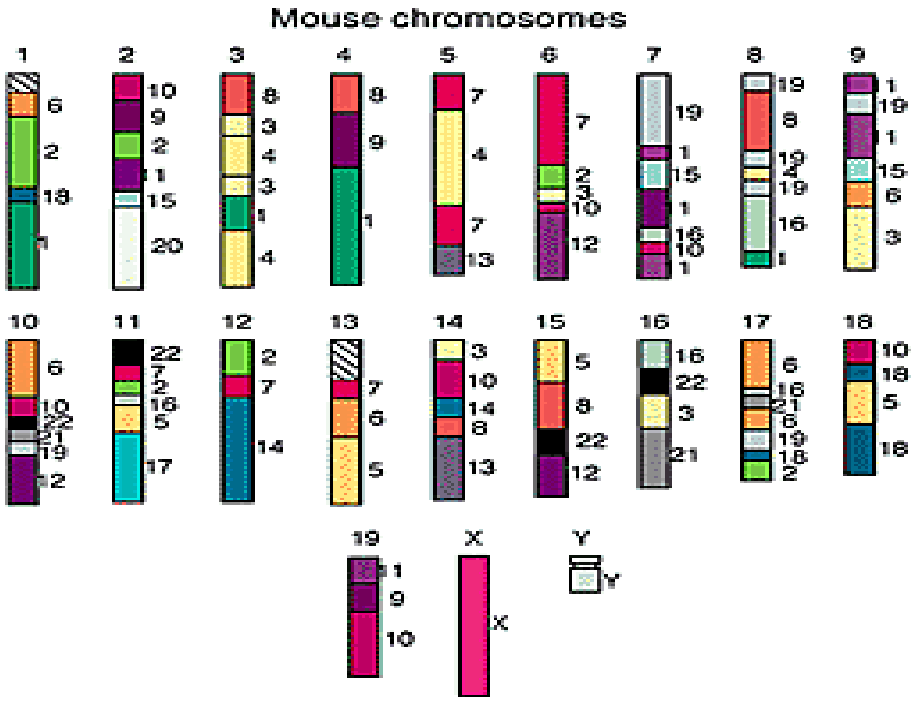
- **Substitution - a base is replaced by another - may or may not alter the protein sequence depending on the place it occurs.**
- **Insertion/Deletion – addition/removal of one or more bases – results in a frame-shift in coding regions.**
- **Rearrangement - a change in the order of complete segments along a chromosome.**



Chromosomal rearrangements occur both within and between chromosomes during evolution



The colors on the mouse chromosomes and the numbers alongside indicate the human chromosomes containing homologous segments.



Mutations

Role of Mutations:

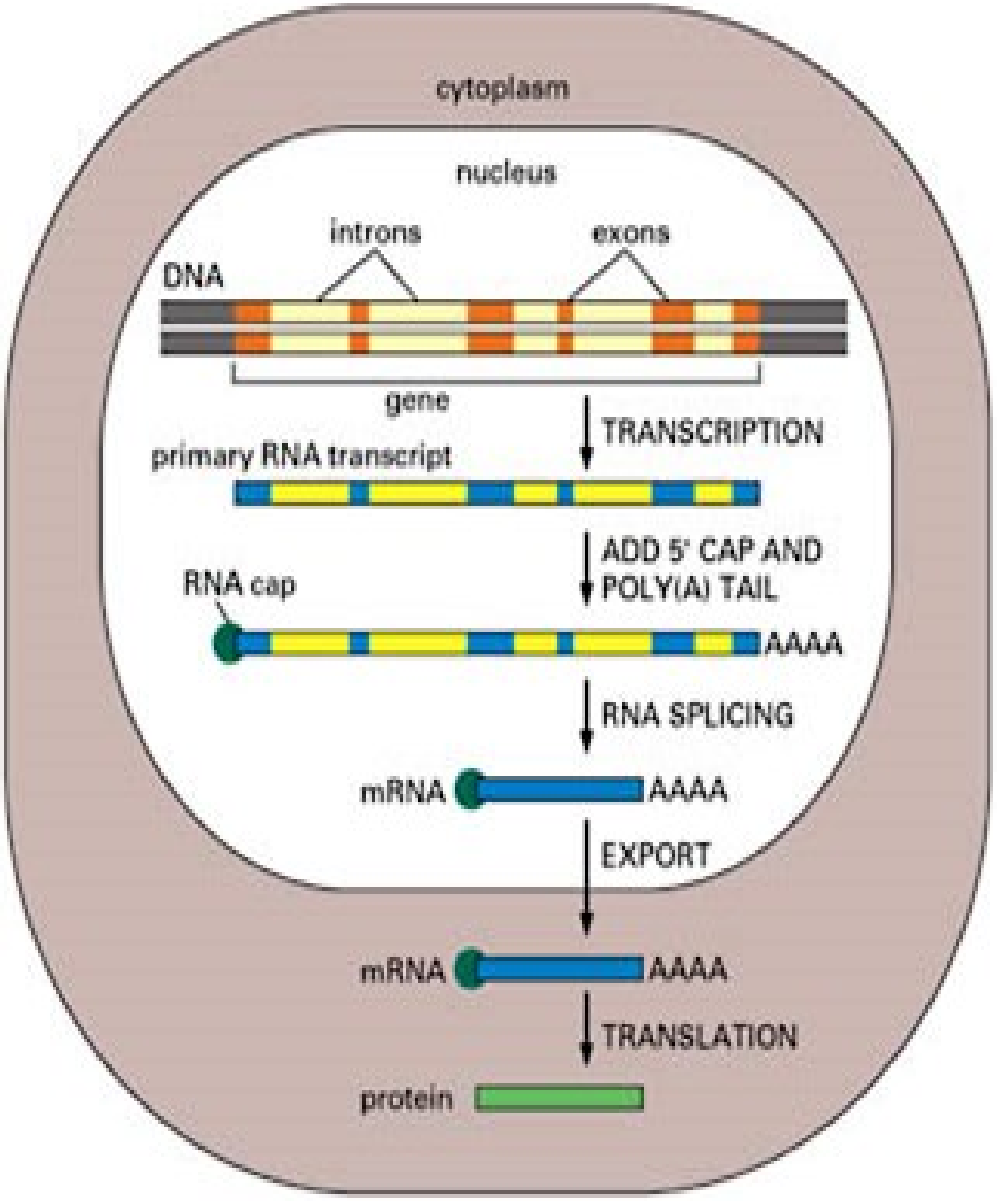
- **Mutations are the source of phenotypic variation on which natural selection acts, creating species & changing them.**

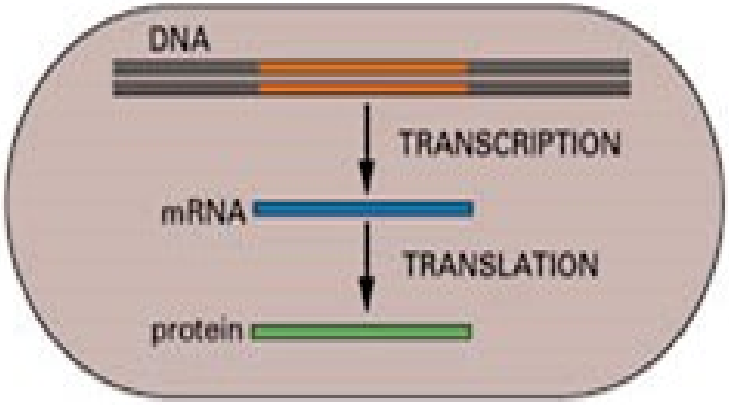
  e.g., the human and mouse genome are very similar – major difference being the internal order of DNA segments.

  Without mutations there wouldn't be any evolution!

- **They are responsible for inherited disorders and diseases, which involve alterations in gene.**

Steps Leading from Gene to Protein

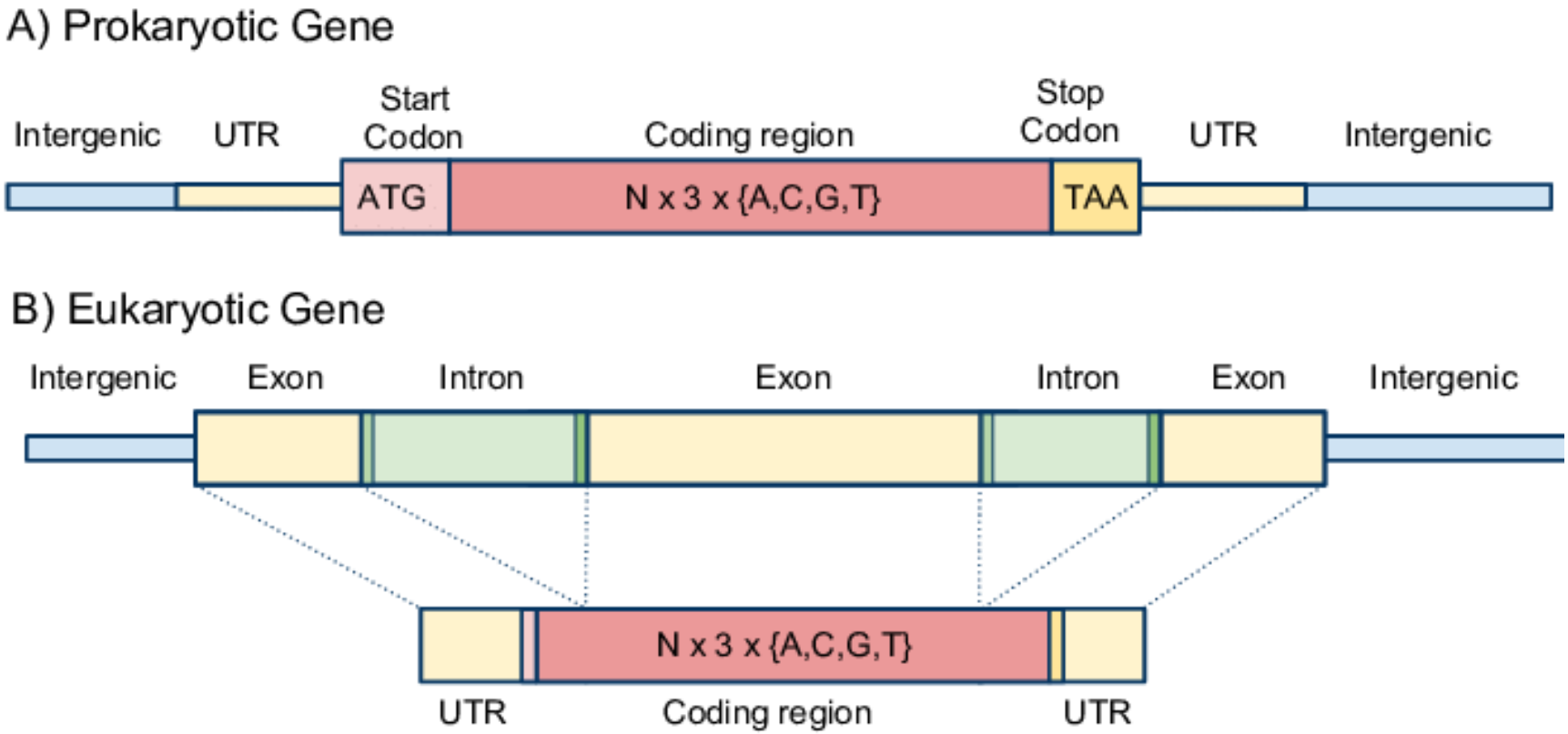




Prokaryotes

Eukaryotes

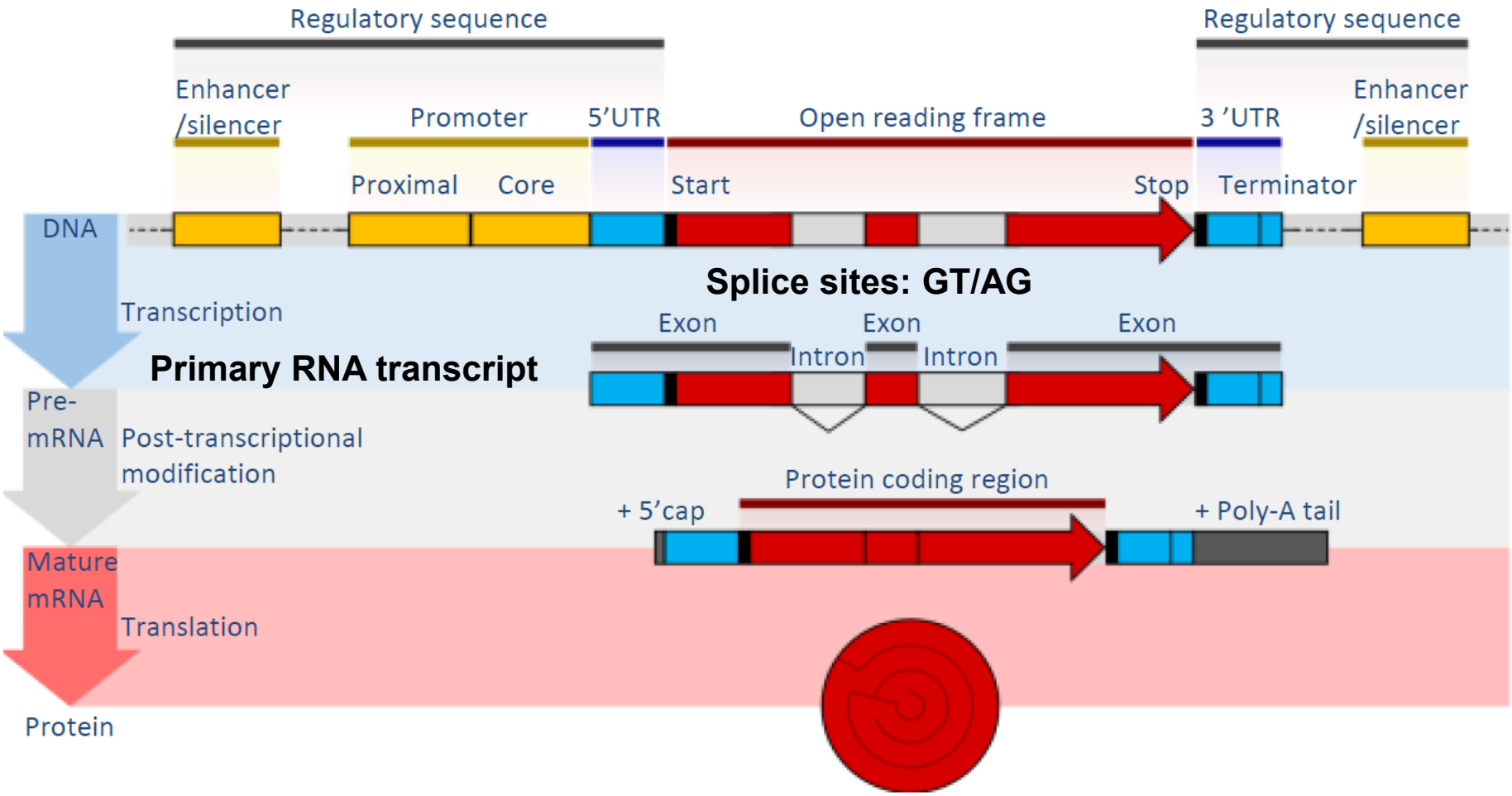
Gene Structure



UTRs – Untranslated Regions – are transcribed, but not translated

Eukaryote Gene Structure

Start codon: ATG **Stop Codon: TAA/TAG/TGA**



Transcription is initiated only at certain specific positions in the sequence, signaling the beginning of genes, called promoters.

Alternative Splicing

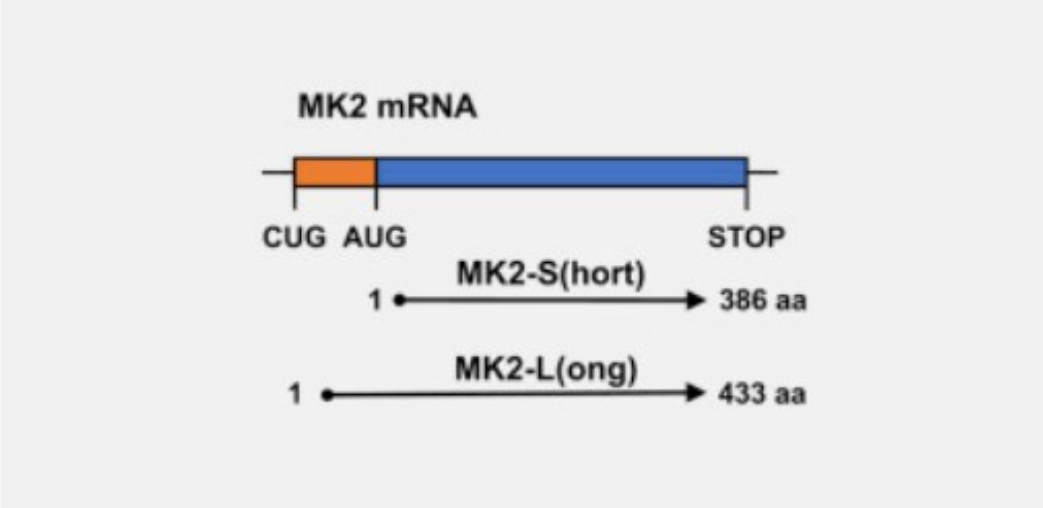
- **In many cases, the pattern of splicing can vary depending on the tissue in which the transcription occurs.**

  e.g., an exon maybe spliced in the gene transcribed in liver but retained when transcribed in the brain.

- **This variation called alternative splicing, contributes to the overall protein diversity in the organism**

Alternative Initiation

- Another type of variation that contributes to protein diversity is alternative initiation



- Alternative translation is an important mechanism of post-transcriptional gene regulation leading to the expression of different protein isoforms originating from the same mRNA

Data Representation

DNA - a complex, dynamic, three-dimensional molecule represented as a string of alphabets

- a perfect representation for computer analysis

Aim: to find grammar & syntax rules of DNA language based on this 4-letter alphabet

- similar to English Grammar to form meaningful sentences

Biological Sequence Analysis

Pattern Recognition:

Assumption in biological sequence analysis:

- strings carrying information will be different from random strings

If a hidden pattern can be identified in a string, it must be carrying some functional information

Biological Sequence Analysis

Order of occurrence of bases:

not completely random

- Different regions of the genome exhibit different patterns of the four bases, A, T, G, C

e.g., protein coding regions, regulatory regions, intron/exon boundaries, repeat regions, etc.

Aim: Identify various patterns to infer their functional roles

Example

❄♒♓♦ ♓♦ ♋ ●♏♍♦◆❒♏ □■
✍♓□♓■♐□❒○♋♦♓♍♦

♋♦♐♎ ●&♐♐● ♐♎♐♎ ♦♐♐♦⛝♏
■❖♍❒□♦ ■⌘♍♎♐♒♦□◆

This is a lecture on bioinformatics

asjd lkjfl jdjd sjftye nvcrow nzcdjhspu

Frequency of letters

A.	7.3%	N.	7.8%
B.	0.9%	O.	7.4%
C.	3.0%	P.	2.7%
D.	4.4%	Q.	0.3%
E.	13.0%	R.	7.7%
F.	2.8%	S.	6.3%
G.	1.6%	T.	9.3%
H.	3.5%	U.	2.7%
I.	7.4%	V.	1.3%
J.	0.2%	W.	1.6%
K.	0.3%	X.	0.5%
L.	3.5%	Y.	1.9%
M.	2.5%	Z.	0.1%

Other statistics

Frequencies of the most common first letter of a word, last letter of a word, doublets, triplets, etc.

20 most used words in written English

- **the of to in and a for was is that on at he with by be it an as his**

20 most used words in spoken English

- **the and I to of a you that in it is yes was this but on well he have for**

Parallels in DNA language

ATGGTGGTCATGGCGCCCCGAACCCTCTTCCTGCTG
CTCTCGGGGGCCCTGACCCTGACCGAGACCTGGGCG
GGTGAGTGCGGGGTCAGGAGGGAAACAGCCCCTGC
GCGGAGGAGGGAGGGGCCGGCCCGGCGGG

GTCTCAACCCCTCCTCGCCCCCAGGCTCCCACTCCA
TGAGGTATTTCAGCGCCGCCGTGTCCCGGCCCGGCC
GCGGGGAGCCCCGCTTCATCGCCATGGGCTACGTGG
ACGACACGCAGTTCGTGCGGTTC

Parallels in DNA language

ATG GTG GTC ATG GCG CCC CGA ACC CTC TTC
CTG CTG CTC TCG GGG GCC CTG ACC CTG ACC
GAG ACC TGG GCG GGT GAG TGC GGG GTC AGG
AGG GAA ACA GCC CCT GCG CGG AGG AGG GAG
GGG CCG GCC CGG CGG…

GTC TCA ACC CCT CCT CGC CCC CAG GCT CCC ACT
CCA TGA GGT ATT TCA GCG CCG CCG TGT CCC
GGC CCG GCC GCG GGG AGC CCC GCT TCA TCG
CCA TGG GCT ACG TGG ACG ACA CGC AGT TCG
TGC GGT TC…

1st exon and 1st intron of Human HLA gene

This task needs to be automated because of the large genome sizes:

Smallest genome:

Mycoplasma genitalium 0.5×10^6 bp

Human genome: 3×10^9 bp – not the largest!

