Lecture Notes

Big Data in Medical Informatics

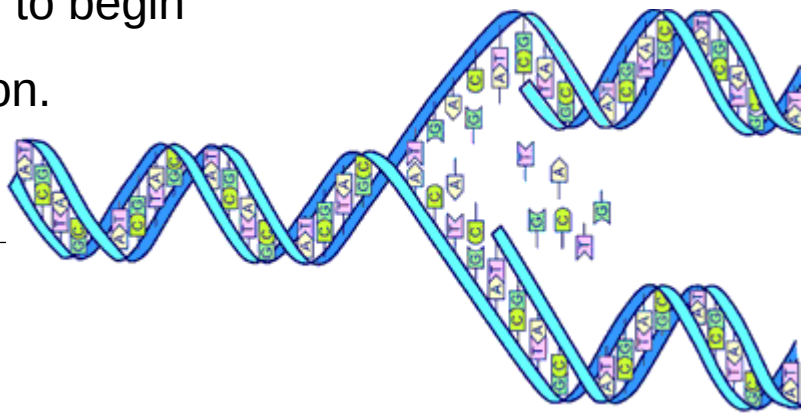
Week 12:

Introduction to Basics of Genomic

RWTH AACHEN UNIVERSITY

The Human Genome Project (HGP)

- The Human Genome Project (HGP) was an international effort started in 1990 to sequence all three billion base pairs that make up human DNA
- The Human Genome Project originally aimed to map the nucleotides contained in a human reference genome (more than three billion).
- The "genome" of any given individual is unique;
- mapping the "human genome" involved sequencing a small number of individuals and then assembling these together to get a complete sequence for each chromosome.
- The finished human genome is thus a mosaic, not representing any one individual.
- The work of the HGP has allowed researchers to begin to understand the blueprint for building a person.



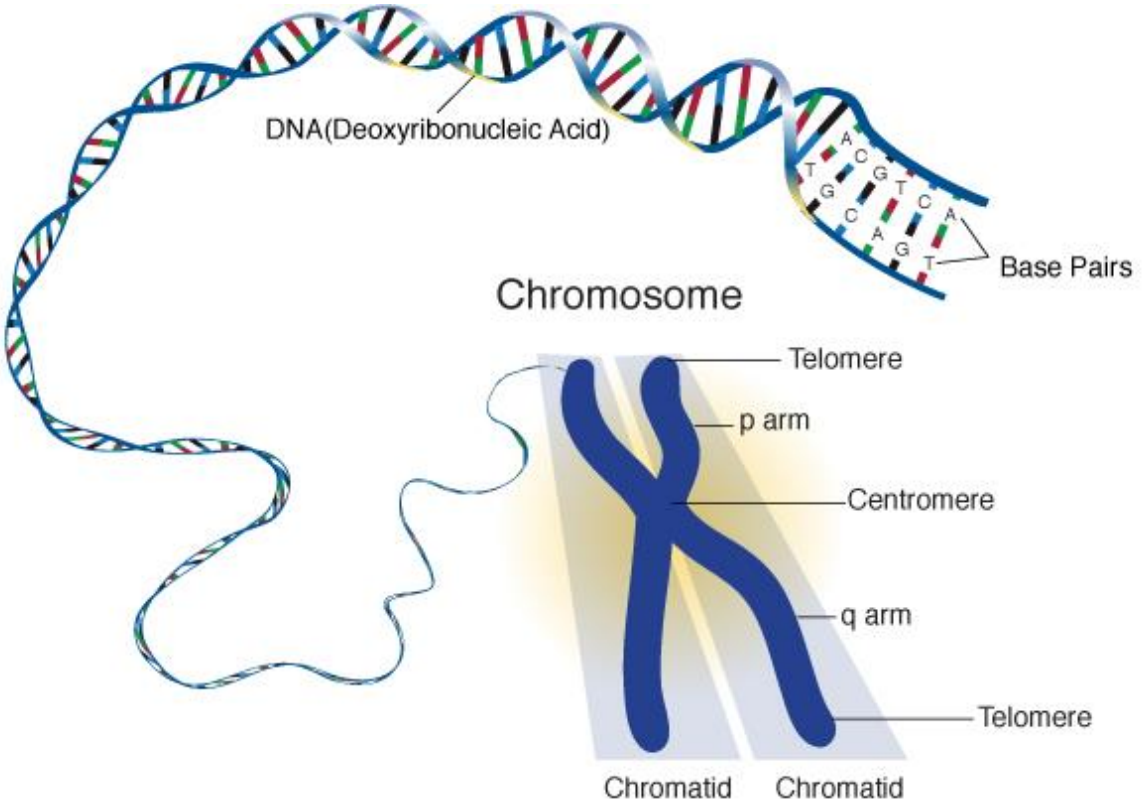
RWTH Informatik 5 | Ahornstr. 55 D-52056 Aachen Germany |
Tel +49/241/8021501 | Fax +49/241/8022321 | http://dbis.rwth-aachen.de/cms

Concepts

- Nucleotides are organic molecules that serve as the monomers, or subunits, of nucleic acids like DNA (deoxyribonucleic acid) and RNA (ribonucleic acid)
- A genome is an organism's complete set of deoxyribonucleic acid (DNA), a chemical compound that contains the genetic instructions needed to develop and direct the activities of every organism.
- The human genome contains approximately 3 billion of these base pairs, which reside in the 23 pairs of chromosomes within the nucleus of all our cells.
- Each of the estimated 30,000 genes in the human genome makes an average of three proteins.
- Sequencing means determining the exact order of the base pairs in a segment of DNA. Human chromosomes range in size from about 50,000,000 to 300,000,000 base pairs

Chromosome

- A chromosome is an organized package of DNA found in the nucleus of the cell.
- Different organisms have different numbers of chromosomes.
- Humans have 23 pairs of chromosomes--22 pairs of numbered chromosomes, called autosomes, and one pair of sex chromosomes, X and Y.

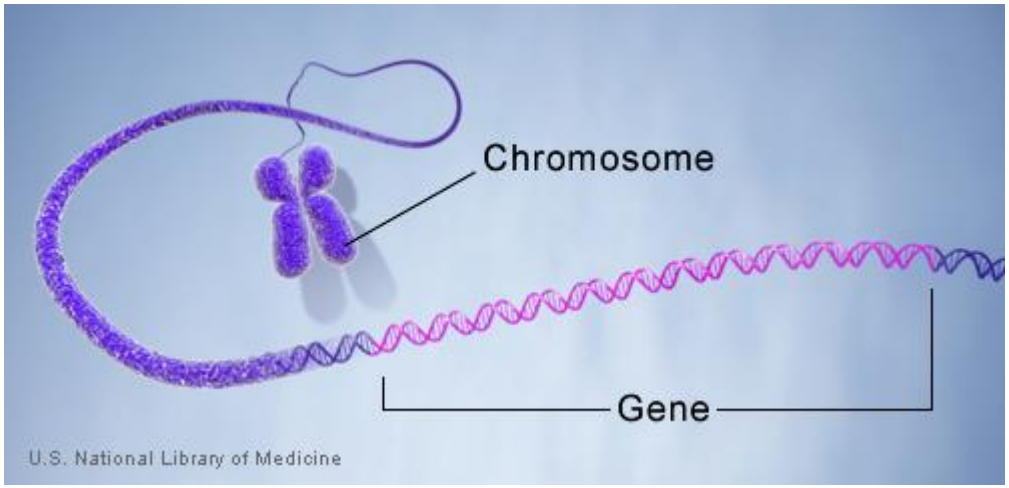


RWTH Informatik 5 | Ahornstr. 55 D-52056 Aachen Germany |
Tel +49/241/8021501 | Fax +49/241/8022321 | http://dbis.rwth-aachen.de/cms

Gene

- This DNA carries the genetic blueprint that is used to make all the proteins the cell needs.
- Genes are subunits of DNA, the information database of a cell that is contained inside the cell nucleus.
- Genes, which are made up of DNA, act as instructions to make molecules called proteins.
- Every gene contains a particular set of instructions that code for a specific protein.

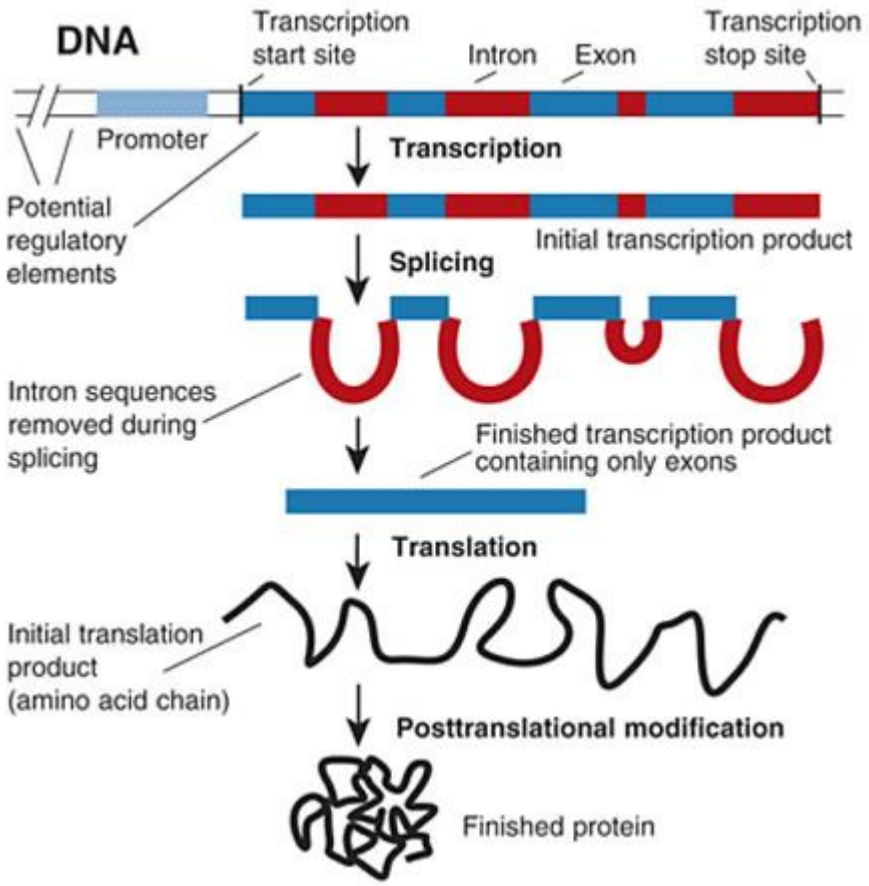
- In humans, genes vary in size from a few hundred DNA bases to more than 2 million bases.
- The Human Genome Project has estimated that humans have between 20,000 and 25,000 genes.



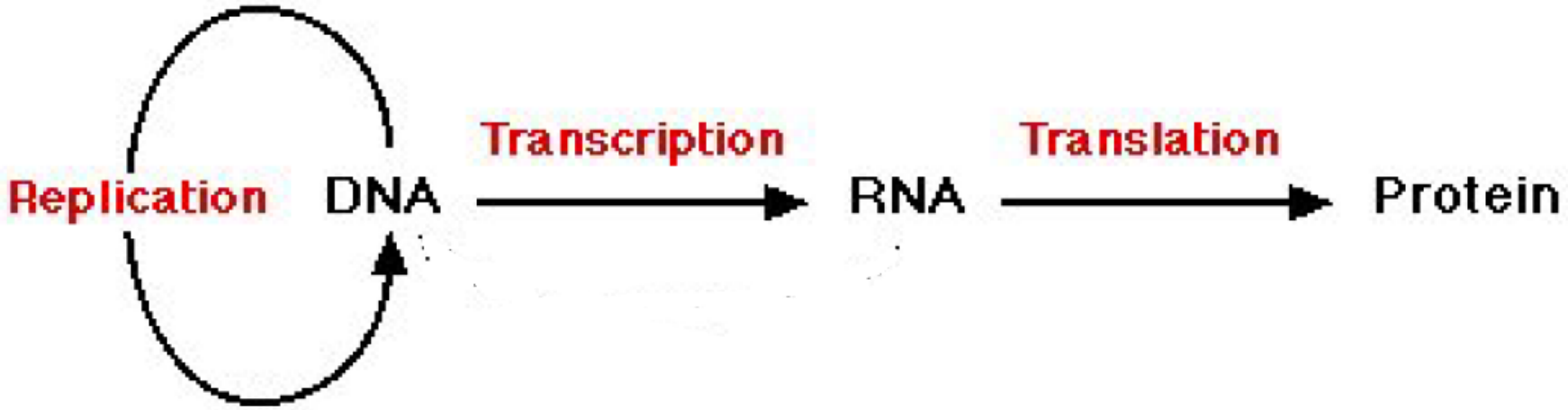
RWTH Informatik 5 | Ahornstr. 55 D-52056 Aachen Germany |
Tel +49/241/8021501 | Fax +49/241/8022321 | http://dbis.rwth-aachen.de/cms

Gene Expression

- Gene expression is the process by which genetic instructions are used to synthesize gene products.
- These products are usually proteins, which go on to perform essential functions as enzymes, hormones and receptors
- Gene expression is a sequence of subcellular complex reactions aiming to convert inherited data (i.e. gene) into functional chemical molecules



Central Dogma of Molecular Biology



Transcription is carried out by **RNA polymerase**

Translation is performed on **ribosomes**

Replication is carried out by **DNA polymerase**

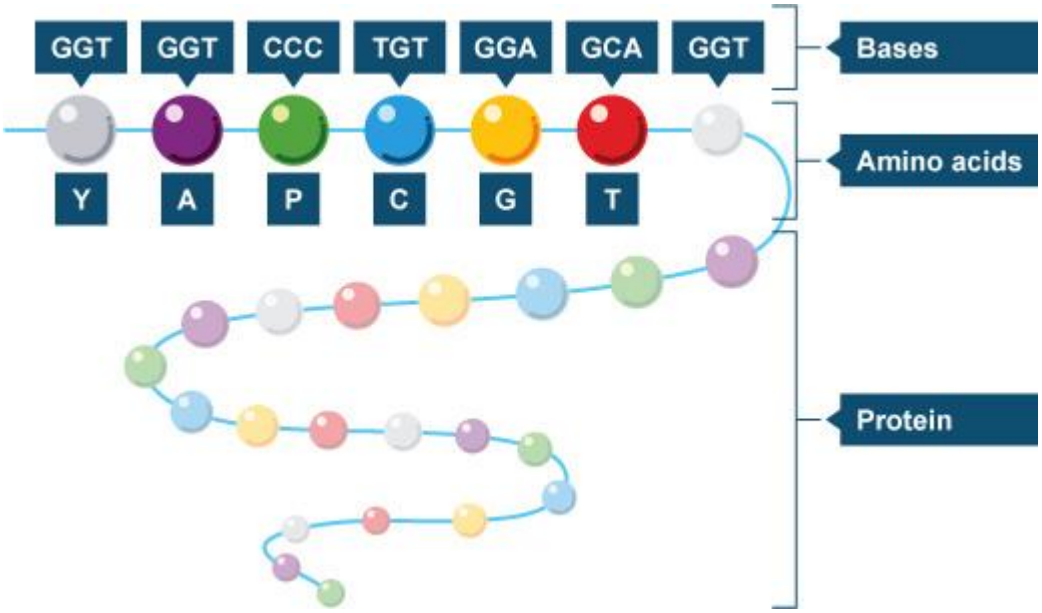
Solid arrow indicate types of information transfers that occur in cells.