~ 10-100 times the Britannica Encyclopedia

Plant genomes are even larger.

DNA Sequence Analysis

- **Evolution has operated on every sequence that we see today - genes and sequences involved in gene regulation are conserved.**
- **these are transferred, like code modules, from one organism to another. Because of evolution, similar sequences have similar functions.**
- **Algorithms for comparing sequences and finding similar regions are at the heart of computational biology.**

Syllabus

Unit 1: Overview – Bioinformatics, Gene & Genome structure

Gene Technology – Restriction Endonucleases, Cloning vectors

DNA sequencing – PCR, cDNA and Whole Genome sequencing, NGS and third generation sequencing technologies

Unit 2: BioDatabases

- **Major Bioinformatics Resources – NCBI, EBI, PubMed,**
- **Primary Nucleotide and Proteins Databases - GenBank, UniProt, PDB,**
- **Genome Browsers – Ensembl, UCSC**
- **k-mer analysis and their significance in biological sequences**

Syllabus

Unit 3: Sequence Alignment:

- **Pairwise Alignment – Types of pairwise alignments – Global, Local and Overlap alignments, Dot Plots, dynamic programming (DP) algorithm,**
- **Scoring matrices for nucleotides and proteins and gap penalties,**
- **Sequence-based Database Search algorithms – BLAST, FASTA,**
- **Multiple Alignment, Algorithms for Global and Local MSA – DP, Progressive based (ClustalX), Iterative methods, Motif search-based methods**

Syllabus

Unit 4: Modeling Molecular Evolution – Phylogeny:

- **Markov models of base substitution,**
- **Computing Phylogenetic Distances,**
- **Phylogenetic Tree Construction Methods, PHYLIP**

Unit 5: Gene Prediction:

Gene Prediction approaches –

- **Open Reading Frames,**
- **Homology search,**
- **Content-based methods,**
- **Markov models**

1

The story in DNA

Or: what kind of information can I get from DNA?

"In nature's infinite book of secrecy
a little I can read."

William Shakespeare (1623) *Anthony and Cleopatra*

What this chapter is about

This chapter gives an overview of the kind of information that can be gained from analysing DNA sequences, including identifying individuals, unravelling social interactions, understanding the evolution of major adaptations, tracing the evolutionary origins of lineages, and investigating the tempo and mode of evolution over all time scales. By taking a broad look at the use of DNA sequences in evolution and ecology, we will set the scene for later chapters; topics only briefly mentioned here will be covered in more detail later in the book.

Key concepts

Evolutionary biology: evolution connects individual lives to population-level processes to large-scale evolutionary change.

Molecular evolution: information on all of these levels is available in DNA.

Techniques: DNA sequencing

The mystery of the Chilean blob

In July 2003, a 13-tonne blob washed up on a Chilean beach (Figure 1.1). With no bones, no skin, not even any cells, there was nothing obvious to identify the origin of the mystery blob. Theories abounded: it was a giant squid, a new species of octopus, an unknown monster from the deep, an alien. This was not the first appearance of a giant 'globster'. In 1896, an 18-metre blob washed up on St Augustine beach in Florida, USA. The identification of the blob varied from giant squid to whale blubber. However, when it was formally described in the scientific literature (albeit sight unseen), it was given the scientific name *Octopus giganteus*. There have been at least half a dozen other reported globsters, including the 1960 Tasmanian West Coast Monster, 9 m long and 2.5 m tall, which, although badly decomposed, was described as being hairy.

The mystery of the Chilean blob was solved in 2004 when researchers sequenced DNA samples from the globster, and compared them to DNA sequences held in the gigantic public database GenBank (see **TechBox 1.1**). The sequences matched that of a sperm whale (scientific name *Physeter catadon*: Figure 1.2, p. 7), The globster was nothing but blubber (see **Case Study 1.1**, p. 5). The team also retrospectively solved previous sea monster mysteries. For example, DNA extracted from samples of the 'Nantucket Blob' of 1996 showed that it had been the remains of a fin whale (*Baleonoptera physalus*). As reported on Unexplained-Mysteries.com: 'One of the myths of the sea has been skewered by gene researchers'.

Figure 1.1 The blob that washed up on a Chilean beach had people guessing: squid, octopus, kraken? Similar blobs have been found on beaches all over the world.

Image courtesy of Elsa Cabrera/ CCC.

The DNA sample taken from the Chilean blob was enough to unambiguously identify it as the remains of a sperm whale. But that DNA sample could do far more. From that one tiny sample, we could identify not only the species it came from, but also, given enough data, which individual whale. We could use that DNA sample to predict where that individual whale was born, and to understand the relationships it had during its lifetime, both with its immediate family and members of its social group. That sample of DNA could allow us to track whale movements across the globe, both in space and in time. We could use it to trace this whale's family history back through the ages, exploring how the whales responded to a changing world as ice ages came and went. The DNA sample could help us reconstruct the evolution of important whale adaptations such as echolocation, and identify the nearest mammalian relatives of the whales, a question that has perplexed biologists for centuries. By comparing this whale DNA to the DNA of other mammals, we could ask whether the rise of modern mammals was contingent on the extinction of the dinosaurs. Deeper still, this DNA gives the chance to look beyond the fossil record, potentially shedding light on one of the greatest biological mysteries, the explosive beginnings of the animals, and back further still to the origin of the kingdoms.

The reason the DNA sequences from the genomes of different species can provide information at all of these different levels of evolutionary history – from individuals to populations to species, families, phyla, and kingdoms of life – is that different parts of the genome evolve at

TECHBOX 1.1

GenBank

KEYWORDS

Entrez

database

NCBI

accession

annotation

FURTHER INFORMATION

The first chapter of the NCBI handbook (freely available online) explains their databases, including GenBank (*www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=handbook*).

RELATED TECHBOXES

TB 6.1: Taxonomy

TB 3.4: BLAST

RELATED CASE STUDIES

CS 1.1: Chilean blob (identifying species)

CS 6.1: Barcoding nematodes (DNA taxonomy)

GenBank is a golden example of the international science community sharing data freely. It is based at the National Center for Biotechnology Information (NCBI) in the US, but is synchronized with European (EMBL) and Japanese (DDBJ) molecular databases so that they all share the same data. GenBank contains most of the DNA sequences that have ever been produced. Whenever a scientist sequences a section of DNA, they should submit the sequence to GenBank so that anyone else can access the sequence and use it in their own research. Submission to GenBank is usually a requirement of publication in the scientific press, in line with the ethos of repeatability of scientific experiments (any scientist should have access to the data and materials needed to check published results). In addition, there are many sequences on GenBank that have never been formally published, but are available for anyone to use. At the time of writing, GenBank contains over 77 billion nucleotides of DNA sequences, from over 200,000 species.

When a researcher submits a DNA sequence, they provide information about the organism it was sampled from, what kind of sequence it is, and other features of the sequence (this information is broadly known as sequence annotation). For example, if the sequence contains part of a protein-coding gene, the information given might include the location of the beginning and end of the coding sequence, the amino acid sequence of the protein product generated from it and the likely role of the protein product. Genome sequencing projects often rely on automated annotation to identify and label features such as coding regions or gene regulation elements.

Each GenBank submission is assigned a unique accession (identification) number that can be used to retrieve that sequence from the database. Sequences can also be accessed by searching for an organism, gene, author name, or key word. For example, if you type 'Chilean blob' into the GenBank search engine, Entrez, you retrieve three sequences from the study described in **Case Study 1.1** with the accession numbers AY582746, AY582747, and AY582748. If you type these accession numbers into query box on the Entrez search engine, it will locate the records of these sequences in the database so that you can access them. **Figure TB1.1** shows the GenBank entry you will retrieve if you enter the accession number AY582746.

Accepting submissions from any individual or laboratory is one of the strengths of GenBank, allowing it to rapidly expand to cover more species and more genes. But it is also a weakness, as it is difficult to guarantee submission quality. It is inevitable that some sequences contain mistakes made in the sequencing process. Worse, some sequences may represent contaminants, rather than the target sequence reported. Sometimes there are errors in annotation (sequences may be listed as the wrong gene or from an incorrect species). So while sequences from GenBank are a boon to biological research, they should not be used uncritically: check your own analyses for aberrant results that could be caused by a mistake in GenBank, and look for alternative sequences if you suspect any problems.

GenBank can be accessed using any internet browser. The easiest approach is to go to the Entrez search engine (www.ncbi.nlm.nih.gov/gquery/) which allows you to search a large

collection of databases at once. In addition to the DNA sequence database GenBank, the search engine Entrez allows you to search a range of other databases, including whole genome sequences, single nucleotide polymorphisms (SNPs: **TechBox 3.3**), population-based datasets, functional and structural information on gene products, and taxonomic information.

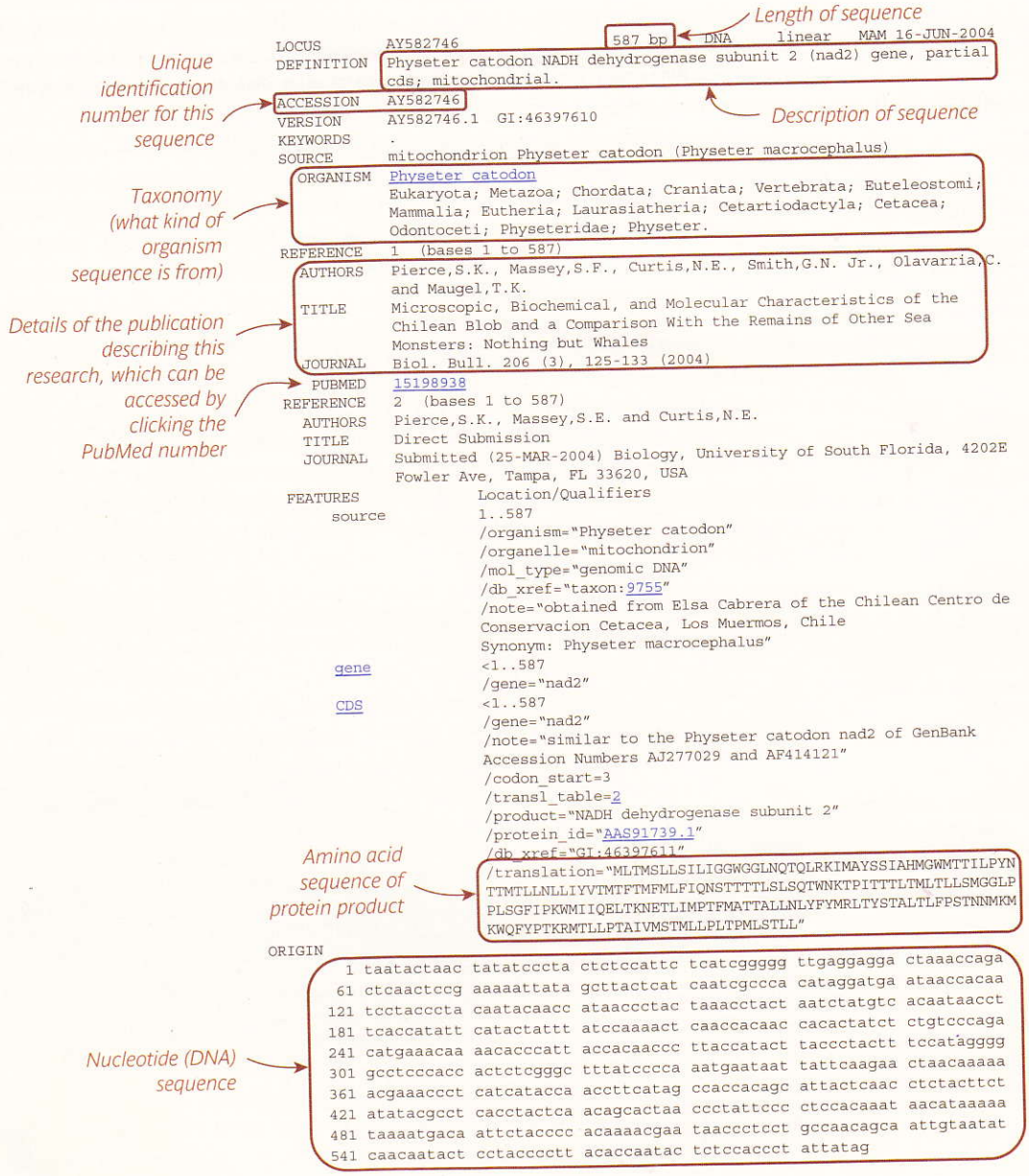


Figure TB1.1 The GenBank entry for a sequence from the Chilean Blob described in **Case Study 1.1**. Important features of the GenBank entry have been labelled in red: all GenBank entries should have the same basic layout.

CASE STUDY 1.1

Solving the mystery of the Chilean blob: identifying species using DNA sequences

KEYWORDS

GenBank
sample preservation
ethanol
formaldehyde
contamination
mitochondrial
DNA barcoding
cryptozoology

RELATED TECHBOXES

TB 1.1: GenBank
TB 6.3: Multiple alignment

RELATED CASE STUDIES

CS 2.1: Origin of faeces (identifying species from remote samples)
CS 2.2: More moa (identifying species from ancient DNA)

41205

UCT-CCMB LIBRARY HYDERABAD 500 007

Pierce, S.K., Massey, S.E., Curtis, N.E., Smith, G.N., Olavarri, C. and Maugel, T.K. (2004) Microscopic, biochemical, and molecular characteristics of the Chilean Blob and a comparison with the remains of other sea monsters: nothing but whales. *Biological Bulletin*, Volume 206, pages 125–133

> *". . . to our disappointment, we have not found any evidence that any of the blobs are the remains of gigantic octopods, or sea monsters of unknown species."* [1]

Background

Strange gelatinous material occasionally washes up on beaches around the world. The large mass that washed up on a beach in Florida in 1896 was too heavy to be removed using a team of horses and so rubbery that axes bounced off it. The material generally lacks cells that would aid its identification, and so these 'globsters' have variously been described as the remains of giant squid, unknown species of octopus, or whale blubber. The lack of a clear species identification has resulted in globsters being listed on websites that discuss unexplained phenomena: the blob that washed up on a beach in Chile in 2003 made international headlines and was discussed by biologists and conspiracy theorists all over the world.

Aim

DNA sequences provide the means to unambiguously identify a biological sample, through comparison of the sequence to that from known species. Even if the sample is from a previously undescribed species, then comparing the sequence to a database of known sequences can show which species is the closest relative (for example, if a globster is a previously unknown species of giant octopus, then DNA sequences from the sample will be more similar to sequences from other octopuses than they are to those of squid or whales). These researchers aimed to use DNA from the Chilean blob to work out what kind of organism the blob had been derived from.

Methods

Small pieces of the blob were frozen, or preserved in ethanol: both procedures protect the DNA in the sample from decaying. DNA is everywhere, so contamination in sequencing labs can be a problem (Chapter 4). These researchers reduced the chance of their sequences being contaminants by analysing two different samples in independent laboratories, to check that they get the same result (this can rule out incidental contamination from the lab, but not contamination of the source of the samples). One lab in the US analysed a frozen sample of the blob, and a lab in New Zealand analysed an ethanol-preserved sample. DNA was extracted from the samples and sequenced. The US team sequenced part of mitochondrial gene called NADH2, which codes for the subunit of a fundamental metabolic enzyme. The NZ team sequenced the mitochondrial control region. These sequences were chosen because they tend to vary between species, and because they are relatively easy to

sequence, so there are a lot of comparable sequences in the international sequence database, GenBank (**TechBox 1.1**).

Results

To identify the origin of the blob, the DNA sequences were compared to existing sequences in GenBank. The sequences of NADH2 from the blob were identical to two NADH2 sequences in the database, which had both been taken from sperm whales (*Physeter catadon*, also known by the species name *Physeter macrocephalus*: **Figure 1.2**). Because this sequence normally varies between species, this is convincing evidence that the blob was from a sperm whale. The mitochondrial control region sequence from the Chilean blob was 99% identical to *P. catadon* sequences in GenBank. The control region tends to evolve faster than NADH2, and may vary between individuals within a species, so it is often used for population-level studies. Although the control region sequence differed from the sperm whale sequence on GenBank, it was only as different as you would expect from two different individuals in the same species.

	1	60
Physeter catadon	TAATACTAACTATATCCCTACTCTCCATTCTCATCGGGGGTTGAGGAGGACTAAACCAGA	
Chilean blob	TAATACTAACTATATCCCTACTGTCCATTCTCATCGGGGGTTGAGGAGGACTAAACCAGA	
	61	120
Physeter catadon	CTCAACTCCGAAAAATTATAGCTTACTCATCAATCGCCCACATAGGATGAATAACCACAA	
Chilean blob	CTCAACTCCGAAAAATTATAGCTTACTCATCAATCGCCCACATAGGATGAATAACCACAA	
	121	180
Physeter catadon	TCCTACCCTACAATACAACCATAACCCTACTAAACCTACTAATCTATGTCACAATAACCT	
Chilean blob	TCCTACCCTACAATACAACCATAACCCTACTAAACCTACTAATCTATGTCACAATAACCT	
	181	240
Physeter catadon	TCACCATATTCATACTATTTATCCAAAACTCAACCACAACCACACTATCTCTGTCCCAGA	
Chilean blob	TCACCATATTCATACTATTTATCCAAAACTCAACCACAACCACACTATCTCTGTCCCAGA	
	241	300
Physeter catadon	CATGAAACAAAACACCCATTACCACAACCCTTACCATACTTACCCTACTTTCCATAGGGG	
Chilean blob	CATGAAACAAAACACCCATTACCACAACCCTTACCATACTTACCCTACTTTCCATAGGGG	
	301	360
Physeter catadon	GCCTCCCACCACTCTCGGGCTTTATCCCCAAATGAATAATTATTCAAGAACTAACAAAAA	
Chilean blob	GCCTCCCACCACTCTCGGGCTTTATCCCCAAATGAATAATTATTCAAGAACTAACAAAAA	
	361	420
Physeter catadon	ACGAAACCCTCATCATACCAACCTTCATAGCCACCACAGCATTACTCAACCTCTACTTCT	
Chilean blob	ACGAAACCCTCATCATACCAACCTTCATAGCCACCACAGCATTACTCAACCTCTACTTCT	
	421	480
Physeter catadon	ATATACGCCTCACCTACTCAACAGCACTAACCCTATTCCCCTCCACAAATAACATAAAAA	
Chilean blob	ATATACGCCTCACCTACTCAACAGCACTAACCCTATTCCCCTCCACAAATAACATAAAAA	
	481	540
Physeter catadon	TAAAATGACAATTCTACCCCACAAAACGAATAACCCTCCTGCCAACAGCAATTGTAATAT	
Chilean blob	TAAAATGACAATTCTACCCCACAAAACGAATAACCCTCCTGCCAACAGCAATTGTAATAT	
	541	587
Physeter catadon	CAACAATACTCCTACCCCTTACACCAATACTCTCCACCCTATTATAG	
Chilean blob	CAACAATACTCCTACCCCTTACACCAATACTCTCCACCCTATTATAG	

Figure CS1.1 Part of an alignment of a sequence from the mysterious Chilean blob with that of a sperm whale (*Physeter catadon*), showing that the sequences are identical for this part of their mitochondrial genome.

Conclusions

The identical (or near-identical) match between the sequences from the blob and whale sequences in GenBank prove beyond doubt that the blob is the remains of a sperm whale.

Limitations

Selection of an appropriate gene to sequence is critical – there has to be enough difference between species to allow discrimination, yet not so much difference that the relationship to other species is unclear (see Chapter 6). Sample quality and nature of preservation are important for DNA extraction. They could not extract DNA from earlier 'blob' samples that had been preserved in formaldehyde (including samples of the relatively recent Tasmanian West Coast Monster), because formaldehyde destroys DNA. This is a shame, because many museum specimens are pickled in formaldehyde.