- DNA directs its own replication to produce new DNA molecule;
- DNA is transcribes into RNA;
- RNA is translated into protein.

Proteins

- Proteins are large, complex molecules that are critical for the normal functioning of the human body.

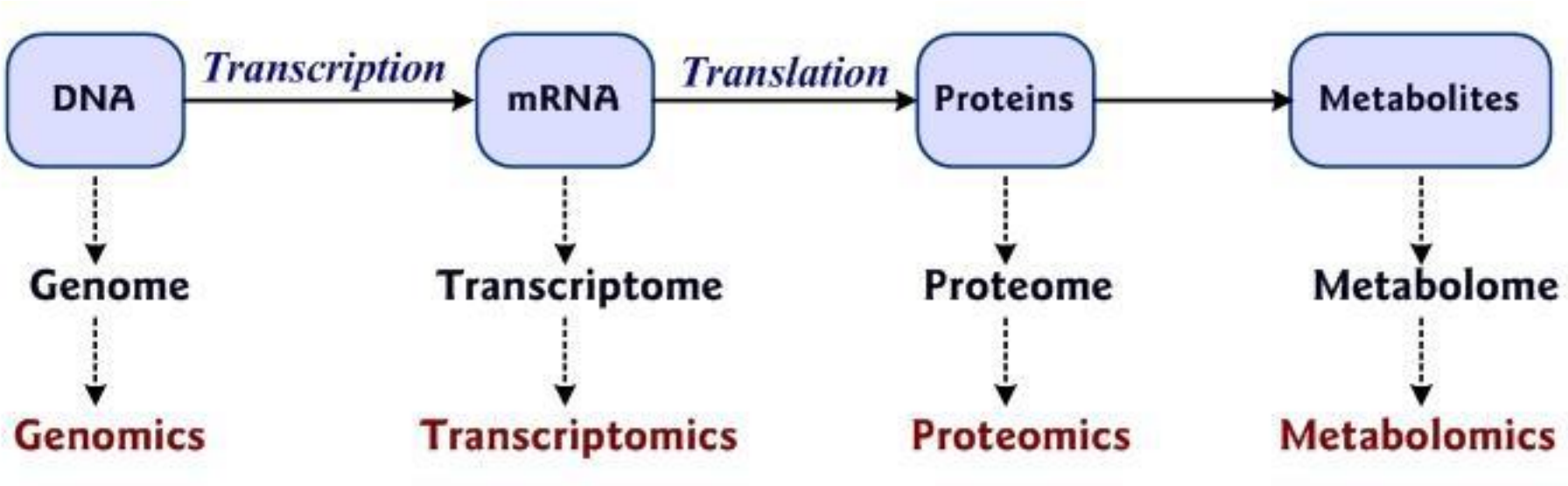
- They are essential for the structure, function, and regulation of the body's tissues and organs.
- Proteins are made up of hundreds of smaller units called amino acids that are attached to one another by peptide bonds, forming a long chain.



RWTH Informatik 5 | Ahornstr. 55 D-52056 Aachen Germany |
Tel +49/241/8021501 | Fax +49/241/8022321 | http://dbis.rwth-aachen.de/cms

Gene Expression

- Conceptual Representation of Gene Expression and Corresponding Genomics Data

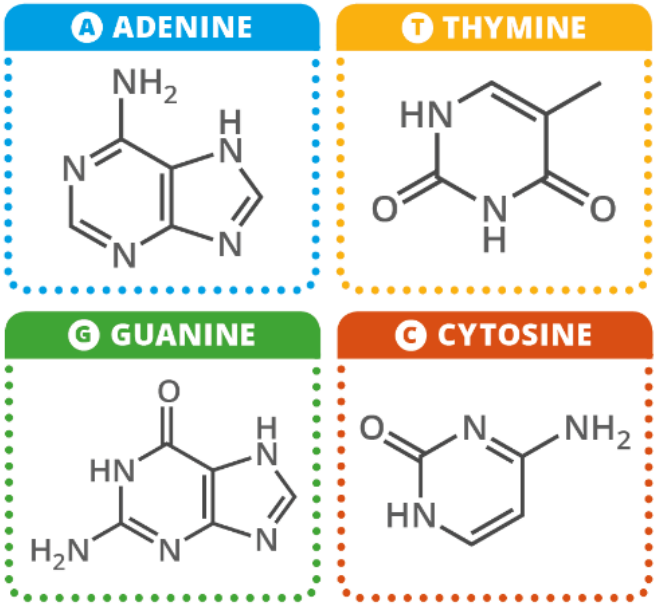


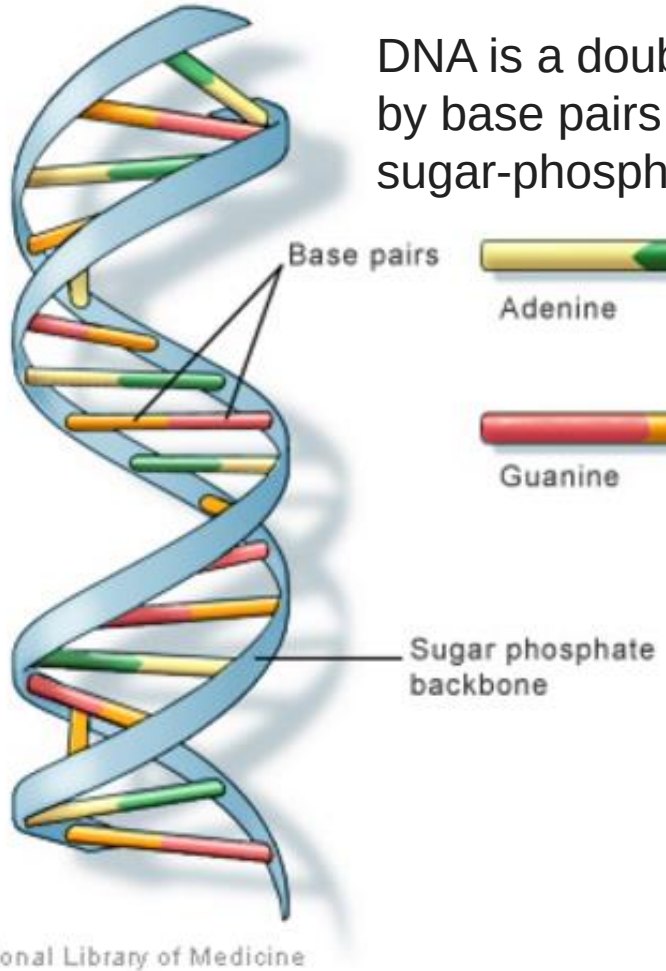
- Gene expression is a sequence of subcellular complex reactions aiming to convert inherited data (i.e. gene) into functional chemical molecules

The Chemical Structure of DNA

DNA is a polymer made of units called nucleotides.
Nucleotides made up three different components:
a sugar group , a phosphate group and a base.