Future work

Because DNA can be extracted from samples such as hair, faeces, or saliva, molecular analysis has the potential to solve many mysteries of cryptozoology (the field that seeks to establish whether apparently mythical creatures are in fact real organisms). For example, DNA from purported Yeti hair samples was surprisingly similar to DNA sequences from horses[2]. The DNA barcoding movement seeks to catalogue DNA sequences of all species so that any unknown sample can be identified.

References

1. Pierce, S.K., Massey, S.E., Curtis, N.E., Smith, G.N., Olavarri, C. and Maugel, T.K. (2004) Microscopic, biochemical, and molecular characteristics of the Chilean Blob and a comparison with the remains of other sea monsters: nothing but whales. *Biological Bulletin*, Volume 206, pages 125–133.
2. Milinkovitch, M.C., Caccone, A. and Amato, G. (2004) Molecular phylogenetic analyses indicate extensive morphological convergence between the 'yeti' and primates. *Molecular Phylogenetics and Evolution*, Volume 31, pages 1–3.

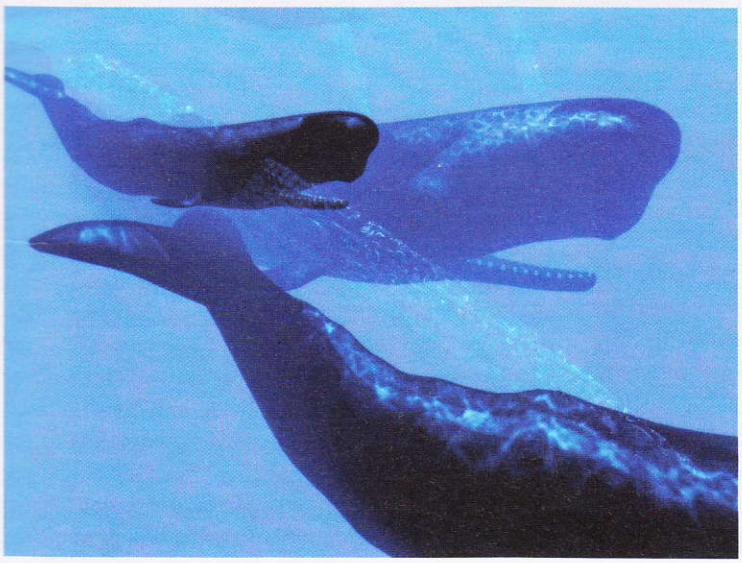


Figure 1.2 Sperm whales (*Physeter catadon*, also known as *Physeter macrocephalus*) derive their common name from spermaceti, a waxy, milky white substance that is found in abundance in sperm whales' heads. The exact function of spermaceti is not known: it might act as a sounding medium for echolocation, or to aid buoyancy, or possibly to give the head extra heft for head-buts in male-to-male combat. When whaling was a global industry, whale oil, derived from spermaceti, had variety of industrial uses, including the production of candles.

Reproduced by permission of Christian Darkin/ Science Photo Library.

different rates. This means that different parts of the genome can be selected to tell different stories. This chapter will use examples from the scientific literature to briefly illustrate what kind of information you can get from a tiny DNA sample. The rest of the book will show you how to do it.

Individuals, families, and populations

The genome of a sperm whale contains over 3,000,000,000 nucleotides of DNA. The nucleotides come in four types, which are given the single-letter codes A, C, G, and T. These four letters make up the DNA alphabet. All of the information needed to make the essential parts of a whale, such as the skin, the eyes, the blubber, and the blood, is coded in these four letters.

The genetic code is covered in Chapter 2

Most of the genome is exactly the same for all sperm whales, because all whales must be able to make functional skin, eyes, blubber, and blood in order to survive. But some of the DNA letters can change without destroying the information needed to make a working whale. Because of this, some DNA sequences vary slightly between individual sperm whales. Most of these differences between genomes arise when DNA is copied from the parent's genome to make the eggs or sperm that will go on to form a new individual. DNA copying is astoundingly accurate, but it is not perfect. In fact, you could probably expect around 100 differences in the nucleotide sequence of the sperm whale genome between a parent whale and its calf, due to mutation. The upshot of this is that, although most DNA sequences are exactly the same in all sperm whales, every individual whale has some changes to the genome that make it unique. So given enough DNA sequence data it would be possible to tell not only which species the Chilean blob had come from, but also which individual whale. The same rationale applies to forensic DNA analysis to identify biological samples left at crime scenes: because each individual has a unique genome, if the DNA from the victim matches the blood on the accused's clothes, then the two must be linked.

Chapter 3 explains how mutation makes individual genomes unique

The sperm whale's genome is copied when it reproduces, and any changes to its genome will also be inherited by the whale's offspring. Therefore, by asking which individuals share particular sequence changes, biologists can use DNA sequences to reveal the relationships between individual whales. Since DNA is inherited from both father and mother with relatively few changes, it is possible to use DNA sequences to conclusively identify an individual's parents (hence the growing number of companies offering 'paternity tests'). The inheritance of specific DNA differences can also be traced back through a whale's family tree, to its parents, grandparents, and great grandparents, and so on. So, in addition to identifying specific individuals, if you take DNA sequences from a whole group of whales you can tell who is whose mother or sister or cousin. More generally, you can start to understand how populations of sperm whales interact and interbreed.

Chapter 4 explains how DNA replication results in related individuals being more genetically similar

Sperm whale social groups

Sperm whales travel in social groups that co-operate to defend and protect each other, and may even share suckling of calves. It is difficult to determine the membership of these groups from sightings alone, because of the practical difficulties of observing whale behaviour, most of which happens underwater. To make things even more difficult, sperm whales can travel across entire oceans and can dive to a depth of a kilometre. Biologists who study whale behaviour generally have to be content with hanging around in boats, waiting for their subjects to surface. But when they do surface, in addition to taking photos which allow individual whales to be identified, biologists can zip over in

Figure 1.3 Research groups that use molecular data to study whales typically wait for whales to surface, then either scoop sloughed-off skin cells from the surface, or shoot biopsy darts that take small tissue samples. This photograph shows Scott Baker and his research student Carlos Olavarria in Tonga approaching a humpback whale in order to take a biopsy.

Reproduced by permission of C. Scott Baker, Marine Mammal Institute, Oregon State University and School of Biological Sciences, University of Auckland.

worryingly small boats and pick up the bits of skin that the whales leave behind on the surface when they resubmerge (Figure 1.3). The DNA extracted from these bits of whale skin not only identifies the individuals in the group, but also reveals their relationships to each other. This has allowed researchers to described sperm whale social groups in detail.

DNA analysis from these skin samples shows that sperm whale social groups are made up of 'matrilines', or female family groups (mothers, daughters, sisters, and so on). Males leave the group before they mature. But not all the individuals in the group are related to each other – there are members of several different matrilines in each group. This suggests that, while adult males come and go and rarely stay with the group longer than it takes to father more offspring, female sperm whales form long-term, and possibly life-long, relationships. By sequencing sperm whale skin samples (**TechBox 1.2**), it is possible to get an insight into the private lives of animals that were previously hard to observe, yet without disrupting their natural behaviour.

Whale populations in space and time

Sperm whales are a global species, found in every ocean. Female social groups inhabit relatively warm temperate and tropical waters and do not tend to move between oceans. But males, who leave the matrilineal social groups at a young age to join roaming 'bachelor schools', travel more widely, going to cooler polar waters to feed, and returning to lower latitudes to mate. Because males move away from their natal group and can travel between oceans, a calf's father could have been born in any ocean of the world. But if a sperm whale calf is born in the Pacific Ocean, then we can be almost certain that its mother was also born in the Pacific and so was its grandmother.

This behavioural difference between males and females is reflected in sperm whale DNA. A sperm whale's nuclear DNA is carried on 44 chromosomes, located in the nucleus of each cell. Each individual whale inherits half of its chromosomes from its mother and half from its father. As males move around the world and mate, they spread their nuclear DNA around.

TECHBOX 1.2

DNA sequencing

KEYWORDS

chain termination
Sanger sequencing
dideoxy
dNTP
primer
gel
pyrosequencing

FURTHER INFORMATION

An explanation of sequencing using an automated sequencing machine, with helpful diagrams, can be found at *http://allserv.rug.ac.be/~avierstr/principles/seq.html*

RELATED TECHBOXES

TB 2.4: DNA extraction
TB 4.2: DNA amplification

RELATED CASE STUDIES

CS 1.1: Chilean blob (identifying species)
CS 1.2: Whale meat (DNA surveillance)

To understand DNA sequencing, it is helpful to have a grasp of DNA structure (**TechBox 2.2**), DNA replication (**TechBox 4.1**), and DNA amplification (**TechBox 4.2**). This means that you may find this box rather challenging if you do not already have a basic understanding of molecular genetics. So why is DNA sequencing introduced in the first chapter? Because it is fundamental to all other topics covered in this book. We have to start somewhere, so let's start with sequencing. Just as this introductory chapter serves as a road map to the topics we will cover in later chapters, this TechBox will briefly touch upon many topics that we will cover in later TechBoxes. My advice is to read through it now without worrying too much about the details, then come back to it later when you have built up a basic understanding of the concepts involved.

1 DNA extraction

Most biological samples, such as a drop of dried blood, a fresh leaf, or the marrow from an old bone, contain DNA. Some tissues are easier to extract DNA from than others, and fresh material is usually easier to work with than old specimens (and some preservation techniques destroy DNA: **Case Study 1.1**). Chemicals, such as phenol and chloroform, are added to the sample to break down the cell membranes and precipitate the DNA so that it is floating in an identifiable layer of liquid that can be pipetted out.

DNA extraction is explained in **TechBox 2.4**

2 Amplification

You need to have a vast number of copies of the DNA you wish to sequence. So you need to take the DNA in your sample and amplify (make many copies of) the sequence that you are interested in. The most common way of doing this is following a series of reactions called the polymerase chain reaction (PCR). The DNA sample is heated (to separate the double strands), mixed with a primer (a short string of RNA that matches the start of the sequence: **TechBox 4.3**), cooled (so that the primer bonds with the DNA from the sample), then warmed again with polymerase (DNA-copying enzyme) which attaches to the primers so that only the target sequence is amplified.

DNA amplification is explained in **TechBox 4.2**

3 Sequencing

Here is an overview of the most common kind of sequencing reaction, which will be described in more detail below. The amplified DNA is heated (to separate the strands) and mixed with polymerase (DNA-copying enzyme), nucleotides (DNA 'letters'), and primers (short DNA sequences that serve as starting blocks for DNA synthesis). Some of the nucleotides are labelled with something that makes them detectable, such as fluorescent dyes or radioactive labels. And some of the nucleotides are modified so that they stop the synthesis of DNA whenever they are added to the new strand. The polymerase makes copies of the DNA in the sample, by constructing chains of nucleotides that are complementary to the template DNA (**TechBoxes 2.2** and **4.1**). While making these nucleotide chains, the

reaction occasionally incorporates a labelled nucleotide. And sometimes incorporates a reaction-stopping nucleotide, at which point the growth of that DNA chain stops. So throughout the reaction mixture, copies of the template DNA are being made by polymerase, and those copies are stopping at different points, whenever the polymerase happens to incorporate a reaction-stopping nucleotide. The result is a solution of DNA sequences of different lengths, each ending at a different 'letter' in the DNA sequence. If these sequences are 'read' in order of length, then the sequence of bases at the ends of the fragments provides the sequence for the template they were all copied from.

Chain-termination sequencing in more detail: The Sanger sequencing method (see **Heroes 1**) is known as the chain-termination method because it uses modified nucleotides to halt DNA synthesis at different points to produce an array of DNA fragments, each of which stops at a different nucleotide in the sequence. This method is also known as dideoxy sequencing because the modified nucleotides are missing their OH group. Since nucleotides are added to the OH group of the last nucleotide in the chain (see **TechBox 4.1**), once a dideoxynucleotide is added, no more nucleotides can be added to that chain, so synthesis stops.

The growth of a DNA chain is described in TechBoxes 4.1 and 4.2

The specific details differ between methods, but the basic approach of Sanger sequencing is:

(i) **Denature:** The amplified DNA is heated, to separate the double-stranded DNA into single strands.

The chemical bonds that hold the DNA helix together are described in TechBox 2.2

Case Study 4.1 describes an application of DNA hybridization (separating and rejoining DNA helices)

(ii) **Prime:** short sequences that have a complementary sequence to the target sequence are added to the single stranded DNA, which is cooled so that the primers can bind to the template (**TechBox 4.2**). The primer provides a starting block for synthesizing a new DNA strand, because it provides a free OH group for newly added bases to bind to (**TechBox 4.3**).

Primers are explained in TechBox 4.3

DNA synthesis in natural systems is covered in TechBox 4.1 and the laboratory methods for DNA amplification are covered in TechBox 4.2

(iii) **Nucleotides:** the amplified DNA with attached primers is split into four samples. To each of these samples is added the four DNA bases (in the form of deoxynucleotides: dATP + dTTP + dCTP + dGTP) and DNA polymerase (the enzyme that makes a new DNA strand to match the template: see **TechBox 4.1**). In addition, each sample has a different chain-terminating dideoxynucleotide (ddATP or ddTTP or ddCTP or ddGTP).

Nucleotides are covered in TechBox 2.2 on DNA structure

(iv) **Grow:** As the polymerase moves along the template, it picks up and adds nucleotides to make a matching DNA strand. If it picks up a deoxynucleotide (dNTP), it continues to add

nucleotides to the growing strand, but if it incorporates a dideoxynucleotide (ddNTP), chain elongation stops at that point. Using the example sequence given in the diagram, in the ddCTP sample, the polymerase will add nucleotides to the growing strand.

When it gets to a C it might incorporate a normal dCTP and keep going, or it might incorporate a chain terminating ddCTP and stop.

The upshot of this is that, in the ddCTP sample, there will be fragments of many different lengths, but they will all stop at a C. Similarly, the ddTTP sample will have fragments of different lengths that all end at a place in a sequence where there is a T.

DNA synthesis is covered in Chapter 4

(v) **Run:** To read the sequence, it is necessary to order the fragments with respect to size and report which nucleotide is at the end of each fragment. In traditional Sanger sequencing, the fragments are labelled radioactively (usually by labelling either the dNTPs or ddNTPs). The fragments are run through a gel, with each of the four samples run in a different lane. Since longer sequences are larger and travel more slowly through the gel, the bands will appear on the gel in order of length, so the nucleotides at the end of each fragment can be read in the order of the bands along the gel. In some automated sequencing reactions fluorescent dyes are added to the primer sequences, so that the fragments can be read by being passed through an optical reader. Alternatively, the ddNTPs can be labelled with different coloured fluorescent dyes: one advantage of this approach is that all labelled dNTPS can be added together, rather than in four separate reactions.

The movement of DNA fragments through a gel is described in Chapter 5

Pyrosequencing: this alternative approach to Sanger sequencing is gaining popularity for large-scale sequencing projects. In standard sequencing reactions, the reporter is some kind of label attached to the products of DNA synthesis. In pyrosequencing, the chain-elongation reaction itself is the reporter. Luciferase (a type of enzyme that produces bioluminescence, such as the enzyme that makes fireflies light up) is included in the reaction mixture, along with DNA, primers, and polymerase. Each dNTP is washed over the reaction mixture in turn. The incorporation of each new nucleotide releases a flash of light, which identifies the position of that particular base in the DNA sequence. The sequence can be read off by monitoring whether light is released as each type of nucleotide is added. 454 sequencing technology, which enables rapid, large-scale sequencing, is based on pyrosequencing.

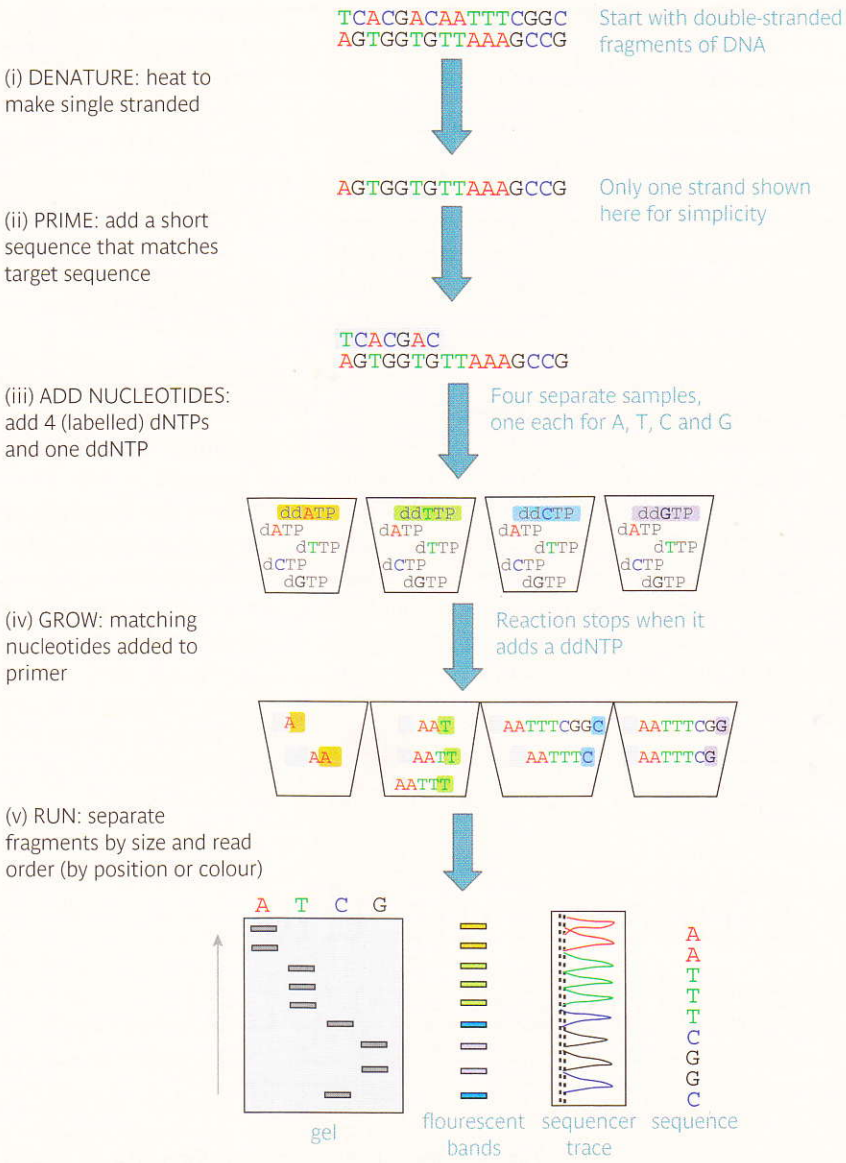


Figure TB1.2 A rough guide to DNA sequencing using the chain-termination (Sanger) method, including several different ways of reading the order of the nucleotides.

But most cells contain another source of DNA, in addition to the chromosomes in the nucleus. Mitochondria – energy-producing organelles found in the cellular cytoplasm – contain their own tiny genomes, less than 0.01% the size of the nuclear genome. In sperm whales, as in most vertebrate species, mitochondria are passed from generation to generation in the egg cells supplied by the mother. Males do not pass their mitochondrial DNA to their offspring, because the mitochondria carried in sperm cells are jettisoned from the fertilized embryo. Therefore, mitochondrial DNA is inherited through the female line. Since females do not roam as widely as males, each whale will tend to have the mitochondrial sequence typical of the ocean in which it was born. This means that if you were to give a sample of sperm whale skin (or blood, blubber, or tooth) to a

biologist, by sequencing the mitochondrial DNA, they could probably tell you in which region of the world that whale was born (see Case Study 1.2).

DNA sequences can reveal not only the current global distribution of sperm whales, but also how the population has changed over time. Changes accumulate in the mitochondrial genome as it is copied and passed from mother to daughter. With every generation, more DNA differences accumulate, so the number of differences between individual genomes increases. When measures of genetic similarity are taken for a whole population, it gives an indication of population size and mating structure. In a small population, there is an increased chance of mating with a relative. This means that in a small, inbred population, you would generally not have to trace two individuals' family trees back many generations before you found a shared ancestor. Since these two related individuals will have both inherited some of the same genetic variants from their common ancestor, you would expect their genomes to be more similar than when you compared two unrelated individuals. In a large population where mating between unrelated individuals is the rule, you would have to go much further back to find an ancestor shared by any two individuals. So two individuals in a large population probably have less of their genome in common than two individuals in a small population. Therefore, the average number of genetic differences between individuals gives an indication of population size. This can be very useful for estimating the numbers of individuals in species that are difficult to survey directly, such as sperm whales.