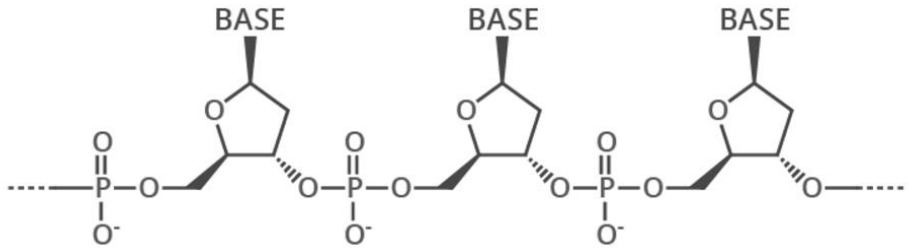
There are four different bases:





DNA is a double helix formed by base pairs attached to a sugar-phosphate backbone

The sugar-phosphate backbone



DNA Replication

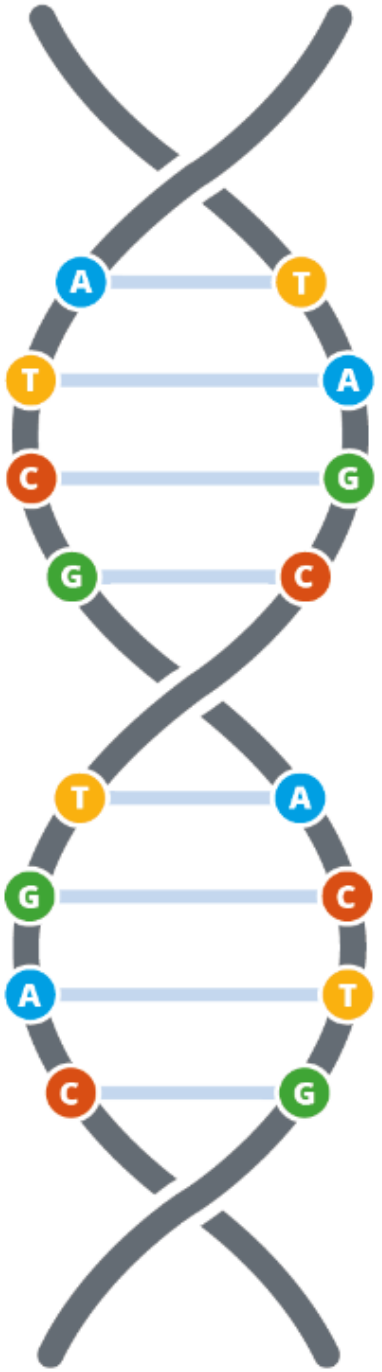
- DNA Can Make Copies of Itself

- Why?
- The cells in your body constantly divide, regenerate, and die, but for this process to occur The DNA within the cell must be able to replicate itself.

- DNA replication is an anabolic polymerization process, that allows a cell to pass copies of its genome to its descendants.

- The key to DNA replication is the complementary structure of the two strands:

- Nucleotides pair in a specific way - called the Base-Pair Rule
  - Adenine pairs to Thymine
  - Guainine pairs to Cytosin

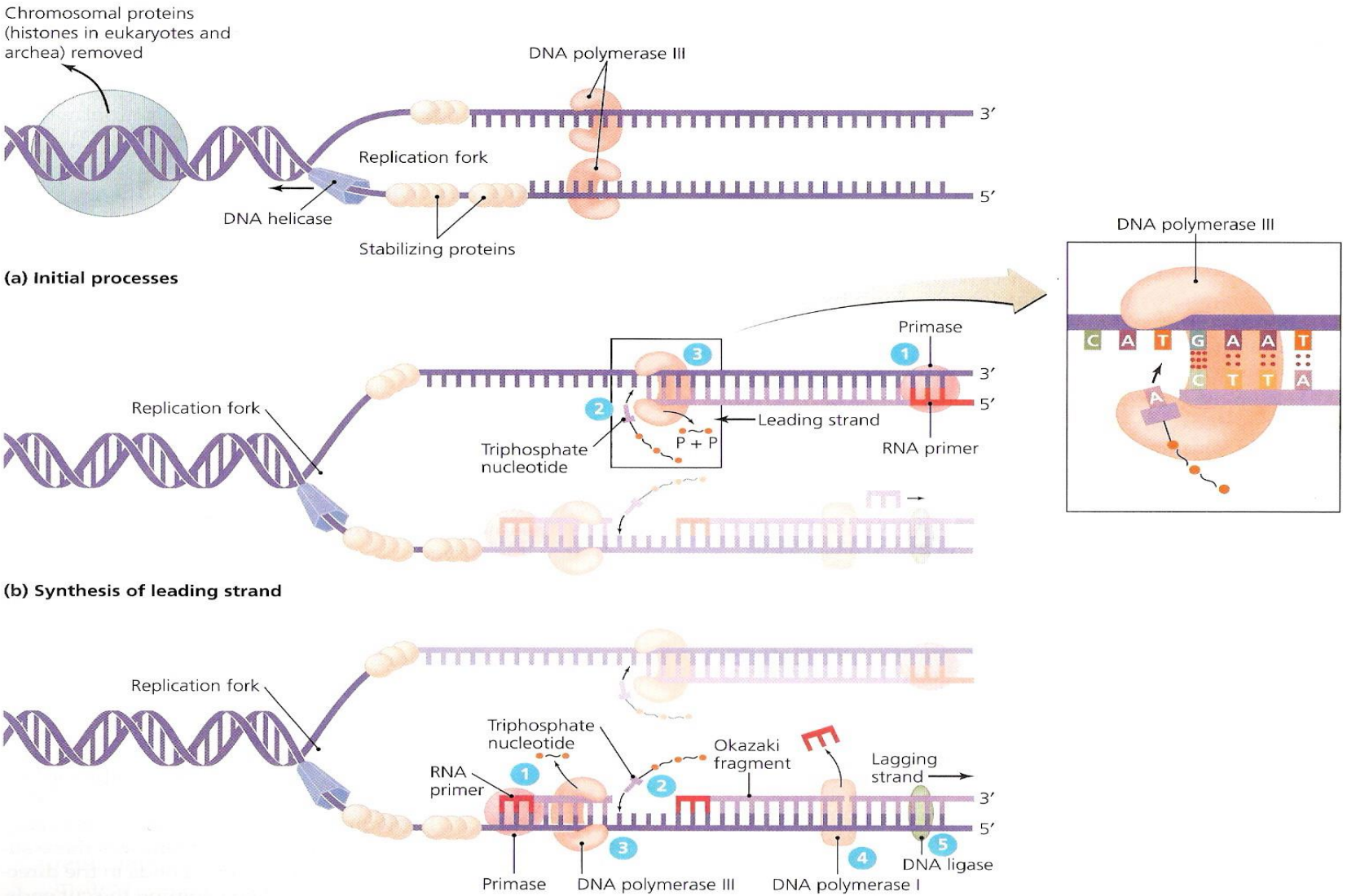


RWTH Informatik 5 | Ahornstr. 55 D-52056 Aachen Germany |
Tel +49/241/8021501 | Fax +49/241/8022321 | http://dbis.rwth-aachen.de/cms