The effect of population size on DNA evolution is covered in Chapter 5

Current estimates suggest that the size of the global sperm whale population is probably around 360,000 individuals. But the genetic diversity of this population is surprisingly low. The low number of differences between the mitochondrial DNA sequences of sperm whales from around the globe suggests that they are all descended from a small number of founding mothers who survived the last ice age. Since sperm whale females prefer warmer waters, their distribution may have shrunk towards the equator as the world's oceans cooled, reducing their population size. As the ice age ended and the climate warmed, the sperm whale population might have expanded and spread out around the globe once more.

Uncovering the evolutionary past

Although the population may have reduced and expanded with the changing climate, sperm whales have swum in the oceans for millions of years. Sperm whales hunt using echolocation: they emit bursts of ultrasonic sound and use the sound reflected back from their surroundings to locate prey (which, in the case of sperm whales, includes giant squid). Echolocation is a characteristic that sperm whales share with other members of the Odontoceti, the group of predatory toothed whales that includes the dolphins and orcas. The other main group of cetaceans, the Mysticeti, have no need for echolocation. They are the baleen whales that use huge filter plates in their mouths (the baleen) to sieve plankton out of the water as they swim. The Mysticeti includes the gigantic blue whale, which, weighing up to 150 tonnes, is the largest animal that has ever lived. It was previously assumed that these two very different types of cetacean, odontocetes and mysticetes, evolved from a primitive whale species over 30 million years ago, and then each lineage developed special adaptations to their different lifestyles: the odontocetes evolved echolocation for hunting prey, and the mysticetes evolved baleens for sieving plankton.

The genome of the sperm whale tells a different story. Sperm whales are currently classified as odontocetes because they have classic toothed whale characteristics, such as teeth and echolocation (Figure 1.4, p. 17). But DNA sequences from sperm whales are more similar to sequences from baleen whales, not toothed whales.

CASE STUDY 1.2

DNA surveillance: using DNA species identification to trace illegal trade in whale meat

KEYWORDS

PCR

conservation

CITES

database

geographic origin

identifying individuals

DNA barcoding

FURTHER INFORMATION

www.dna-surveillance.auckland.ac.nz is the web-based species-identification site that takes a DNA sequence and aligns it against a curated database of reference sequences, and calculates genetic distance to known species.

RELATED TECHBOXES

TB 1.1: GenBank

TB 4.2: DNA amplification

RELATED CASE STUDIES

CS 2.1: Origin of faeces (identifying species from remote samples)

CS 6.2: Keeping the pieces (DNA and conservation)

Baker, C.S. and Palumbi, S.R. (1994) Which whales are hunted? A molecular-genetic approach to monitoring whaling. *Science*, Volume 265, pages 1538–1539

> *"These results confirmed the power of molecular methods in monitoring retail markets and pointed to the inadequacy of the current moratorium for ensuring the recovery of protected species."*[1]

Background

Following dramatic falls in global whale populations, commercial hunting of whales was banned by international treaty in 1986. Hunting of whales is now only conducted by a small number of countries. While some whale populations are increasing, there is a great deal of concern that many species are threatened with extinction. Protection for vulnerable whale species is therefore a conservation priority.

Aim

Japan continues to hunt minke whales (*Balaenoptera acutorostrata* and *Balaenoptera bonaerensis*) for scientific research, and, while import of whale meat is prohibited, there is no law against the sale of whale meat on the domestic market. So whale meat that is a by-product of the scientific catch can be legally sold in Japan. These scientists set out to test if all of the whale meat available in Japanese markets was sourced from the reported scientific catch of minke whales[2].

Methods

The UN Convention on Trade in Endangered Species (CITES), which prohibits international trade in rare animal products, does not allow whale meat to be taken across national boundaries without a permit. But amplified DNA (PCR product: **TechBox 4.2**) does not come under CITES legislation, because it is a synthetic copy of the DNA from the original sample. So the researchers developed a mobile PCR kit that they took to Japan. They would buy whale meat products from the market, then surreptitiously extract and PCR the DNA in their hotel room (**Figure CS1.2**). They took the PCR products back to universities in the US and New Zealand and sequenced them, and compared the whale meat sequences to sequences from known cetacean species.

Results

In the past 8 years, the team have identified the species and in some cases the geographic origin of more than 1100 whale meat products from at least 28 different species of whales and dolphins, including protected humpback, western gray, fin, sei, and Bryde's whales. Some whale meat was from species that could not have been caught in Japanese waters, such as a type of humpback whale found only in Mexican coastal waters, implying that whale meat was being moved between countries. They have even identified a particular individual, a rare fin whale/blue whale hybrid, killed off Iceland in 1989 and purchased in an Osaka market in 1993[3].

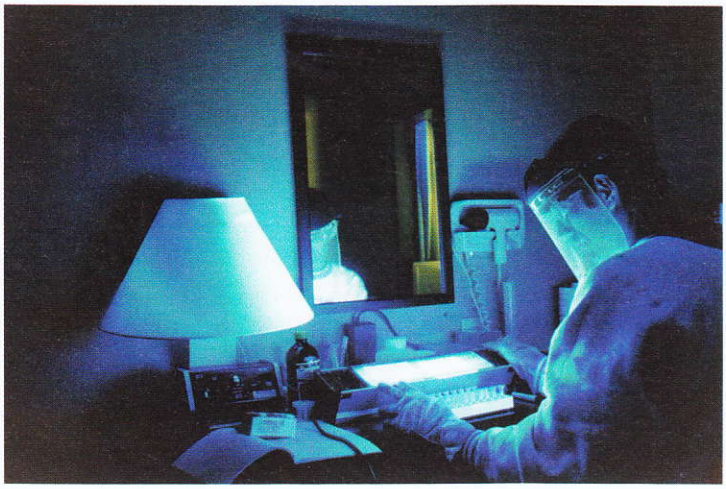


Figure CS1.2 Scott Baker using a 'portable laboratory', consisting of a PCR machine plus chemical reagents which were capable of being carried in a suitcase, to extract and amplify DNA in a hotel room. New DNA-amplification methods that do not rely on cycles of heating and cooling may make mobile DNA testing much easier (see **TechBox 4.2**).

Reproduced by permission of C. Scott Baker, Marine Mammal Institute, Oregon State University and School of Biological Sciences, University of Auckland.

Conclusions

Whale meat purchased in Japan came from a variety of species, including species that have been banned from hunting for over three decades, and species that cannot have been caught in Japanese waters. It was therefore not entirely by-products of the scientific catch.

Limitations

The DNA surveillance database is currently limited to two mitochondrial sequences for cetacean species, but could be expanded to other genes and other taxa (for example, it has recently been extended to include 'What Rat is That?', to identify rat species which can be hard to identify on morphological grounds alone). DNA surveillance reports the closest match between the database and the sample DNA: in the case of a poor match to the database, the sample should be checked against GenBank (**TechBox 1.1**, p. 3).

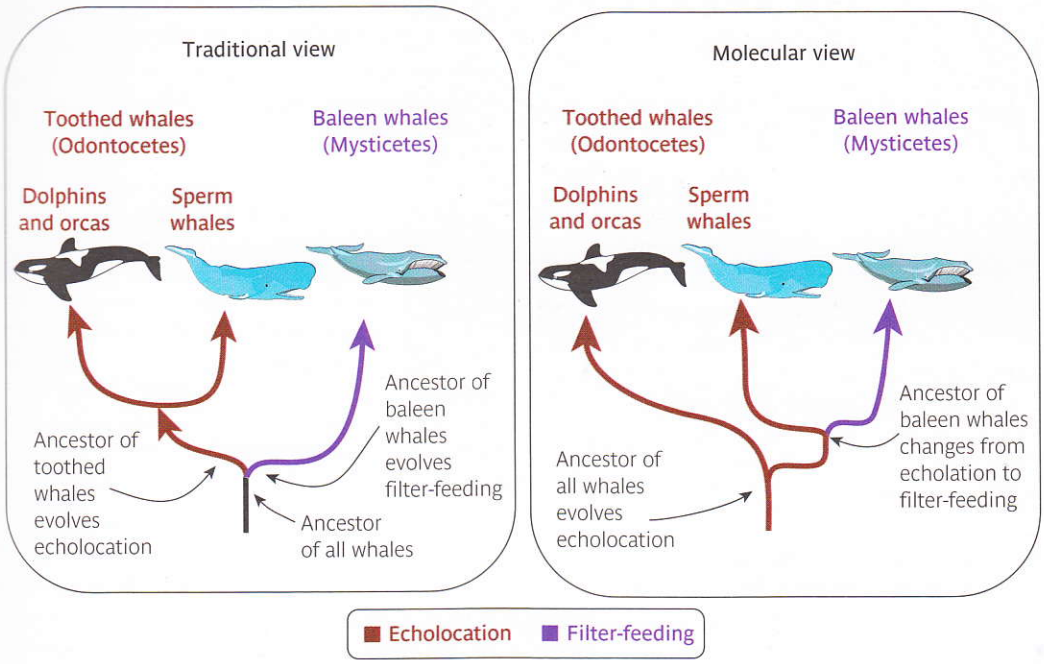
Future work

The team aims to develop 'Same-Day Whale Identification by DNA': using species-specific primers to test whether a sample is from the target species (**TechBox 4.3**; **Case Study 2.1**). This approach is being extended from identification of protected species to monitoring the demographic impact of hunting on species such as the minke whale[1]. The researchers have also used microsatellites to track multiple products from the same whale, in order to estimate the number of individual whales ending up in the market[4].

References

1. Baker, C.S., Lento, G.M., Cipriano, F. and Palumbi, S.R. (2000) Predicted decline of protected whales based on molecular genetic monitoring of Japanese and Korean markets. *Proceedings of the Royal Society London B*, Volume 267, pages 1191–1199.
2. Baker, C.S., Cipriano, F. and Palumbi, S.R. (1996) Molecular genetic identification of whale and dolphin products from commercial markets in Korea and Japan. *Molecular Ecology*, Volume 5, pages 671–685.
3. Cipriano, F. and Palumbi, S.R. (1999) Genetic tracking of a protected whale. *Nature*, Volume 397, pages 307–308.
4. Baker, C.S., Cooke, J.G., Lavery, S., Dalebout, M.L., Ma, Y.-U., Funahashi, N., Carraher, C. and Brownell, R.L. Jnr (2007) Estimating the number of whales entering trade using DNA profiling and capture-recapture analysis of market products. *Molecular Ecology*, Volume 16, pages 2617–2626.

Figure 1.4



This suggests that sperm whales are more closely related to the filter-feeding odontocetes than they are to the other echolocating mysticetes. This may seem like a fairly unimportant distinction, but it has important implications for the interpretation of whale evolution. If sperm whales belong with the mysticetes, then both major branches of the whale tree contain classic odontocete characters such as echolocation (Figure 1.4).

If echolocation is present in both major branches of the whale lineage, then the most likely explanation is that both lineages inherited echolocation from their common ancestor. This means that the ancestral whale must have had echolocation, but it was lost in the baleen whales as they adapted to a new way of life. This hypothesis, built upon DNA evidence, has gained some support from morphological studies, such as the evidence of vestigial 'melons' (echolocation sounding chambers) in baleen whales, remnants left from their predatory ancestor.

DNA sequences are valuable sources of information for ascertaining the relationships between living species. And these relationships can tell us a lot about the evolutionary past. The relationships between whale lineages, revealed by analysis of sperm whale DNA, shows that evolution both adds and takes away complex characteristics. In this way, evolutionary changes can sometimes obscure the history of species. But while a species' appearance changes, the genome continues to record evolutionary history. Because of this, molecular data can sometimes give a clearer view of a species' evolutionary past than can its highly modified morphology.

The importance of inherited similarities (homologies) for uncovering evolutionary relationships is covered in Chapter 6

Four-legged ancestors

Whales may look like giant fish, but inside they are typical mammals, with mammalian blood, bones, and organs. So although adaptations to life at sea have, in many ways, erased the signal of the whale's past, traces of the whales' ancestry can be seen in the way that the mammalian finger bones have been modified into flippers. Some whale species even have the remains of a pelvis left over from an ancestor that walked on four legs. But what exactly was this four-legged ancestor like? While fossil data are the ultimate source of

information on the morphology of ancestral species, molecular data can provide important clues by revealing which living mammals are the whales' closest relatives.

The whale skeleton has been so highly modified by evolution that biologists have argued over whether the whale is most closely related to artiodactyls (such as cows, camels, pigs), perissodactyls (horses, rhinos, tapirs), or carnivorans (cats, dogs, bears, etc.). But even when morphology changes dramatically, the genome continues, by and large, to steadily accumulate changes. Just as changes to the genome every generation allow the relationships between individual whales to be traced, so the sum of these changes over longer time periods allows the evolutionary relationships between species to be uncovered. By comparing the similarities and differences in the DNA sequences of different species it is possible to reconstruct their history as an evolutionary tree.

Chapter 7 explains how to estimate phylogenies (evolutionary trees) from DNA sequence data

When DNA sequences from whales are compared with those from other mammals, it is clear that they are most similar to artiodactyls, something that had been suspected for some time. But, more surprisingly, the DNA suggested that the whales' closest living relative is the hippo (Figure 1.5), and that whales and hippos share a more recent common ancestor that either do with the rest of the artiodactyls, such as pigs, goats, and camels. This initially startling idea came to be known as the Whippo Hypothesis. Although whales and hippos share some unusual characteristics, such as thick hairless skin insulated with layers of fat, these were previously considered to be convergent adaptations. That is, it was assumed that both whales and hippos independently evolved the same solutions to the shared problems of being warm-blooded mammals living in cold water. Molecular phylogenies suggest that these shared traits are not coincidental, but may have been inherited by both whales and hippos from a shared, semiaquatic ancestor. The DNA evidence suggests that hippos might provide clues to the evolution of fish-like whales from their hairy, four-legged, land-dwelling relatives.

The first whales

The fossil record of whales has improved dramatically in the last decade, with new finds providing more information on stages in the evolutionary series. The oldest whale fossils are 'walking whales' from the beginning of the Cenozoic period (which runs from 65 million years (Myr) ago to the present, and is sometimes given the romantic name 'Age of Mammals'). These four-legged mammals took to the water not long after the oceans had been vacated by the great marine reptiles, the fish-like ichthyosaurs, the serpent-like mososaurs, and the Loch Ness monster-like plesiosaurs. Fossils of these great aquatic reptiles are known from throughout the Mesozoic (from 250 to 65 Myr ago, known as the 'Age of Reptiles'). But, along with the dinosaurs, the icthyosaurs, mososaurs, and pleiosaurs all disappear from the fossil record by the beginning of Cenozoic era.

You can find the names of evolutionary eras on the geological timescale in Appendix III

The mammals that took over from the great sea-dwelling reptiles evolved similar adaptations to aquatic life, such as streamlined bodies and flippers rather than legs. Because of this, many odontocete whales (such as dolphins) look remarkably similar to the fish-like reptilian ichthyosaurs. The independent acquisition of similar traits is called evolutionary convergence (Chapter 6). The fossil record tells a similar story of convergence for other major groups of mammals which appear after the dinosaurs disappeared: hoofed mammals replaced browsing sauropods, carnivorans replaced predatory dinosaurs.

This picture of an evolutionary scramble to fill a world vacated by the dinosaurs has strongly influenced biologists' views of the relationships between the major groups of mammals, such as primates, artiodactyls, and bats. If, on being released from the tyranny of the giant reptiles, mammal groups all evolved simultaneously from a common ancestral stock, then rather than a serially branching evolutionary tree, mammalian relationships may resemble a 'bush', with all branches arising at once from a common root. This conclusion was supported by morphological studies that often failed to resolve any clear relationships between the mammalian orders.

But as the amount of DNA sequence data increases, the phylogeny of mammals is being resolved, and surprising new relationships have been suggested. For example, the base of the new molecular tree for mammals seems to be firmly rooted in Africa. DNA sequences

Figure 1.5 Hippos *(Hippopotamus amphibious)*, equally at home in the water and on the land, may provide clues to the origins of the fully aquatic Cetacea (whales and dolphins). Although hippos spend most of their time in shallow water, and can dive for 5 minutes or more, on land they can run faster than humans.

Reproduced by permission of Paul Martitz.

analysis has united an unlikely group of mammals, including aardvarks, elephants, and tenrecs, into a group now known as the Afrotheria (Figure 1.6, p. 20). When the evolutionary tree of mammals is constructed from molecular data, it is these Afrotherian lineages that are amongst the earliest lineages to branch off from the other mammals. This has been used to suggest that the early diversification of placental mammals occurred in ancient Africa, when the continent was isolated from the rest of the world, then spread out from there in the Cenozoic. There is currently little available fossil evidence that could shed light on land vertebrates in Africa during the Mesozoic, so an African genesis of mammals would essentially hide their early evolutionary history from the known fossil record.

Furthermore, the DNA evidence suggests that many of the major branches of the mammalian evolutionary tree stretch back into the time of the dinosaurs. Because changes to the genome accumulate continuously, the longer two lineages have been evolving separately the more differences you expect to see between their genomes. For example, the genomes of two species of baleen whale are more similar than either is to a sperm whale's genome, because the baleen whales' genomes were more recently copied from the same shared ancestor. If changes to DNA accumulate at a predictable rate in all species, then we can use a measure of genetic difference to estimate when two species last shared a common ancestor. When DNA from mammals such as whales, cats, monkeys, and rats are compared, the results are surprising: there are far more DNA differences between the major mammal groups than you would expect if their common ancestor had lived less than 65 Myr ago. Instead, the molecular data suggest that major branches of mammalian

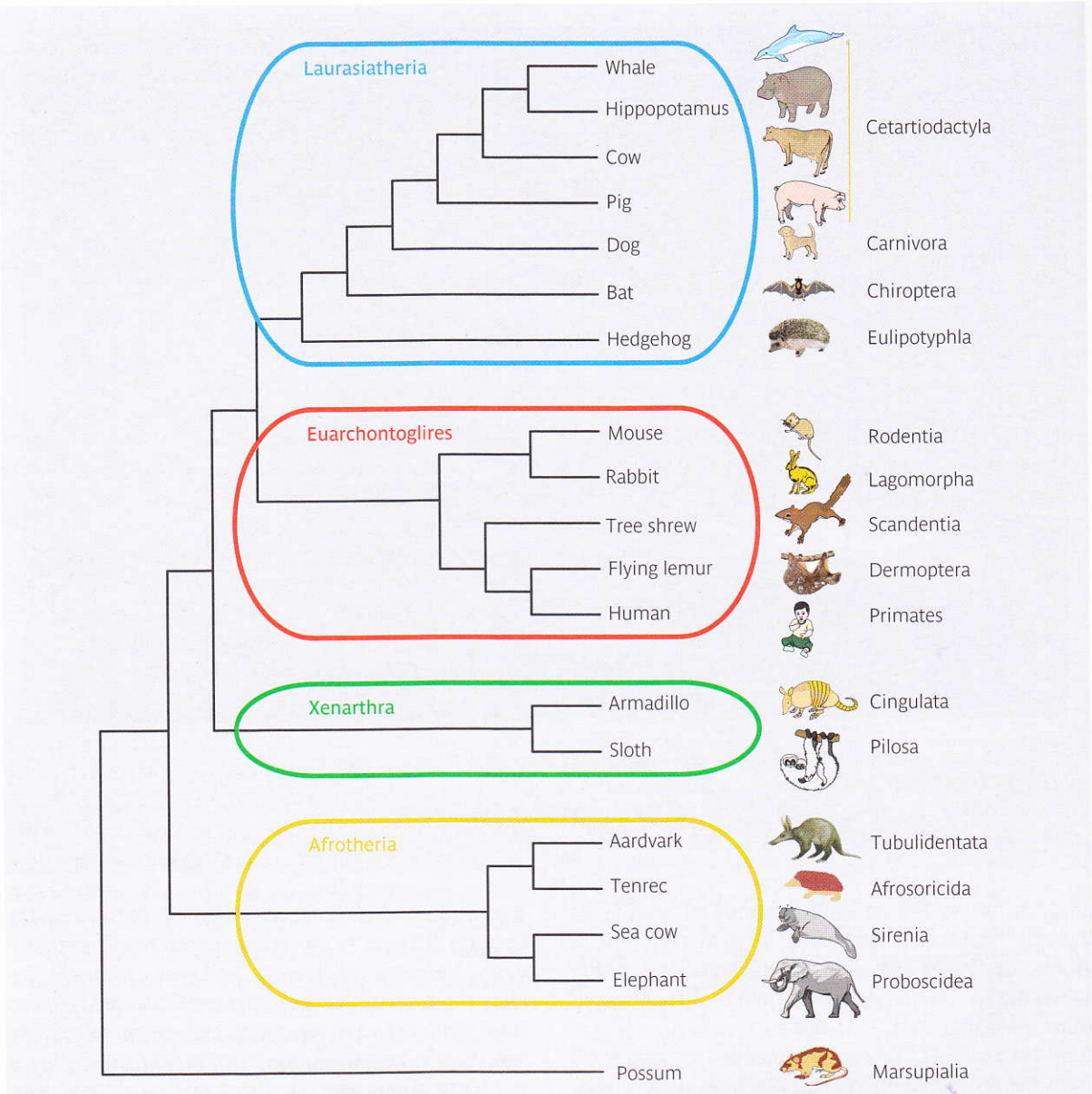


Figure 1.6 The phylogeny of placental mammals has been revised in the last decade, reshaping ideas about mammalian evolution. In particular, molecular data suggest a new grouping of 'Afrotherians' as one of the earliest lineages to emerge from the mammalian radiation.

evolutionary tree arose deep in the Mesozoic, long before the final extinction of the dinosaurs.

Estimating evolutionary time from DNA sequences is covered in Chapter 8

So analyses of DNA sequences paint a very different picture of mammal evolution, not an explosive post-dinosaur radiation, but a gradual Mesozoic diversification. But these molecular date estimates are controversial. If changes to DNA happen when it is copied

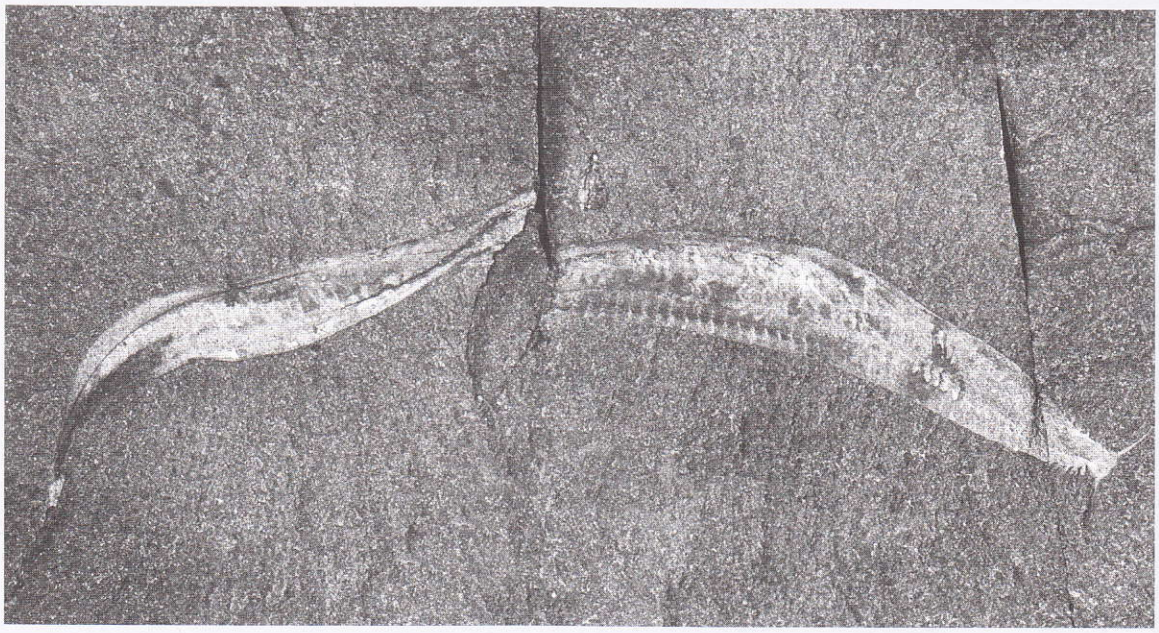
each generation, then the gigantic sperm whales that take a decade to mature might accumulate DNA differences at a slower rate than their diminutive artiodactyl ancestors that might have bred every year. If that was the case, then assuming that DNA differences accumulated at the same rate in the ancient walking whales as they do in their gigantic sperm whale descendants could lead to incorrect estimates of the time of origin of the whale lineage. As our understanding of molecular evolution grows, as statistical methods become more sophisticated, and as computers get more powerful, we will become better at deciphering the story in the DNA.

Evolution of animal body plans

The DNA of a sperm whale can reach even further back into the whales' history, back to its vertebrate ancestors. A sperm whale's flipper looks nothing like a sheep's hoof, a monkey's hand, or a bat's wing, and yet we can recognize the same underlying structure, not only in all mammals but also in a gecko's toes and a parrot's wing. These disparate animals also share the 'head and tail' body plan that unites the members of the phylum Chordata, which includes the mammals, birds, reptiles, and fish. The group gets its name from the central nervous cord which runs from the head (which holds the sense organs, mouth, and respiratory equipment) to the muscular tail. It is possible to piece together a more-or-less continuous series of forms that illustrate the evolution of the chordate body plan, linking gradual transitions along the evolutionary paths that connect the sperm whale to the monkey, the parrot, and the gecko.

But it is not so easy to follow the evolutionary paths that connect the chordates to other types of animals, for example linking the sperm whale to the barnacle stuck on its fin and the flatworm inside its guts. All animals are united by common features that reveal a shared ancestry, such as cell junctions that allow their multicellular bodies to be co-ordinated, absence of cell walls which permits movement and flexibility, and heterotrophy (consuming biological material as food, rather than producing their own energy from light or chemicals). But beyond the basic shared features of multicellularity, locomotion, and heterotrophy, the main groups of animals are strikingly different. The major divisions of the animal kingdom – the phyla – are often considered to represent different body plans, or basic ways of constructing animals. For example, the arthropod body plan consists of a segmented body with jointed appendages, all clothed in a jointed exoskeleton of protein and chitin (which is what makes bugs crunch underfoot). The echinoderm body plan, on the other hand, has pentameral symmetry (like a five-pointed star), an exoskeleton made of hard calcite, and a water vascular system that, in addition to transporting nutrients around the body, can power locomotion through hundreds of soft, hydraulically operated feet (turn a starfish over to see these tube feet in action).

These animal body plans are so different from each other that there has been an intense (and often rather lively) debate about how they evolved. We can trace the whale and barnacle lineages back half a billion years, to Cambrian-age rocks that contain fossilized animals with recognizable chordate and arthropod body plans (**Figure 1.7**: alas, flatworms do not tend to leave fossils). But there the trail ends rather abruptly. There is no continuous series of fossils showing the different animal body plans diverging gradually from each other, slowly modifying existing features to form the special characteristics of their phylum. Instead, a great diversity of different body plans appear almost simultaneously in the early Cambrian, complete with eyes, limbs, segments, or armour, in an explosion of animal forms. Some interpret this Cambrian explosion as the result of an imperfect fossil record, that failed to preserve the earlier animal ancestors, perhaps because they were small and squishy and lacked the skeletons, shells, and spikes that would have granted them geological immortality. DNA evidence has been used to support this interpretation. When genes shared by whales, barnacles, and flatworms are compared, there are more differences in the DNA sequences than expected from half a billion years of evolution, suggesting that these major



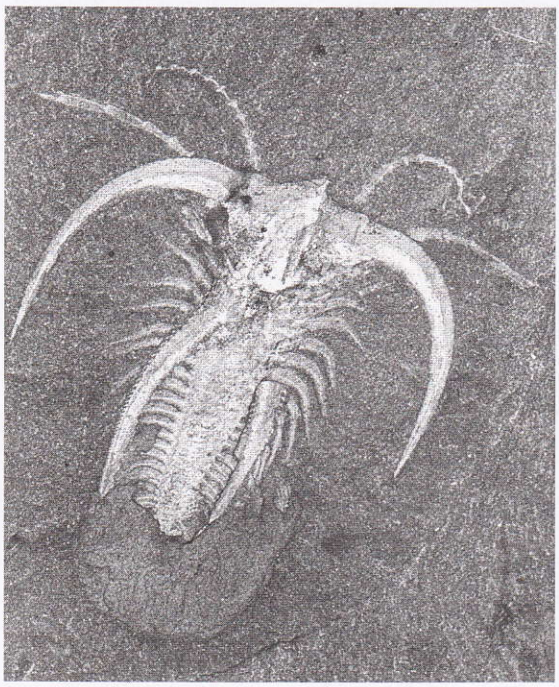


Figure 1.7 The arthropod and chordate 'body plans' have a long history and can be found in Cambrian age rocks, over half a billion years old. Fossil beds such as the Burgess Shale in Canada and Chenjiang in China contain exquisitely preserved examples of these aquatic animals, such as *Marrella* (tiny arthropod 'lace crabs') and *Pikaia* (an early chordate, something like the modern amphioxus).

© 2007 Smithsonian Institution

branches of the evolutionary tree of animals stretch back beyond the first fossils, deep into the Precambrian era. If the beginnings of the animal kingdom are before the fossil record starts, then DNA evidence will be essential in revealing its ancient history.

The role of DNA sequences in understanding the diversification of animals is covered in Chapter 8

But there is an alternative interpretation that is gaining ground, fuelled by the growth in developmental biology.

Could it be possible that body plans as different as chordates, arthopods, and flatworms evolved so rapidly that it requires a radically different evolutionary explanation than the gradual Darwinian plodding that transformed the ancestral chordate into lungfish, parrots, geckos, and whales? There has been a great deal of excitement surrounding the discovery of potential 'body-plan genes', which, with a single mutation, can have dramatic effects such as causing a second pair of wings to grow, or switching on the formation of extra eyes. Could the Cambrian explosion represent a remarkable period when changes to body-plan genes created animal diversity in evolutionary leaps, rather than taking a long series of tiny evolutionary steps? We cannot directly study the genes that were present in long-extinct animals, but we can use DNA sequence analysis to reconstruct the likely state of ancient genomes. The DNA sequences of body-plan genes from animals alive today can be used to trace the history of the genes themselves, as they were copied and changed and shaped to take on new functions.

Back to the beginning

Some sequences in the sperm whale genome code for such fundamental properties of life that they are shared not only with other animals, but with all living things, including oak trees, mushrooms, seaweed, and bacteria. The DNA sequences of these basic genes reveal the deep history of all life. These shared parts of our genome demonstrate that all life on Earth has a single common origin. The original genome, present in the last common ancestor of all plants, animals, fungi, algae, and bacteria, has been modified and expanded, but there is a part of that ancestral genome in all of us. The DNA from the globster was enough to prove that it was the remains of a sperm whale, and not (as some hoped) an extraterrestrial visitor. But if biologists ever do get their hands on a sample of alien life, the first thing they will do is check to see if it has DNA. If it does, it is almost certainly our distant relative. And using the techniques described in this book, we would be able to use the alien DNA sequence to investigate the history and biology of our cousins from space.

A small piece of the mystery blob that washed up on a beach in Chile contains enough DNA to tell a long and wonderful story. The story begins with a sperm whale, born in the same ocean that its mother had been, and its grandmother, and great grandmother, and so on back into history. Our sperm whale's ancestor was one of the founding mothers who survived the last ice age, although she may have had to retreat to tropical waters to do so. The story reaches back in time to tell of the origins of the sperm whales, whose predatory ancestors used echolocation to swim after their prey, leaving their four-legged relatives behind, wallowing in the mud. These ancestors were part of a radiation that exploded onto the world as the dinosaurs left the stage, and yet the roots of this radiation were planted firmly in the time of the reptiles. Both the reptiles and the mammals were themselves products of the diversification of the successful chordate body plan, which appeared in the fossil record half a billion years ago, but may have more ancient beginnings. And using the DNA from the blob, we can follow the story right back, beyond the limits of the fossil record, to the origins of the animal kingdom and, ultimately, back to the last living ancestor of all life. Not bad for a piece of blubber.

About this book

What is this book for?

The aim of this book is to provide a from-the-ground-up introduction to the use of DNA sequence data in evolutionary biology. Obviously, it is not exhaustive: there are many fascinating and useful ideas and techniques that are not covered here. And, in such a fast moving field, it is inevitable that this book will be out-of-date almost as soon as it is printed. So rather than trying to give an exhaustive introduction to the field, the aim is to give you the kind of fundamental information and intellectual tools you need to understand the way DNA sequences are used in biology.

That is why you won't find instructions for specific programs or particular laboratory protocols in this book. A program that is all the rage today is likely to be superseded by improved methods next year. Lab

Assignment

Q1. It is known that the initiation of replication is mediated by DnaA, a protein that binds to a short segment within the ori known as a DnaA box, which is a 9-mer. Is it possible to say which of the sequences given below is likely to be a DnaA box sequence? Give reasons.

CTCTTGATC ATGATCAAG, TCTTGATCA, and CTTGATCAT

Assignment

Q2. Two individuals are taken from two different species, say A and B. If the similarity between the individuals of species is A is more compared to that between species B, what can you say about the population size of the two species? Give reason to support your answer.

Sequence Comparison

DotPlots & Alignments

Computational Molecular Biology

Genome Analysis/ Sequence Analysis

- involves identifying characteristic features in a genome

Some important analytical approaches involve:

- **Sequence Alignment - to identify regions of similarity (Pairwise & Multiple)**
- **Pattern search - identifying repeats, motifs, CDS, etc.**
- **Database search - sequence/pattern-based search to identify similar sequences in the database**
- **Statistical measures – *ab initio* methods based on certain characteristic features of sequence (e.g., gene prediction), evaluating significance of alignment/motifs in Db search.**

Types of Mutations

- **Mutations - are local changes in DNA content, caused by inexact replication. There are various kinds of mutations:**
- **Substitution - a wrong base is incorporated instead of a true copy. A substitution may or may not alter the protein sequence depending on the place it occurs, e.g., GUU, GUC, GUA, GUG all code for Valine, GGU – Glycine, CUU – Leucine; Val & Leu – non-polar, Gly - polar**
- **Insertion / Deletion - addition/removal of one or more bases - leads to frame-shift in coding regions.**
- **Rearrangement - a change in the order of complete segments along a chromosome, e.g., human and mouse genome are very similar – major difference being the internal order of DNA segments.**

Mutations are important for several reasons:

- **are the source of phenotypic variation on which natural selection acts, creating species & changing them, allowing them to adapt to changes in the environment, etc.**
- **are responsible for inherited disorders and diseases including cancer, which involve alterations in gene.**

To understand evolution we need to know the various types of mutations that occur, frequency/distribution of their occurrence, and their effect.

For disease diagnosis, we need to understand the types of mutations, their inheritance pattern, their phenotype, etc.

Sequence Comparison

Why compare sequences?

Why Compare Sequences?

Sequencing of genomes – has outputted an enormous amount of sequence data on new proteins

Fundamental problem – determination of the function of a new protein

If there is significant sequence similarity between a pair of sequences, we can extrapolate the functional annotation of one sequence to the other.

Any other reasons for Sequence Comparison?

Comparison of Sequences

- **Identifying species – as in the case of DNA barcoding**
- **Phylogenetic analysis – to find evolutionary relatedness between species**
- **Genome comparison between individuals in a population – for structural variation analysis**
- **Genome comparison - diseased (e.g., cancer) *vs* normal cells – for identifying variations responsible for the disease**
- **Genome comparison between species – for understanding genome evolution**
- **Identifying overlapping regions – for genome assembly**
- **Identifying repeats, multiple copies of domains**
- **Identifying self-complementary regions in RNA sequences for structure prediction**

Computational Methods in Sequence Comparison:

- **Graphical methods - visual /qualitative comparison - dotplots**
- **Sequence Alignment: Determine residue-residue comparison to identify patterns of conservation and variability.**

  - pairwise alignment

  e.g., identify genes/proteins belonging to the same family.
- **Database Search: Look for homologs of query genes/proteins in the database**
- **Knowledge-based prediction: extract empirical rules from known examples representing sequence-structure or sequence-function relationships.**

  – multiple alignment

  e.g., motif identification, identifying remote homologs

Dot Plots - Graphical Comparison of Sequences

One of the simplest method for comparing two sequences, described by Gibbs & McIntyre (1970)

A dot plot is a visual representation of the regions of similarity within a sequence/between two sequences.

A dot plot can identify

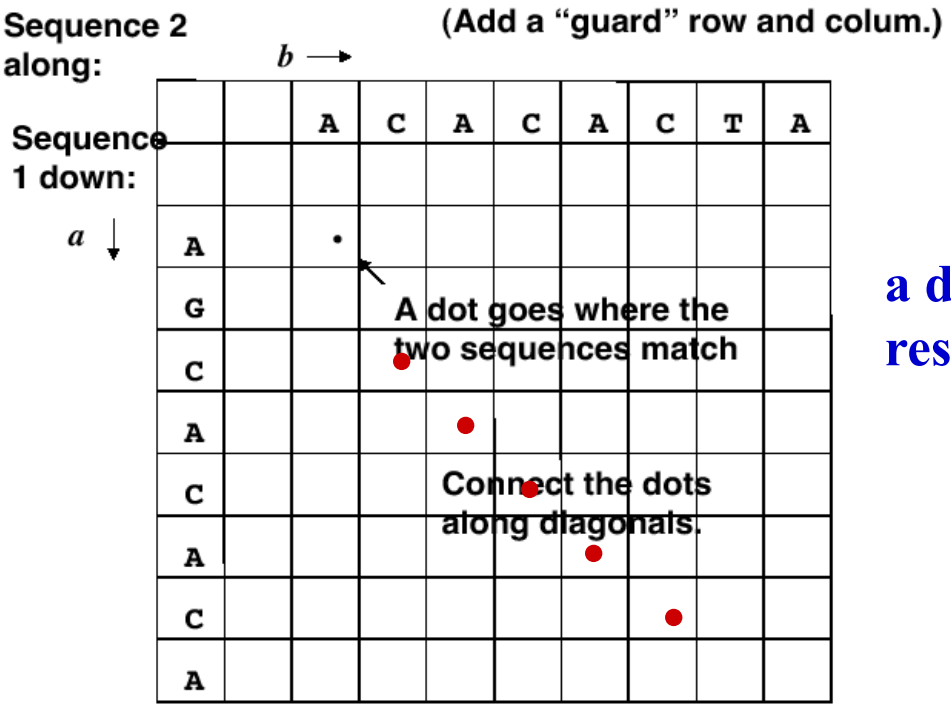
- **regions of similarity**
- **overlap regions**
- **rearrangement events**

Comparing two/more sequences

- **internal repeats, multiple copies of domains**
- **self-complementary regions in RNA sequences**

Self-comparison

Dot Plots



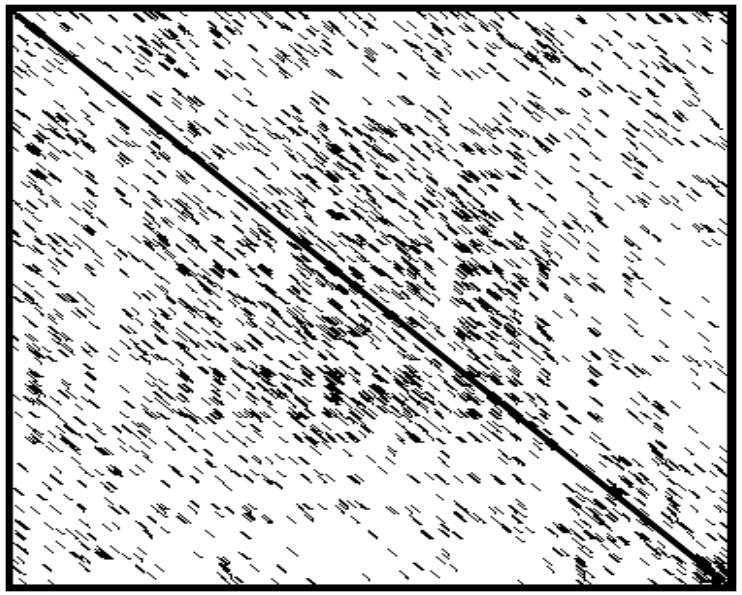
a dot is drawn for residue-residue match

Where the two sequences have substantial regions of similarity, many dots align to form diagonal lines

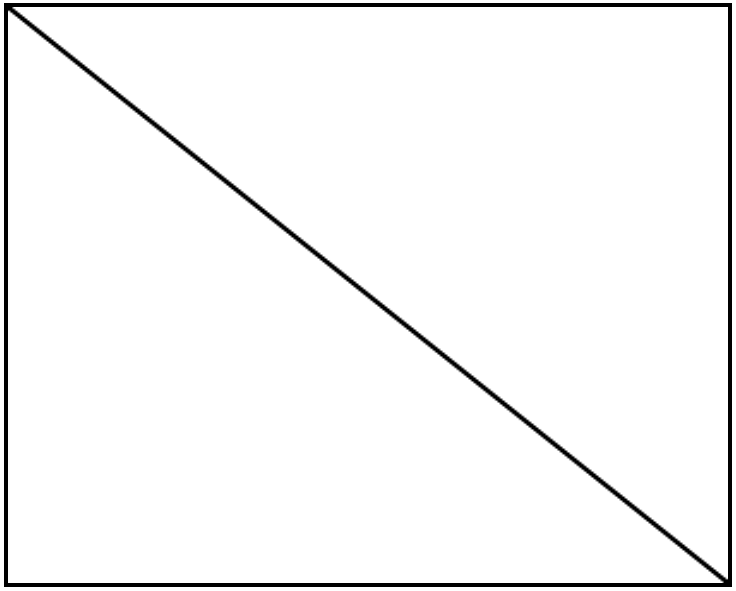
Dot Plots

When two sequences share similarity over their entire length, a diagonal line will extend from one corner of the dot plot to the diagonally opposite corner.