DNA Replication

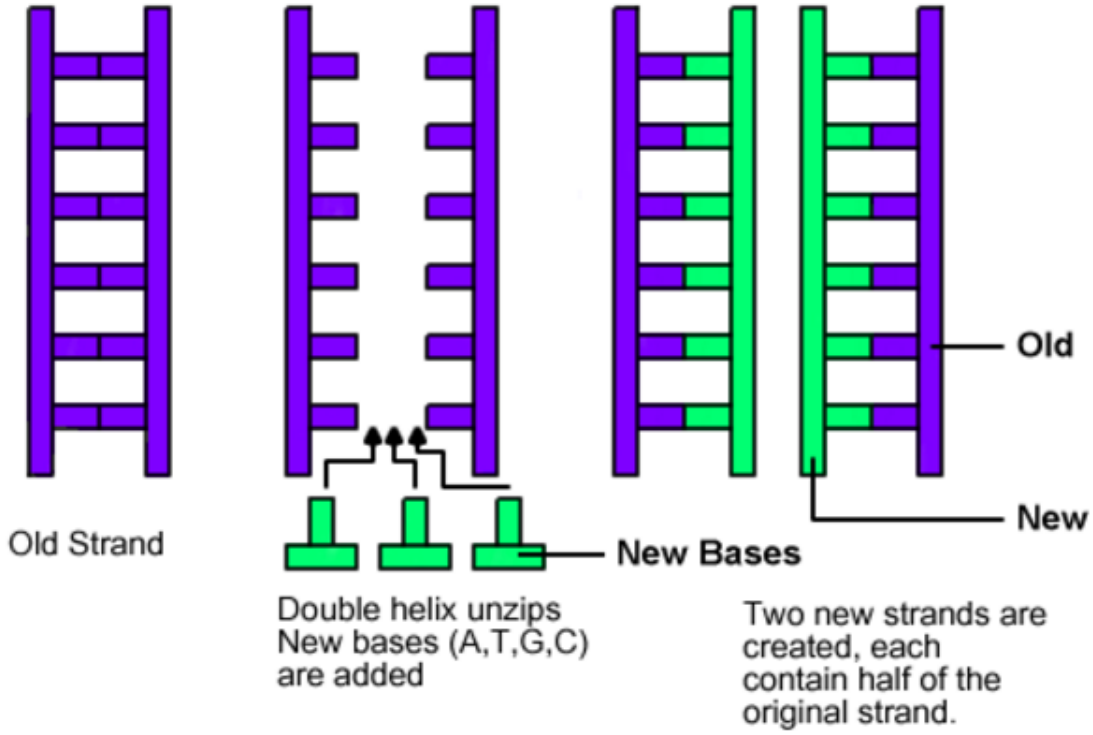
- Each double stranded DNA molecule holds the same genetic information.
- A cell separates the two original strands and uses each as a template for the synthesis of a new complementary strand.
- DNA replication begins at a specific sequence of nucleotides called an *origin*.
- First, a cell removes chromosomal proteins, exposing the DNA helix.
- Next, an enzyme called *DNA helicase* locally "unzips" the DNA molecule by breaking the hydrogen bonds between complementary nucleotide bases, which exposes the bases in a *replication fork*.
- Other protein molecules stabilize the single strands so that they do not rejoin while replication proceeds
- After helicase untwists and separates the strands, a molecule of an enzyme called *DNA polymerase* III binds to each strand.
- DNA polymerases replicate DNA in only one direction - 5' to 3' -

DNA Replication



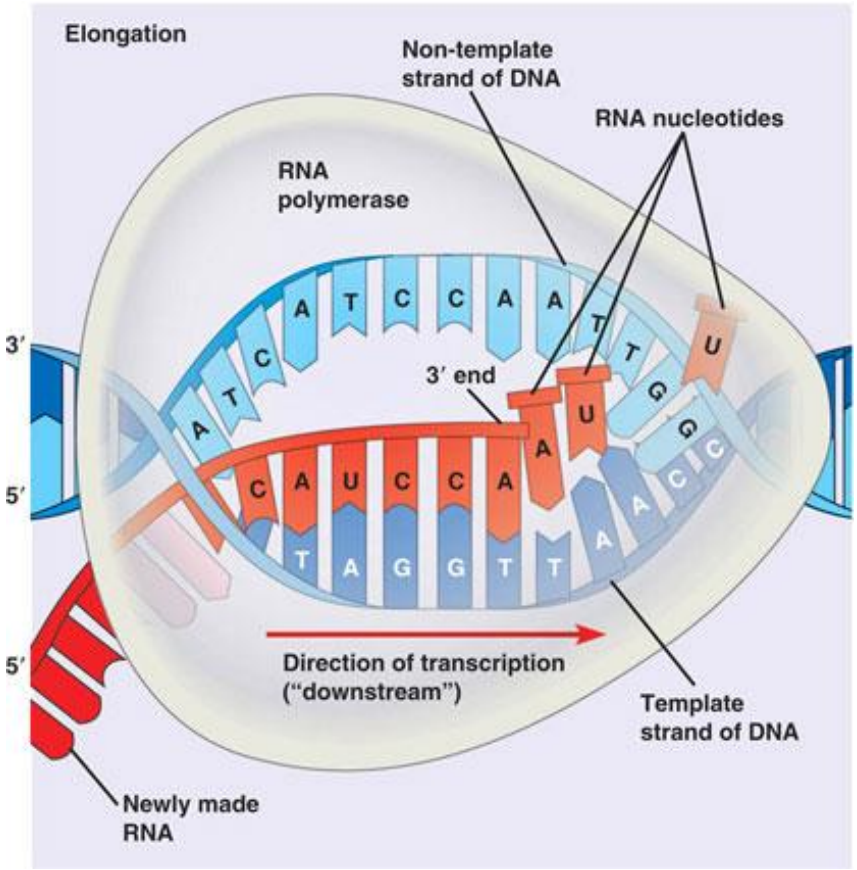
DNA Replication

- The template (original) strands are separated and preserved, while the new strands are assembled from nucleotides.
- This is called semi-conservative replication, since the each of the two resulting DNA molecules consist of one conserved old strand and one brand new strand.
- DNA replication is **semi-conservative**. That means that when it makes a copy, one half of the old strand is always kept in the new strand. This helps reduce the number of copy errors



Transcription

- During gene expression, the info in DNA is first transcribed as mRNA and then translated via tRNA and used to build a protein.
- In the transcription phase one strand of DNA molecule is copied into a complementary pre-mRNA (or nuclear RNA).
- During this process the two-stranded DNA double helix is unwound and information is read only from one strand.



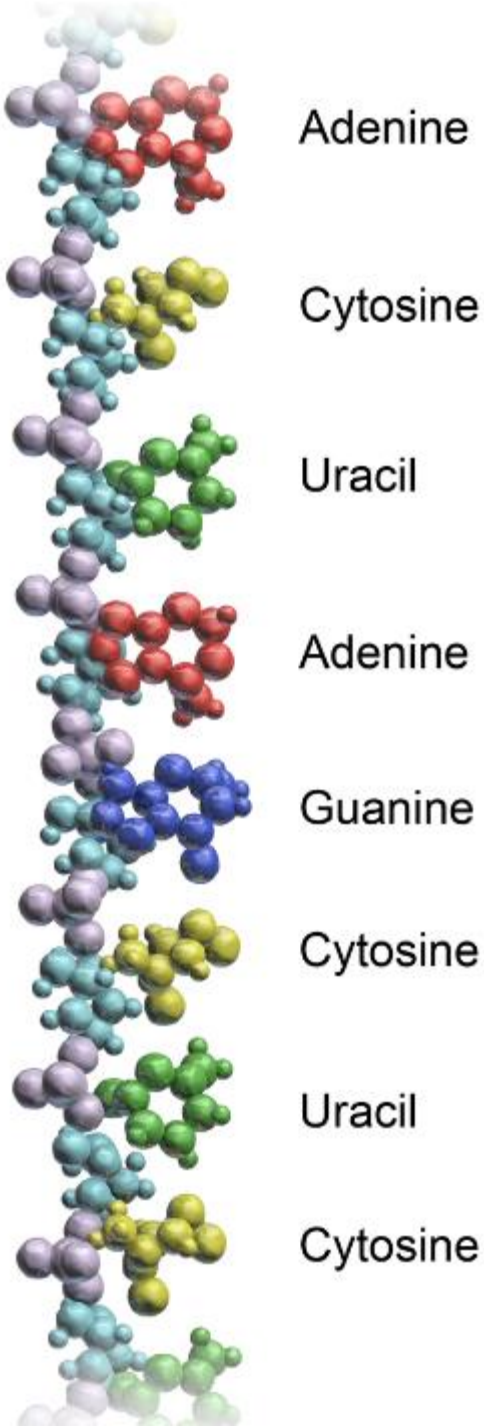
RWTH Informatik 5 | Ahornstr. 55 D-52056 Aachen Germany |
Tel +49/241/8021501 | Fax +49/241/8022321 | http://dbis.rwth-aachen.de/cms