Non-stringent, self-dot plot



Very stringent, self-dot plot



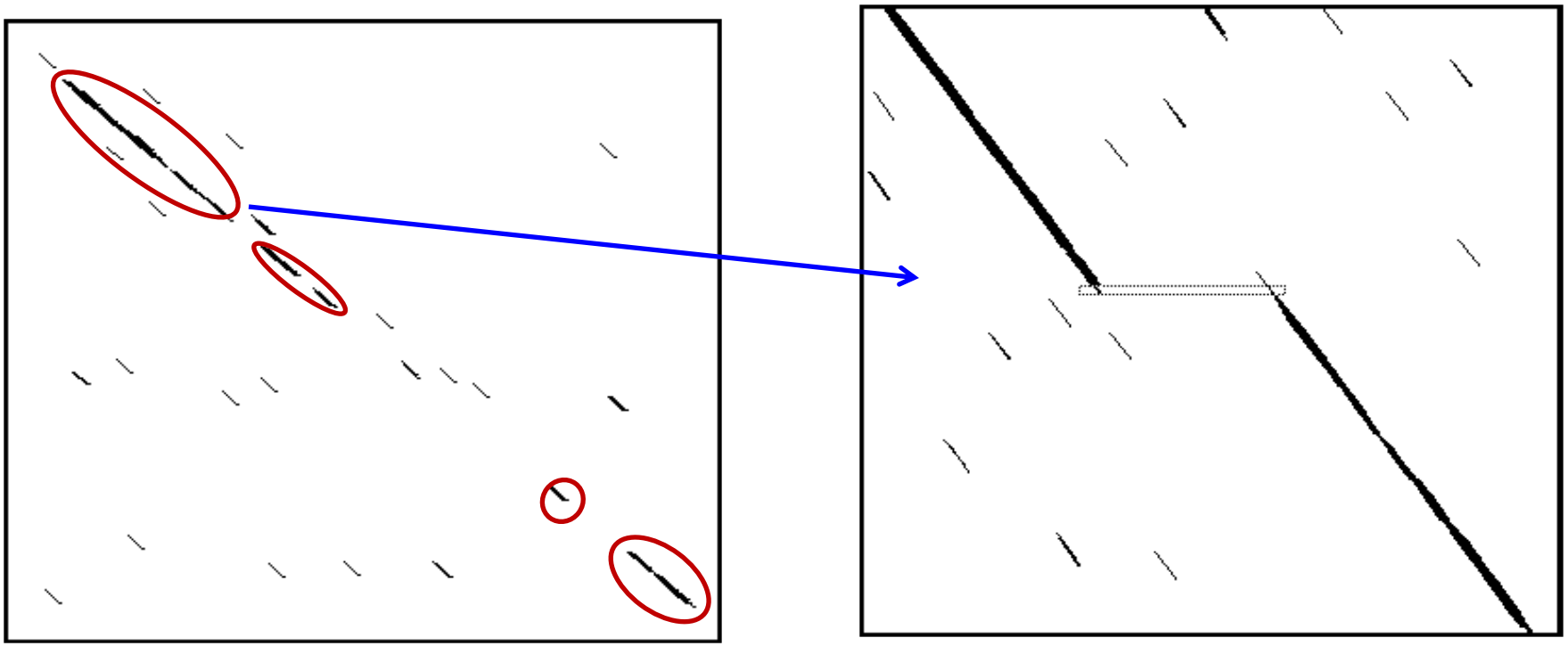
Every residue in one sequence is compared to every residue in the other sequence - nothing is missed

Dot Plots

If two sequences only share patches of similarity this will be revealed by short diagonal stretches.

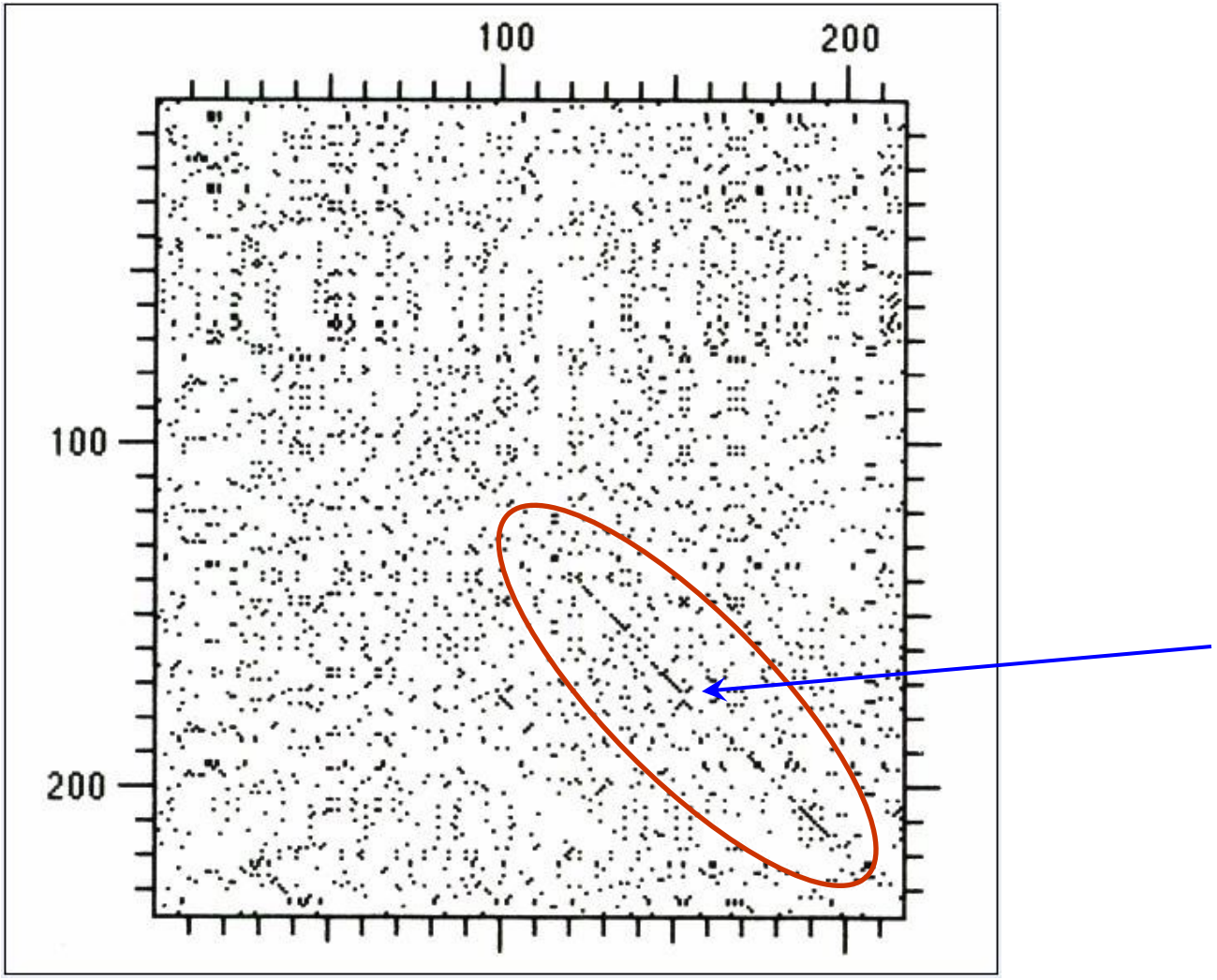
Two similar, but not identical sequences

Insertion or Deletion



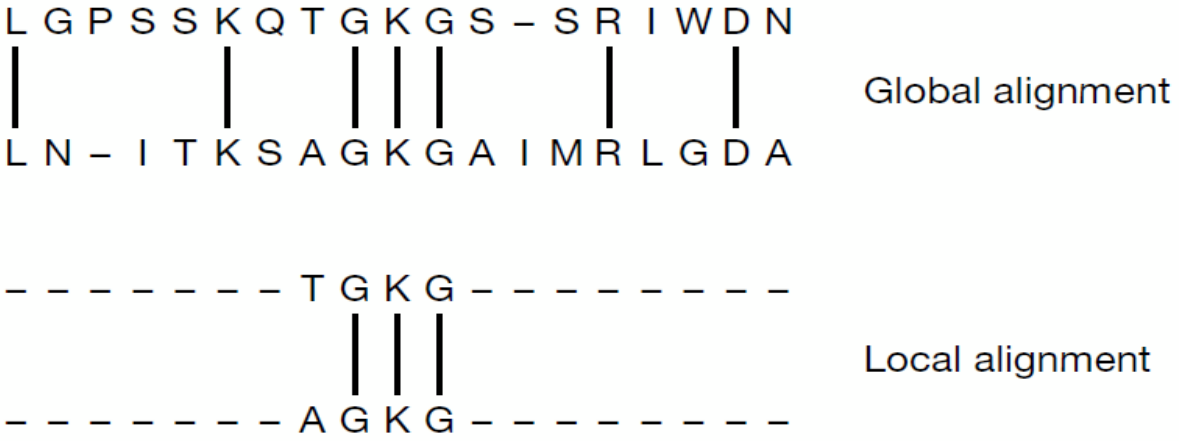
Homologous sequence comparison

Dot matrix analysis of amino acid sequences of the phage λ cI and phage P22 c2 repressors



Dot Plots

- **Major advantage of dot matrix method for finding sequence alignment - all possible matches of residues between two sequences are found, leaving investigator choice of identifying the most significant ones**
- **Based on the dot plot, user can decide whether one is dealing with a case of global (end-to-end), local, or overlapping (similarity at the ends) similarity**



Dot Plots

Detection of matching region is improved by filtering out random matches in a dot matrix - by using a sliding window to compare the two sequences.

Instead of comparing every base, a window of adjacent positions in the two sequences is compared and a dot is printed only if a certain minimal number of matches occur.

Extensions of Dot Plots

Thus, for window analysis of dot plots we define:

- **Window: size of diagonal strip centered on an entry, over which matching is accumulated, and**
- **Stringency: the extent of agreement required over the window, before a dot is placed at the central entry.**

- increasing window size would result in a faster search, but at the cost of sensitivity

Dot Plots

A large window size is generally used for DNA sequences.

- typically a window size of 15 and a suitable match requirement of 10.

For protein sequences, the matrix is often not filtered, but a window size of 2 or 3 and a match requirement of 1or 2 will highlight matching regions.

Why?

Dot Plots

A large window size is generally used for DNA sequences.

- typically a window size of 15 and a suitable match requirement of 10.

For protein sequences, the matrix is often not filtered, but a window size of 2 or 3 and a match requirement of 1or 2 will highlight matching regions.

Why?

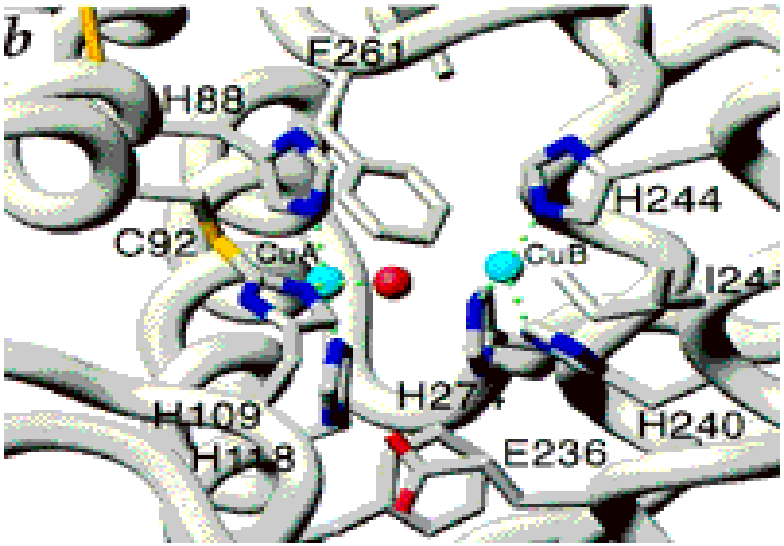
- the no. of random matches is more in case of DNA due to the use of 4 nucleotides symbols as compared to 20 amino acid symbols for proteins.

Dot Plots

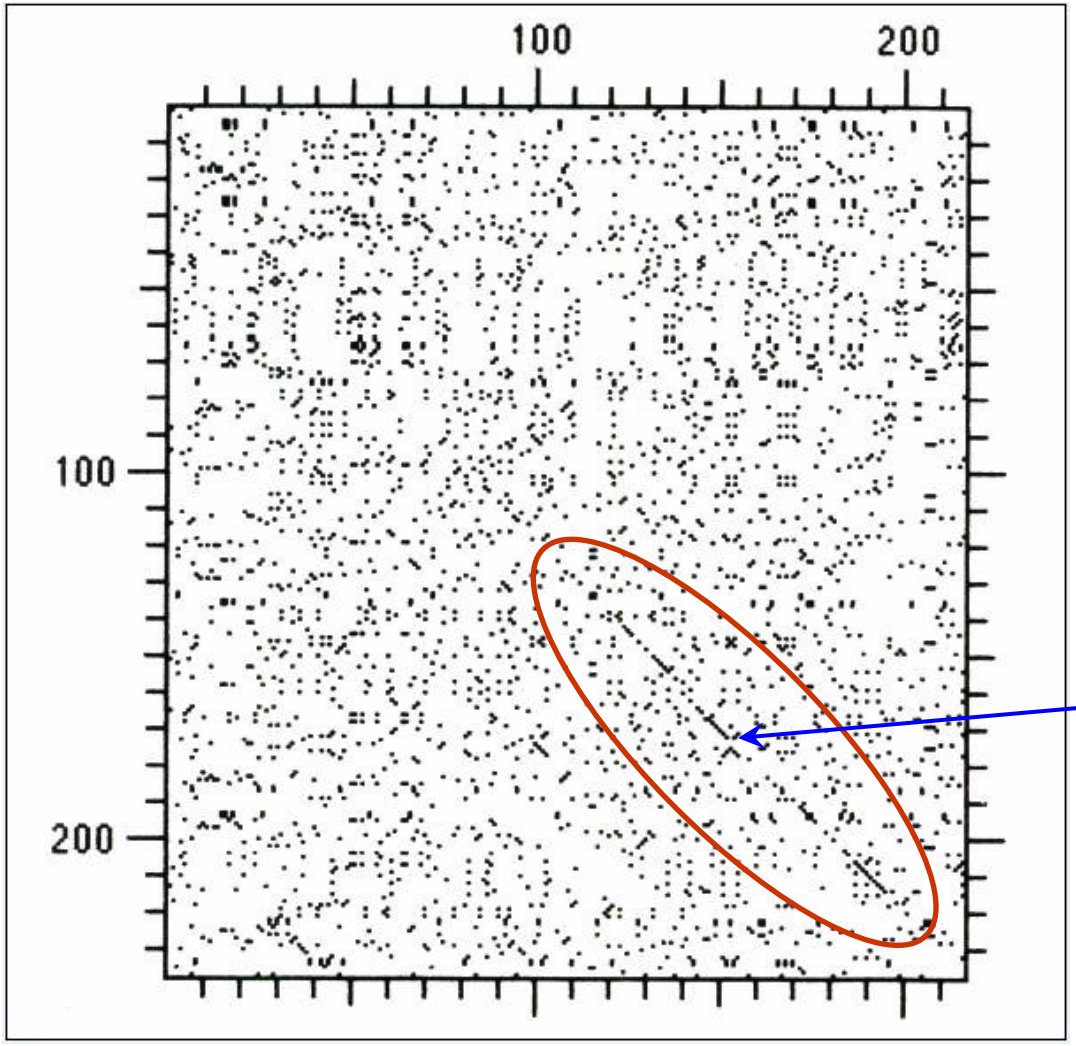
If two proteins are expected to be related but have long regions of dissimilar sequence with only a small proportion of identities, such as similar active sites,

- a large window, e.g., 20, and a small stringency, e.g., 5, should be useful for seeing any similarity.

- the reason being, residues in an active site are not necessarily contiguous in the sequence, and only the positions involved in interaction are conserved.

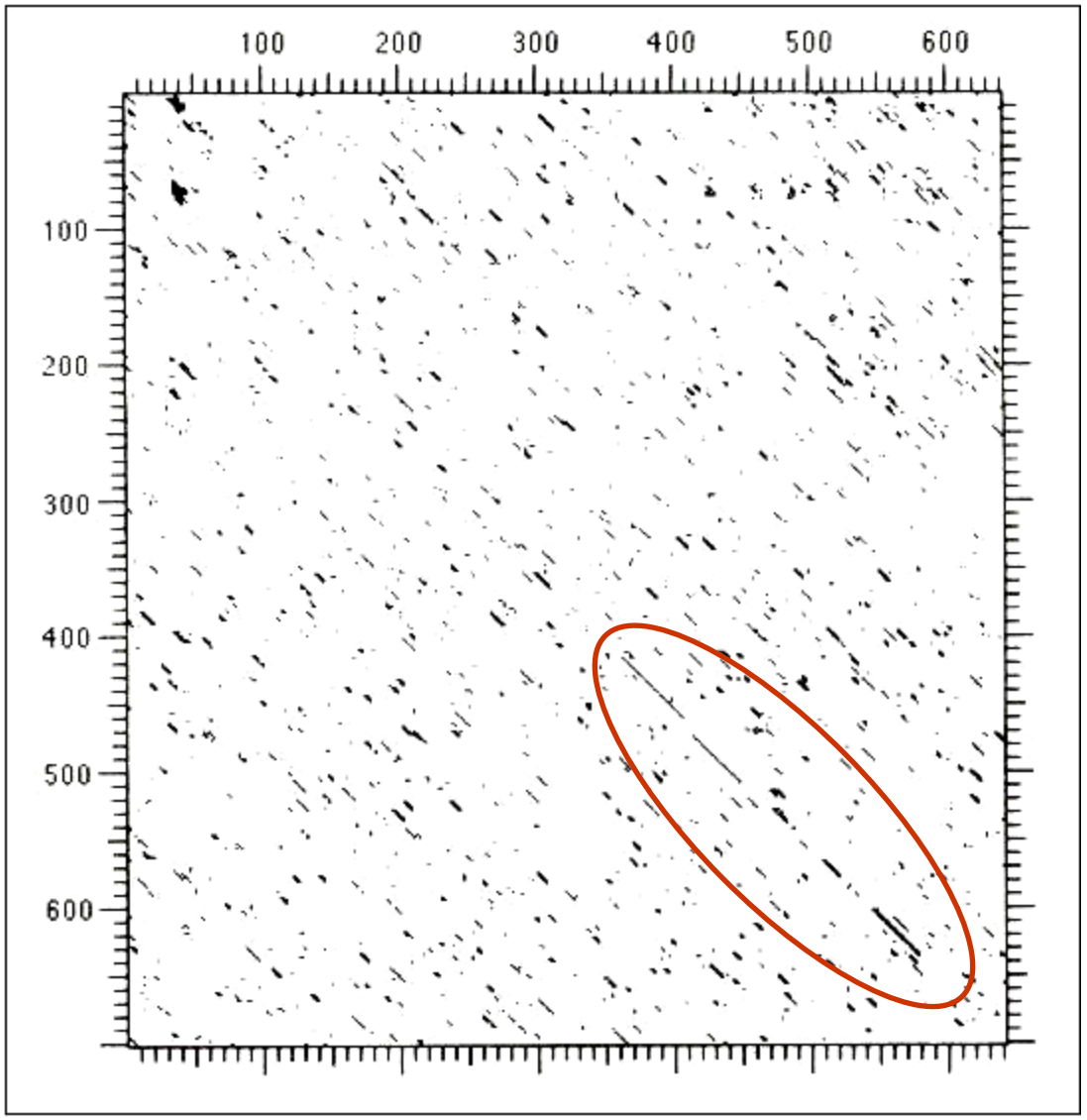


Dot matrix analysis of amino acid sequences of the phage λ cI and phage P22 c2 repressors



Window size: 1
Stringency: 1

Dot matrix analysis of DNA sequences encoding the E. coli phage λ cI (horizontal) & phage P22 c2 (vertical) repressors



Window size: 11
Stringency: 7

similarity in the C-terminal domains of the encoded proteins clearly seen

There are three types of variations in the analysis of protein sequences by the dot matrix method.