RWTH AACHEN UNIVERSITY

RNA

- Ribonucleic acid (RNA) is a molecule similar to DNA.
- RNA is constructed from nucleotides, but instead of the Thymine (T), it has a similar molecule, Uracil (U), which is not found in DNA.
- Because of this difference RNA does not form a double helix, instead they are usually single stranded, but may have complex spatial structure due to complementary links between the parts of the same strand.
- RNA has different functions in the cell. Mainly, we are interested in its role as an intermediate between DNA and proteins.
- 3 forms of RNA
  - mRNA (messenger)
  - tRNA (transfer)
  - rRNA (ribosomal)

- RNA is used to take the information in DNA and make proteins (gene expression)



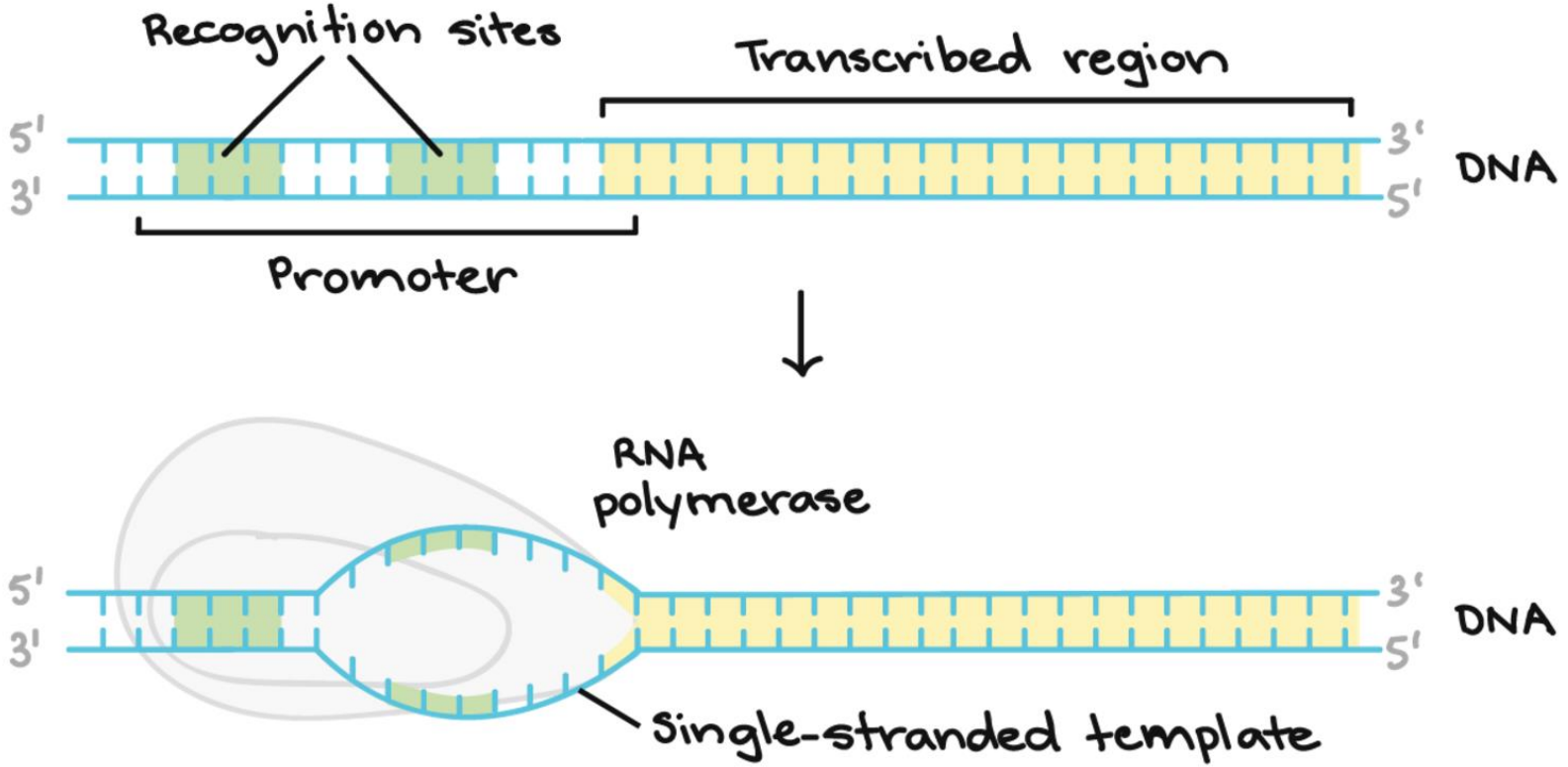
RWTH Informatik 5 | Ahornstr. 55 D-52056 Aachen Germany |
Tel +49/241/8021501 | Fax +49/241/8022321 | http://dbis.rwth-aachen.de/cms

Transcription

- Three main steps to the process of DNA transcription:

1. RNA Polymerase Binds to DNA

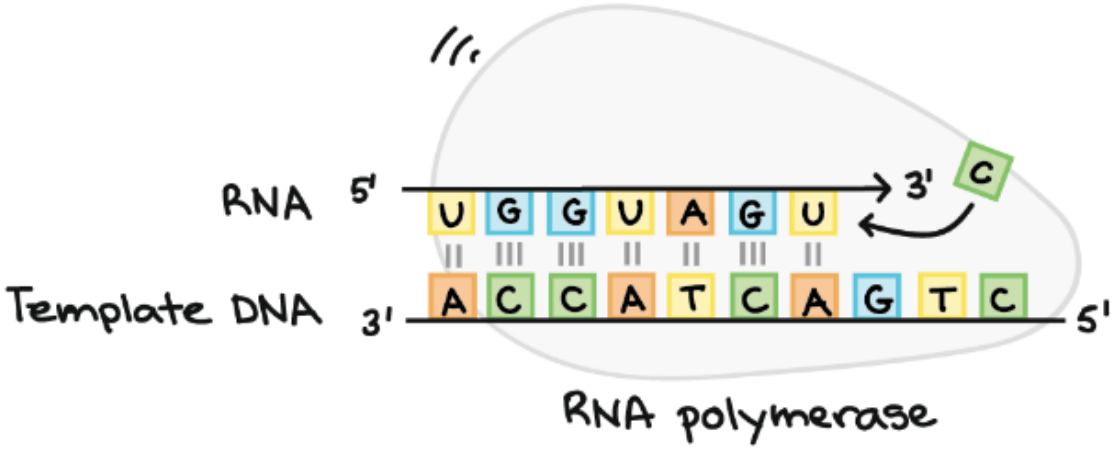
- DNA is transcribed by an enzyme called RNA polymerase.
- Specific nucleotide sequences tell RNA polymerase where to begin & where to end.
- RNA polymerase attaches to the DNA at a specific area called the promoter region.
- The DNA in the promoter region contains specific sequences that allow RNA polymerase to bind to the DNA



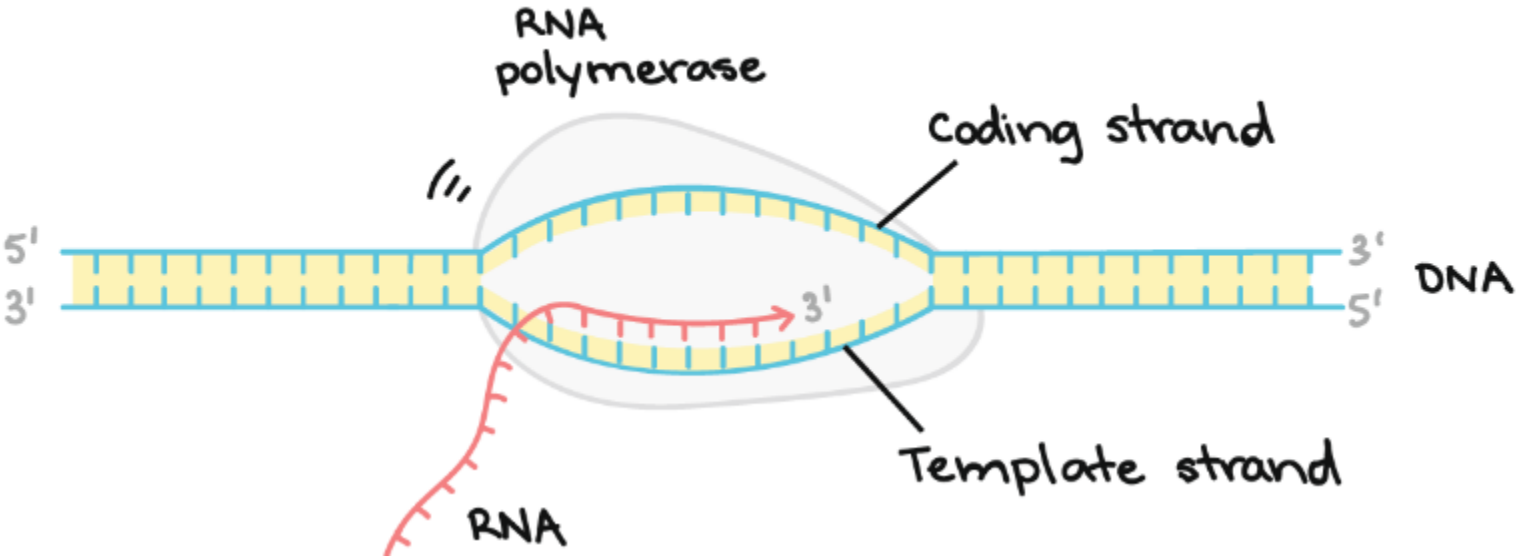
Transcription

2. Elongation

- Certain enzymes called transcription factors unwind the DNA strand and allow RNA polymerase to transcribe only a single strand of DNA into a single stranded RNA polymer called messenger RNA (mRNA).



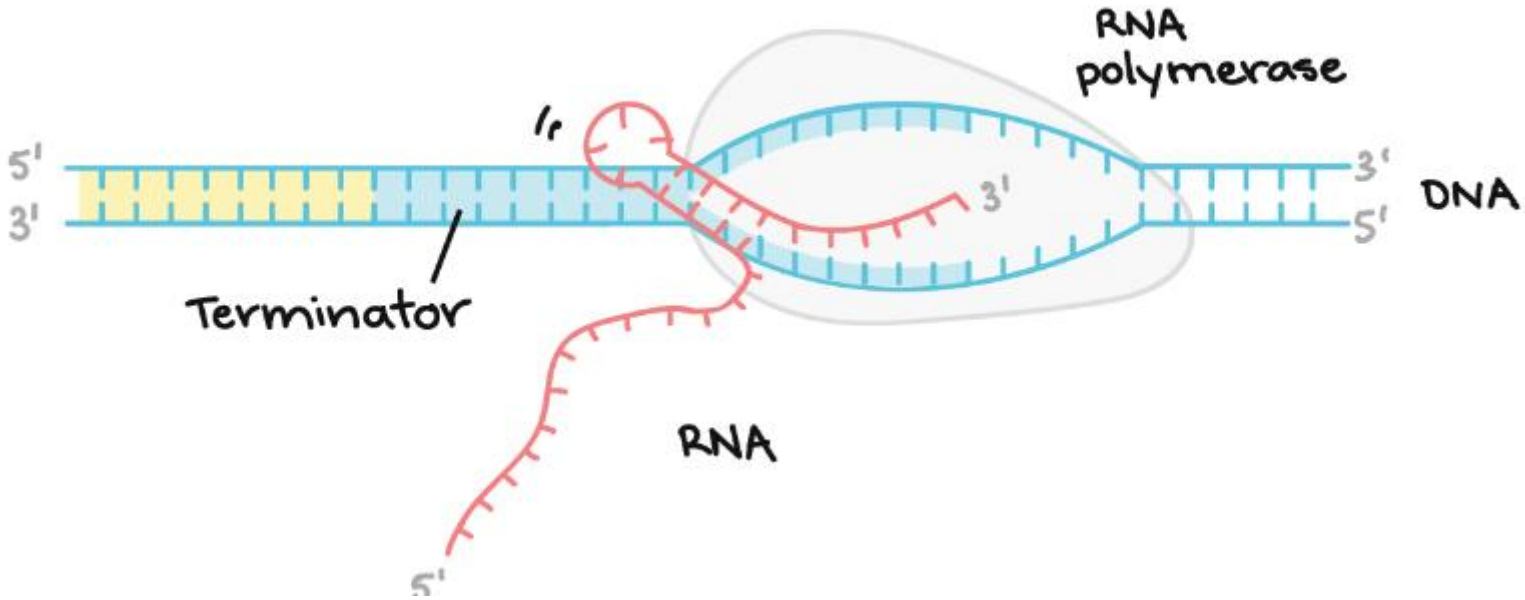
RNA polymerase transcribes the DNA, guanine pairs with cytosine (G-C) and adenine pairs with uracil (A-U)



Transcription

3. Termination

- RNA polymerase moves along the DNA until it reaches a terminator sequence.
- At that point, RNA polymerase releases the mRNA polymer and detaches from the DNA.
- An example of a termination mechanism involving formation of a hairpin in the RNA is shown below.

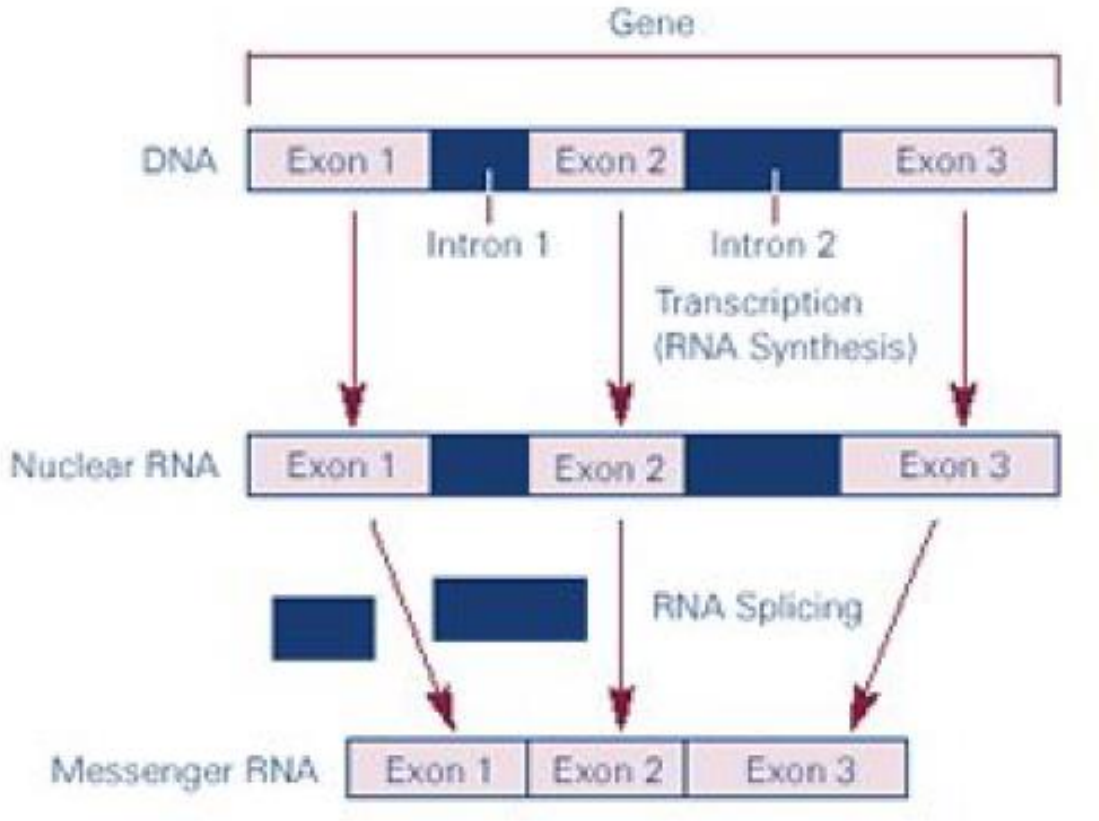


RWTH Informatik 5 | Ahornstr. 55 D-52056 Aachen Germany |
Tel +49/241/8021501 | Fax +49/241/8022321 | http://dbis.rwth-aachen.de/cms

Splicing

- Splicing removes some stretches of the pre mRNA, called introns.
- The remaining sections, called exons, are then joined together.