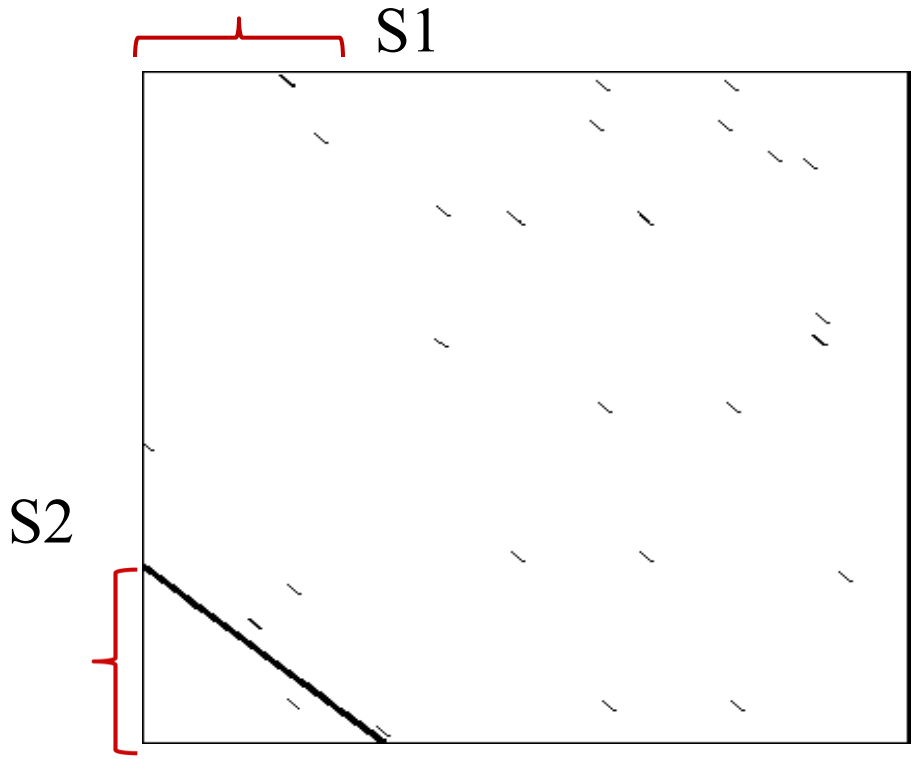
- **First, chemical similarity or some other feature for distinguishing AAs may be used to score similarity.**
- **Second, scoring matrices may be used to provide scores for matches based on their occurrence in aligned protein families.**

When these tables are used, a dot is placed in the matrix only if a minimum similarity score is found.

These table values may also be used in a sliding window option, which averages the score within the window, and prints a dot only above a certain average score.

- improves the sensitivity of a dotplot while comparing protein sequences

Identifying Overlapping Sequences Dot Plots



When do we expect to find overlapping sequences?

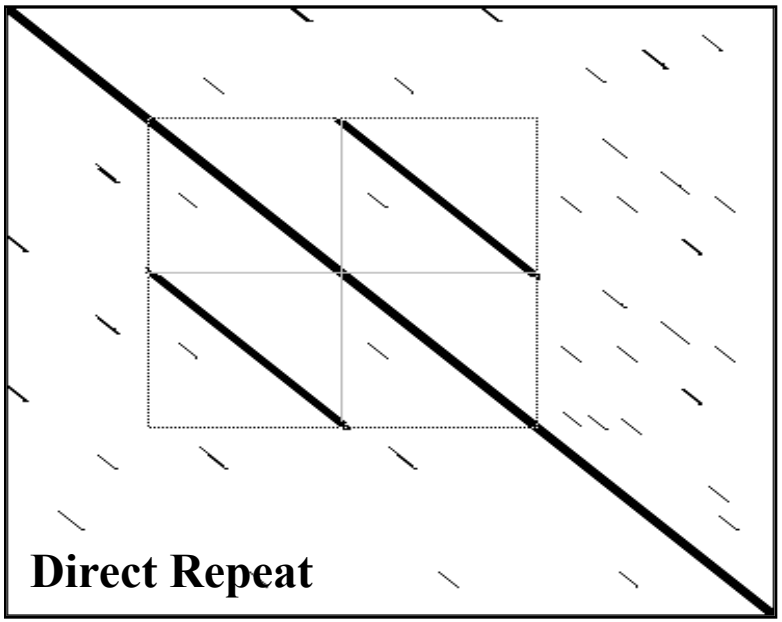
- during sequence assembly, aligning ESTs to gene / genomic sequences

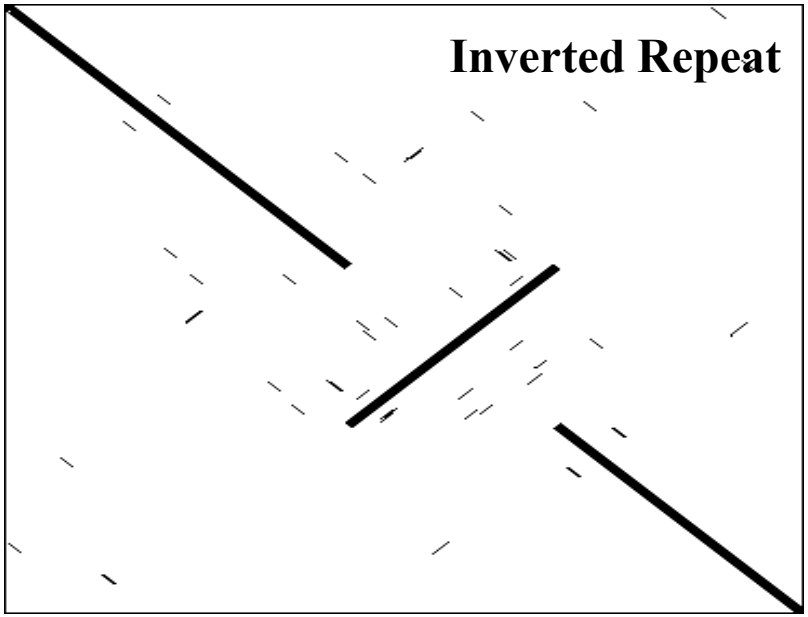
Dot Plots

- **Sequences may contain regions of self-similarity termed internal repeats. A dot plot comparison of the sequence with itself will reveal internal repeats by displaying several parallel diagonals.**
- **Presence of repeats of the same character many times (low-complexity regions) appear as - horizontal or vertical rows of dots that sometimes merge into rectangular or square patterns**

Dot Plots

Self-dot plot of a tandem duplication





We can compare a sequence to itself - it reveals repeat regions in the sequence

Sequence Repeats

Identifying direct and inverted repeats within sequences using Dot matrix analysis.

Sequence is aligned against itself and the presence of repeats is revealed by rows of dots parallel to the diagonal

	A	G	G	C	G	C	G	C
A	•							
G		•	•		•		•	
G		•	•		•		•	
C				•		•		•
G		•	•		•		•	
C				•		•		•
G		•	•		•		•	
C				•		•		•

	G	A	T	T	A	G
G	•					•
A		•			•	
T			•	•		
T			•	•		
A		•			•	
G	•					•

Repeats of a Single Sequence Symbol

A dot matrix analysis can also reveal the presence of repeats of the same sequence character many times.

- these repeats become apparent on the dot matrix as horizontal or vertical rows of dots, merging into rectangular or square patterns.

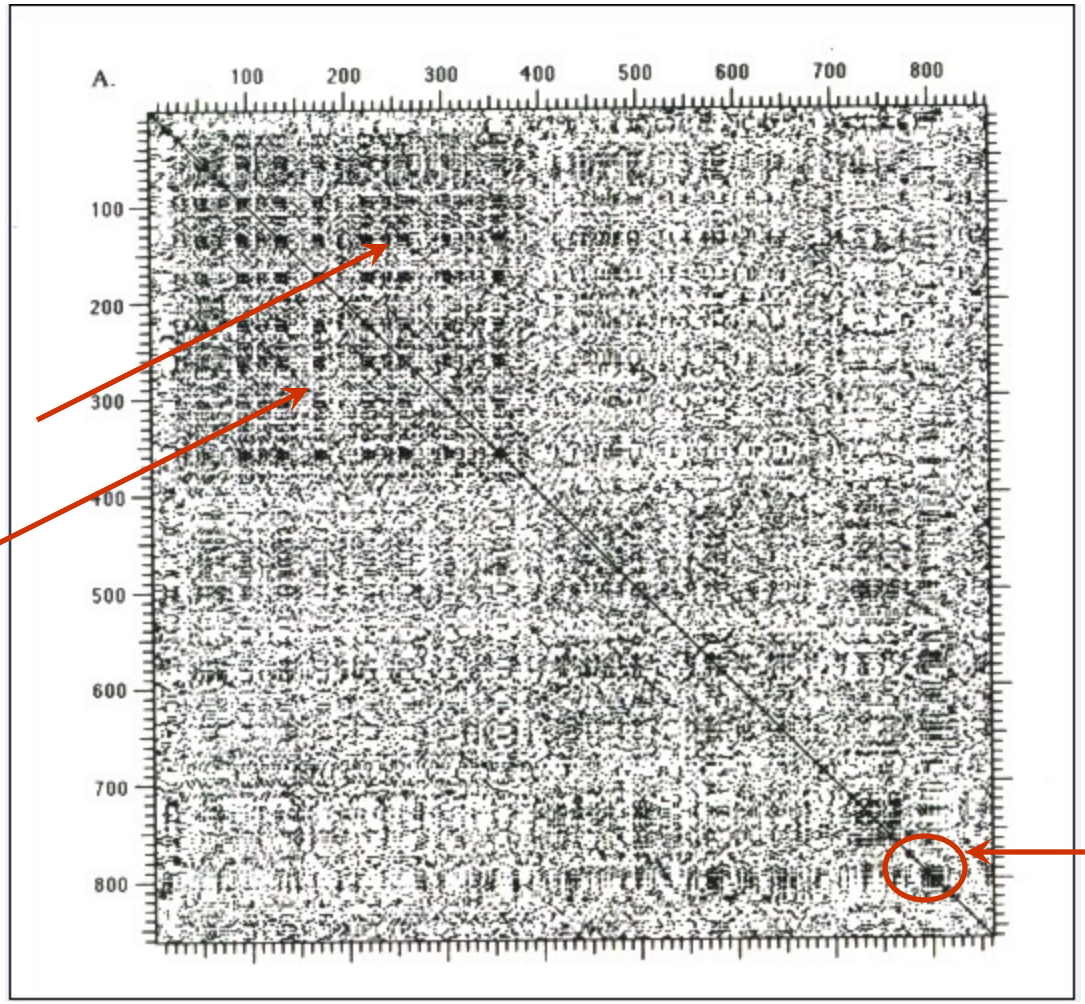
- as seen in the lower-right regions of the dot matrix of the human LDL receptor

Occurrence of such repeats of the same character increases the difficulty of aligning sequences as they create alignments with artificially high scores

- Mask these repeats during database searches

Programs: DUST (DNA), SEG (Protein)

Dot matrix analysis of the human LDL receptor against itself

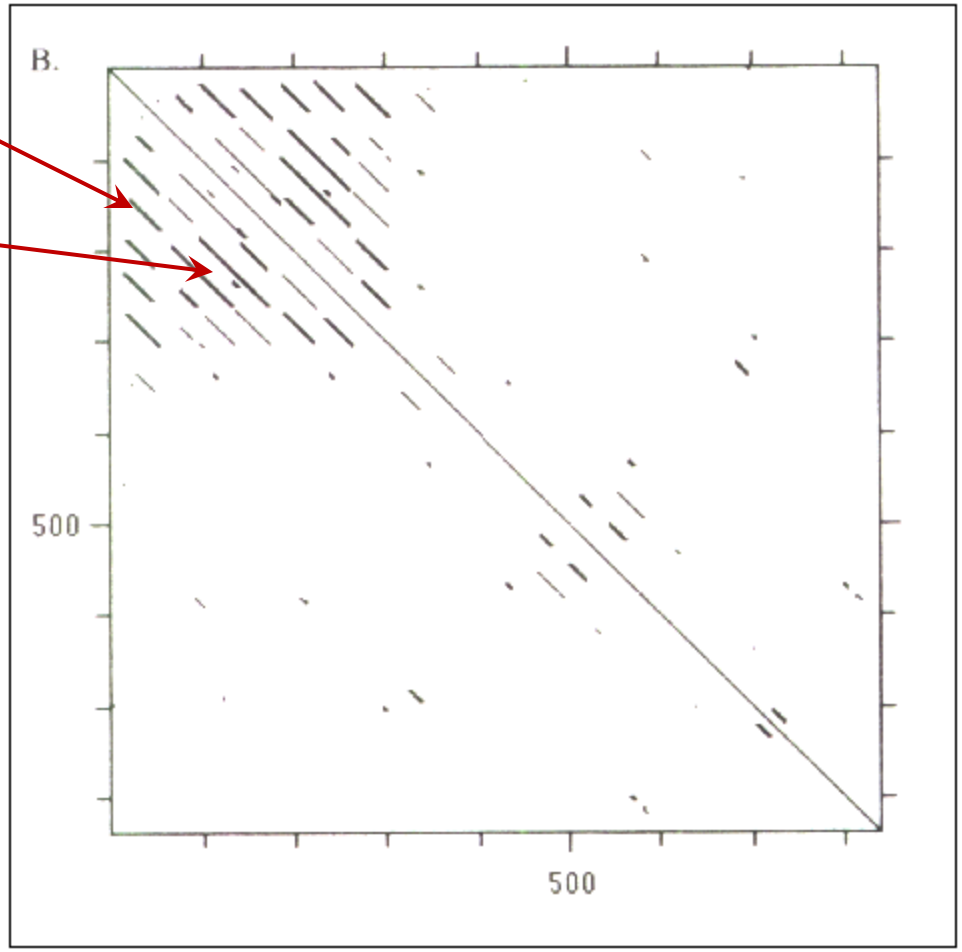


Window size: 1
Stringency: 1

Low-complexity region

Dot matrix analysis of the human LDL receptor against itself

Repeats of different lengths

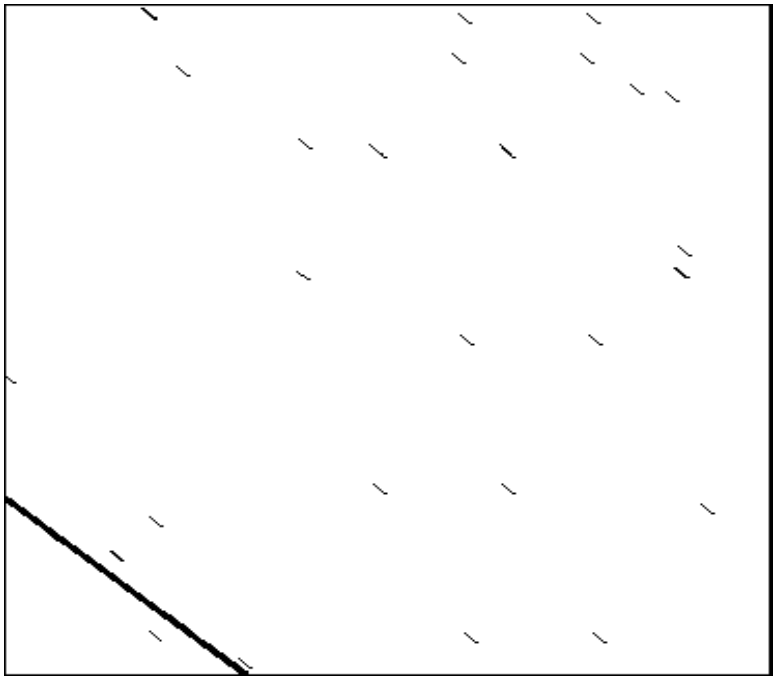


Window size: 23
Stringency: 7

Proteins composed of multiple copies of a single domain can be identified by dot plots

Dot Plots

Overlapping Sequences



When do we expect to find overlapping sequences?

Self-Complementary Regions in RNA Sequences

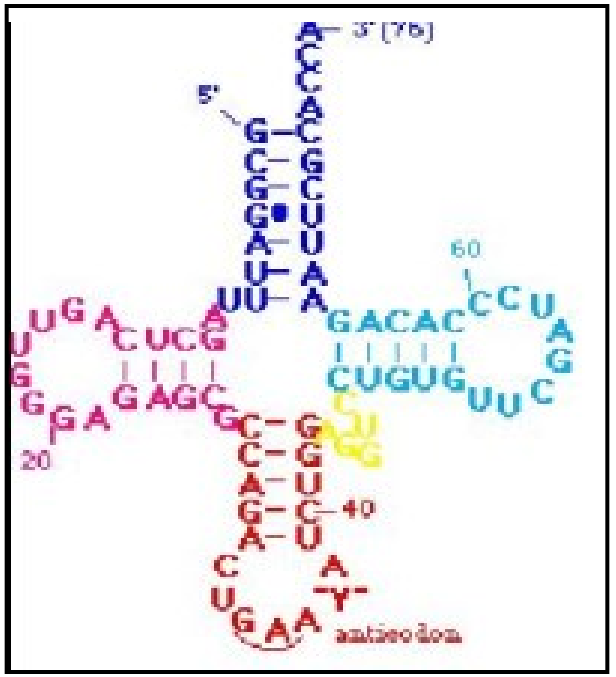
RNA secondary structure analysis begin with the identification of self-complementary regions

- these represent regions that can potentially self-hybridize to form RNA double strands

- once identified, the compatible regions may be used to predict a minimum free-energy structure.

- simplest way of identifying stretches of self-complementary regions in RNA sequence is a dot plot analysis

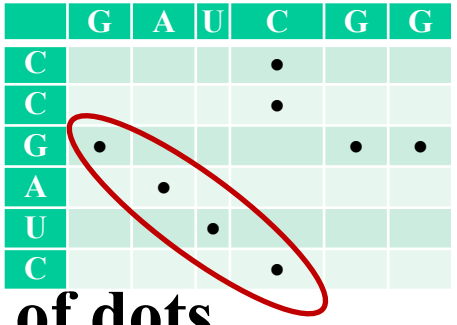
– there are two approaches.



Self-Complementary Regions in RNA Sequences

Method-1: Sequence is listed in 5′ to 3′ direction along the horizontal axis and its complementary sequence is listed along the vertical axis, also in the 5′ to 3′ direction.

Matrix is then scored for identities



Self-complementary regions appear as rows of dots going from upper left to lower right.

For RNA, these regions represent sequences that can potentially form A/U and G/C base pairs

- G/U base pairs not included in this simple analysis because they play a less significant role in base-pairing.

Self-Complementary Regions in RNA Sequences

As with matching DNA sequences, there are many random matches between the four bases in RNA, and the diagonals are difficult to visualize.

A long nucleotide window and a requirement for a large number of matches within this window are used to filter out the random matches.

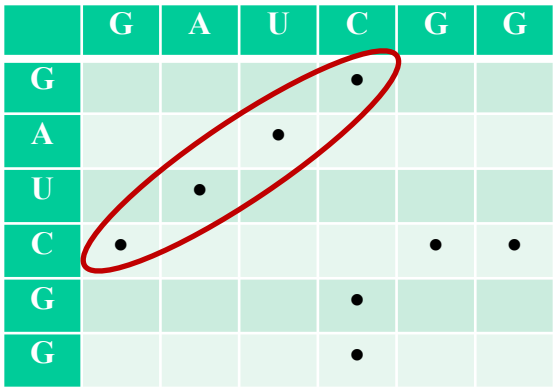
Self-Complementary Regions in RNA Sequences

Method-2: **Alternative approach - list the RNA sequence along the horizontal axis and also along the vertical axis,**

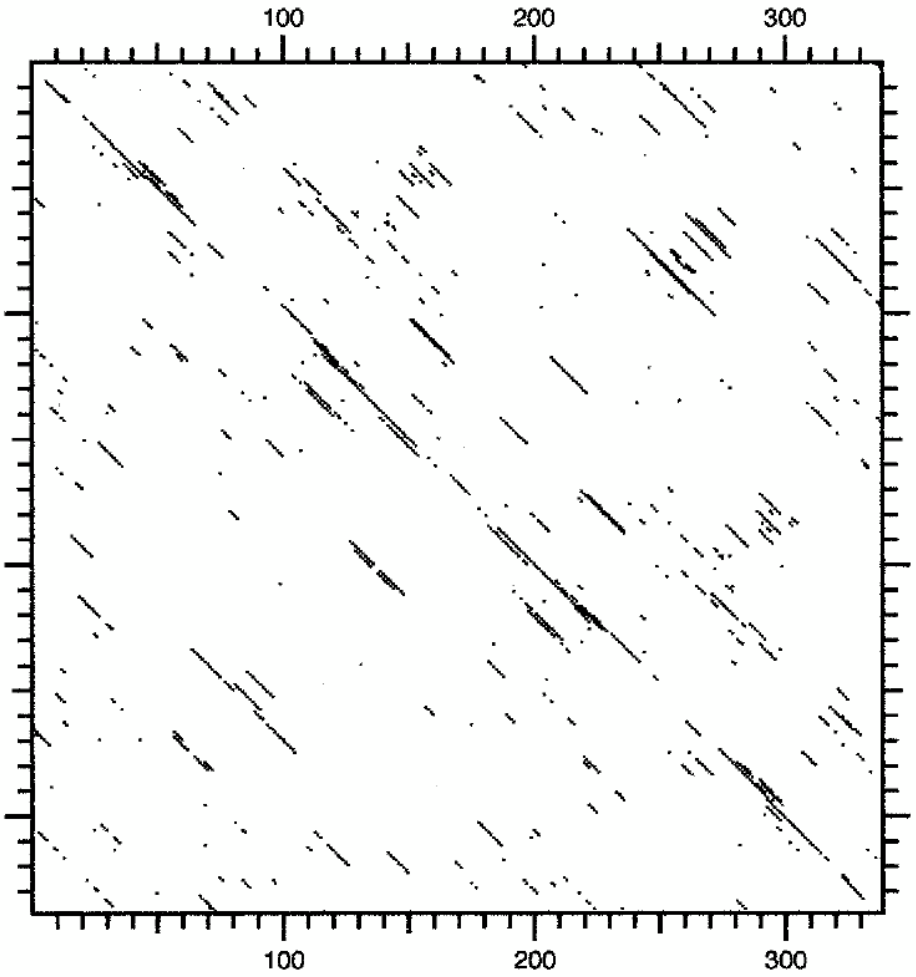
- Score matches of complementary bases G/C, A/U, and G/U instead of identities (as in the earlier method)

Diagonals indicating complementary regions will go from upper right to lower left in this matrix.

This type of matrix is used to produce an energy matrix for RNA secondary structure prediction.



Dot matrix Analysis of Potato Spindle Tuber Viroid for RNA Secondary Structure Analysis



Window: 15
Stringency: 11

Note: mirror image of diagonal from center to upper left and from center to lower right

Tools for Dot Plots

- **Dotter**
- **Dottup - EMBOSS (dotmatcher, dotpath, polydot)**
- **Diagon**
- **Compare & dotplot - GCG package**

EMBOSS

EMBOSS - European Molecular Biology Open Software Suite

- is a suite of free software tools for sequence analysis. It consists of a wide variety of programs ranging in application from database search to presentation of sequence data.

https://www.ebi.ac.uk/Tools/emboss/

dottup

EMBOSS dottup - displays a wordmatch dotplot of two sequences

It looks for places where words (tuples) of a specified length have an exact match in both sequences and draws a diagonal line over the position of these words.

Using a longer tuple size displays less random noise, runs extremely quickly, but is less sensitive.

Shorter word sizes are more sensitive to shorter or fragmentary regions of similarity, but also display more random points of similarity (noise) and runs slower

For what tasks is this program suitable?

dottup

For what tasks is dottup program suitable?

- **When comparing a cDNA sequence (mRNA sequence converted to double stranded DNA sequence) to the genomic sequence, we expect an exact match, and dottup is suitable in such situations.**
- **Comparing very closely related sequences, when we expect a large no. of exact matches.**

Other Dot Plot programs in EMBOSS:

- **dotmatcher – displays a threshold dotplot of 2 Seqs**

  - a sliding window analysis along the diagonal; displays a line over the window if the sum of the comparisons (using a substitution matrix) exceeds a threshold. It is slower but much more sensitive.

- **dotpath - Displays a non-overlapping wordmatch dotplot of two sequences**

  - suitable for moderately distant sequences, for multiple domains in a protein, etc.

- **polydot - Displays all-against-all dotplots of a set of sequences**

Difference between dottup and dotpath?

Assignment:

Find out the functionalities of the various dotplot programs in EMBOSS.

About EMBOSS

Overview • Uses • FAQ Citing EMBOSS

A high-quality package of free, Open Source software for molecular biology ... *more >*

Applications

EMBOSS • EMBASSY • Groups Proposed

Hundreds of useful, well documented applications for molecular sequence and other analyses ... *more >*

GUIs

Jemboss • GUIs • Web • Others

We support the Jemboss GUI but many others are available... *more >*

Servers

Portals • Servers • Mirrors • Misc

Many EMBOSS portals, servers and mirrors are available ... *more >*

Downloads

Stable release • Developers (CVS) version • Getting started

EMBOSS is open source software and is freely available to all ... *more >*

Licence

Licensing terms

EMBOSS uses the General Public Licence (GPL) and Library GPL ... *more >*

EMBOSS

vers

[sort alphabetically]

ALIGNMENT CONSENSUS

cons
megamerger
merger

ALIGNMENT DIFFERENCES

diffseq

ALIGNMENT DOT PLOTS

dotmatcher
dotpath
dottup
polydot

ALIGNMENT GLOBAL

alignwrap
est2genome
needle
stretcher

ALIGNMENT LOCAL

DOTTUP

(Displays a wordmatch dotplot of two sequences)

Help ?

Fields with a coloured background are optional and can safely be ignored...

[Hide optional fields]

1. SET THE PARAMETERS FOR THE RUN (OR ACCEPT THE DEFAULTS...)

input section

Select an input sequence.

Use one of the following three fields:

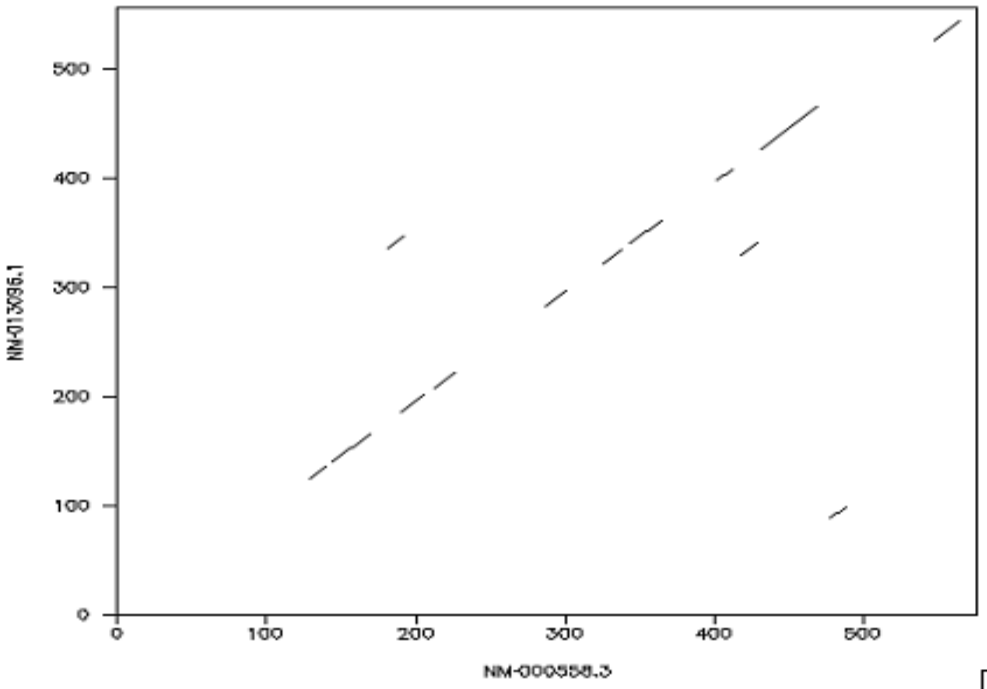
1. To access a sequence from a database, enter the USA path here: *(dbname:entry)*

2. Or, upload a sequence file from your local computer here:

Browse...

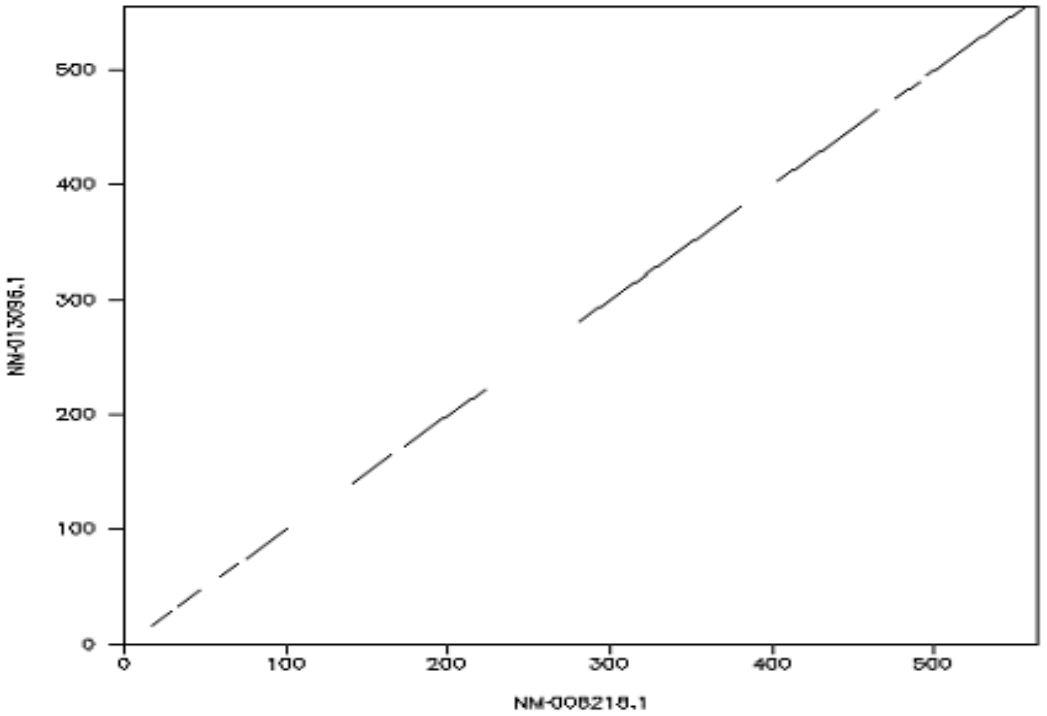
3. Or enter the sequence data manually here:

1. **>gi|14456711|ref|NM_000558.3| Homo sapiens hemoglobin, alpha 1 (HBA1), mRNA**
2. **>gi|6981009|ref|NM_013096.1| Rattus norvegicus hemoglobin alpha, adult chain 1 (Hba-a1), mRNA**
3. **>gi|6680174|ref|NM_008218.1| Mus musculus hemoglobin alpha, adult chain 1 (Hba-a1), mRNA**



Homo sapiens
vs
Mus Musculus

Mus Musculus
vs
Rattus norvegicus



Summarize

By analyzing the diagonal segments, dot plots can be used:

- **to find local regions of similarity, i.e., conserved and less conserved parts of homologous proteins**
  - as long diagonal lines
- **to identify domain homologies between proteins not homologous overall**
- **to identify overlapping sequences, e.g., in sequence assembly**
  - as a diagonal on a corner of the plot
- **to identify internal repeats and duplications**
  - as lines parallel to the diagonal
- **to identify insertions and deletions**
  - as breaks or discontinuities in the diagonal lines
- **to identify self-complementary regions**
  - in RNA secondary structure analysis

Summarize

- **For DNA sequence dot matrix comparisons, use long windows and high stringencies, e.g., 11 & 7, 15 & 11.**
- **For protein sequences, use short windows, e.g., 2 & 1 for window and stringency, respectively.**
- **When looking for a short domain of partial similarity in otherwise not-similar protein sequences, e.g. sharing similar active sites**

  - use a longer window and a small stringency, e.g., 15 & 5, for window and stringency, respectively.

Assignment

Q1. Obtain a dotplot when a sequence A is completely contained in sequence B.

Q2. Obtain a self-dotplot of the sequence:

ATGCGCGCTG

Sequence Alignment

Sequence alignment - a scheme of writing one sequence on top of another where the residues in one position are deemed to have a common evolutionary origin

If the same letter occurs in both sequences then this position has been conserved in evolution.

If the letters differ it is assumed that the two derive from an ancestral letter (could be one of the two or neither)

Comparison of Sequences

Sequence alignment of two sequences basically involves

- identifying regions of similarity, i.e., *conserved regions*, between them

- to find out if the two sequences are related or not

- enable us to extrapolate knowledge of the known sequence, or family, to the unknown query sequence

Any other reasons for Sequence Comparison?

Comparison of Sequences

Sequence alignment of two sequences basically involves

- identifying regions of similarity, i.e., *conserved regions*, between them

- to find out if the two sequences are related or not

- enable us to extrapolate knowledge of the known sequence, or family, to the unknown query sequence

- identifying species, evolutionary analysis

Statistical measures have been proposed to evaluate the significance of alignment, i.e.,

- decide whether the alignment is more likely to have occurred because they are related, or just by chance

Sequence Alignment

A letter or a stretch of letters may be paired up with dashes in the other sequence to signify an insertion or deletion event.

Since an insertion in one sequence can always be seen as a deletion in the other, one frequently uses the term "*indel*"

I

BANANA-
-ANANAS

Score: 10

BANANA
PANAMA

Score: 2

II

Sequence Alignment

Using a simple evolutionarily motivated scoring scheme, an alignment mediates the definition of a distance for two sequences:

Assign 0 to a match, some positive number (say, +1) to a mismatch and a larger positive number (say, +5) to an *indel*.

By adding these values along an alignment one obtains a score for this alignment:

BANANA-	BANANA
- ANANAS	PANAMA
Score: 10	Score: 2

Sequence Alignment

A distance function for two sequences can be defined by looking for the alignment which yields the *minimum score*

Using dynamic programming this minimization can be effected without explicitly enumerating all possible alignment of two sequences.

The idea of assigning a score to an alignment and then minimizing over all alignments is at the heart of all biological sequence alignments.

Sequence Alignment

Note: one may either define a distance or a similarity function to an alignment.

- difference lies mainly in the interpretation of the values

A distance function defines 0 for a match and positive values for mismatches or gaps, and then aims at minimizing this distance

A similarity function assigns high positive values to matches and negative values to mismatches and gaps, and then maximize the resulting score.

Basic structure of the algorithm is the same for both cases.

When would you use a distance function and a similarity function for scoring an alignment?

Sequence Alignment

Thus, an alignment is:

- **a mutual arrangement of two sequences**
- **It exhibits where the two sequences are similar, and where they differ**
- **An 'optimal' alignment is one that exhibits the most correspondences, and the least differences**
- **‘Optimal' alignment need not reflect the true evolutionary relationship between two sequences, though it usually does**

Similarity ⇒ Homology

Why is this not true?

Sequence Alignment

Differences between similarity and homology:

- **Similarity is simply a measure of expression how alike two sequences are**
- **Homology means there is an evolutionary relationship between two sequences - there are no degrees of homology.**
- **Extending this to individual residues they are 'identical' or 'similar' residues - similar implies that they share certain physicochemical properties**
- **Homology cannot be observed, it is only an inference**

Differences between similarity and homology

Identical protein sequences result in identical 3-D structures - similar sequences may result in similar structures, and this is usually the case.

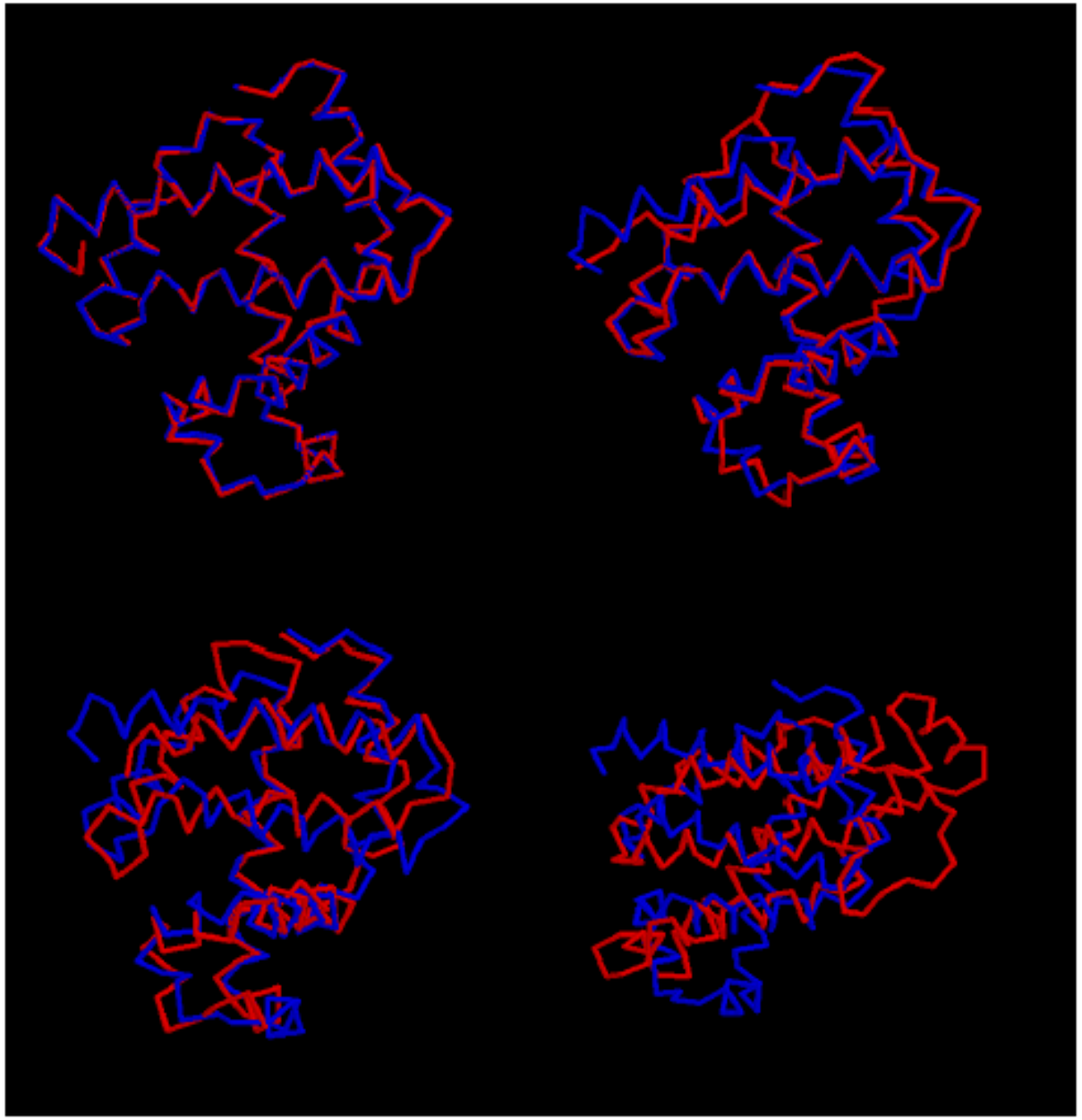
The converse is not true: identical 3-D structures do not necessarily indicate identical sequences. It is because of this that there is a distinction between "homology" and "similarity".

There are examples of proteins in the databases that have nearly identical 3-D structures, and are therefore homologous, but do not exhibit significant (or detectable) sequence similarity

Sequence identity and rmsd of Sperm Whale myoglobin

myoglobin
pig

rmsd = 0.5 Å
id = 86%



haemoglobin
pig

rmsd = 1.5 Å
id = 28%

globin-3
P. piclitum

rmsd = 2.2 Å
id = 18%

phycocyanin
F. diplosiphon

rmsd = 3.3 Å
id = 8%

Summarize

- Comparison of an unknown sequence to an annotated sequence permits us to infer structural, functional & evolutionary relationships
- Wherever possible use the protein sequence since this confers more information
- Substitutions, deletions and insertions all occur as part of the natural evolutionary process
- Homology implies an evolutionary relationship between two sequences