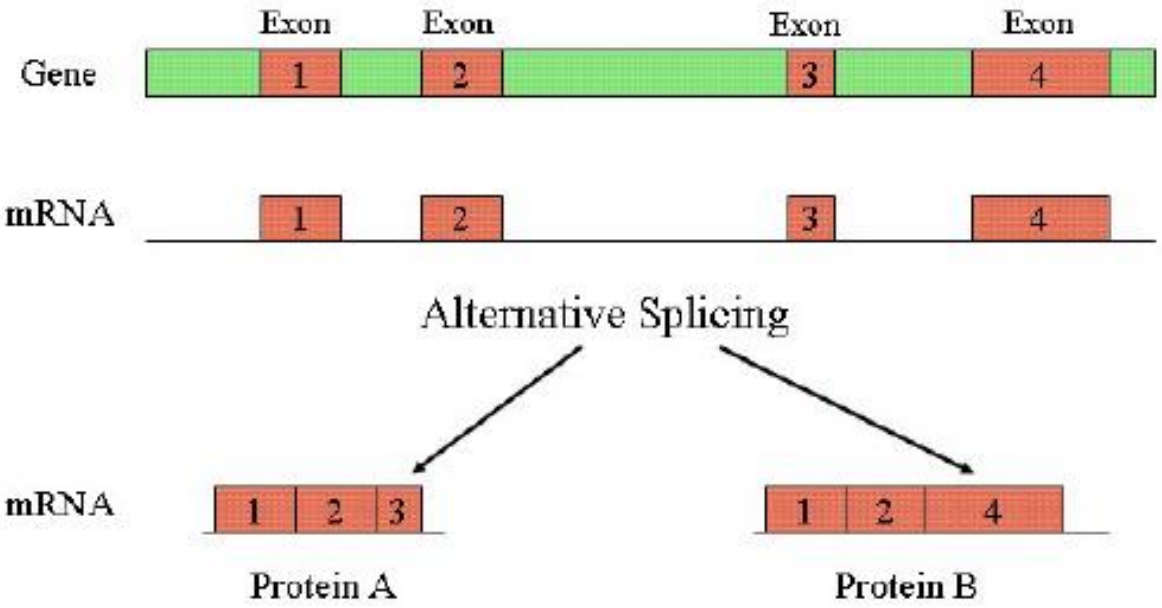
- Exons are the part of the gene that code for proteins and they are interspersed with non coding introns which must be removed by splicing.
- The number and size of introns and exons differs considerably among genes and also between species.
- The result of splicing is mRNA.



RWTH Informatik 5 | Ahornstr. 55 D-52056 Aachen Germany |
Tel +49/241/8021501 | Fax +49/241/8022321 | http://dbis.rwth-aachen.de/cms

Splicing

- Many genes are known to have different alternative splice variants,
- i.e. the same pre-mRNA producing different mRNAs, known as
- alternative splicing

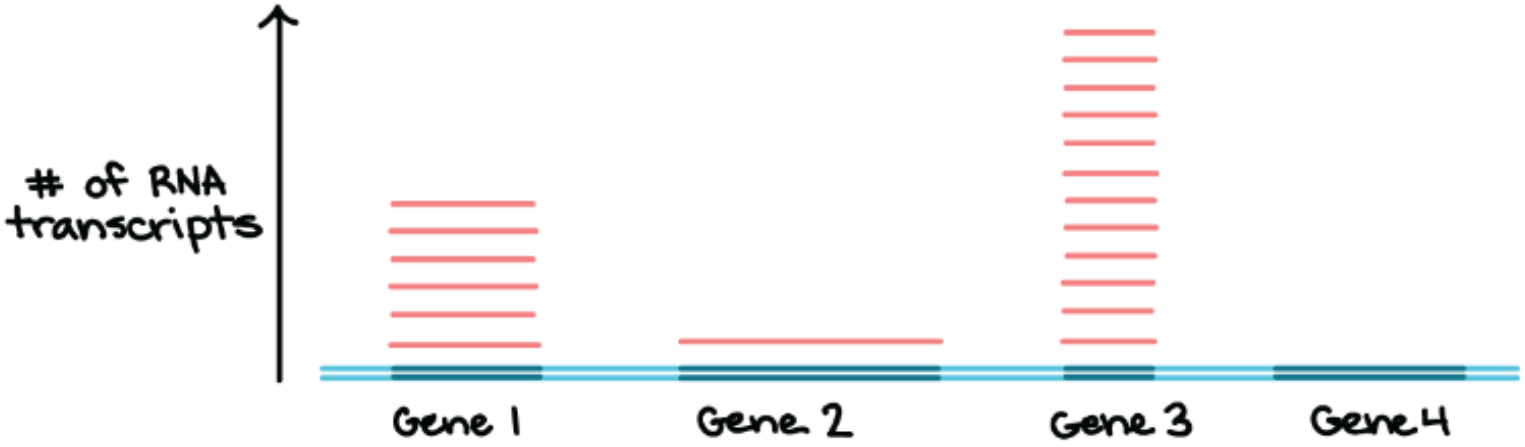


- alternative splicing promotes diversity of proteins produced from a single gene
- enables formation of new proteins from combinations of different genes separated by long noncoding regions

Transcription

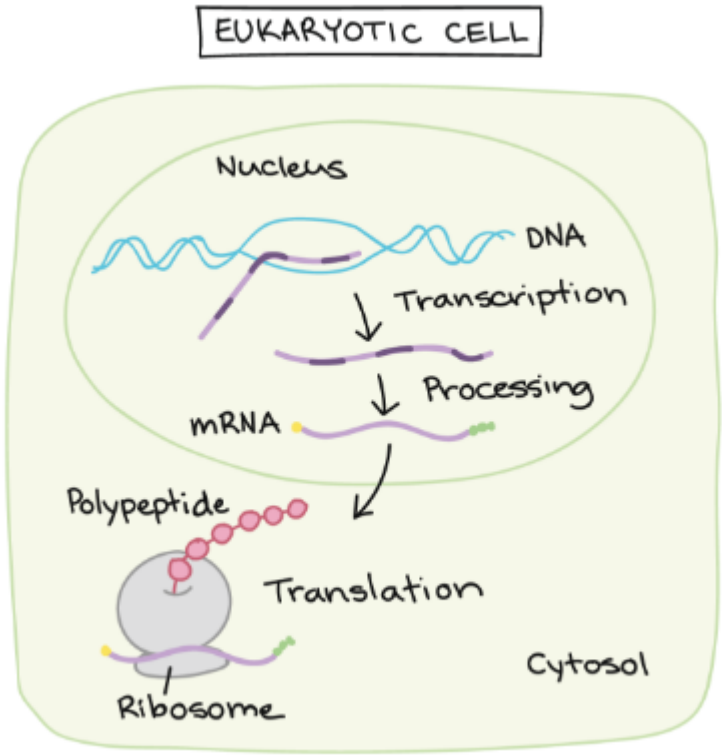
- Messenger RNA, or mRNA, is the RNA “copies” of genes ultimately used to synthesize proteins, although some RNA are the final product themselves

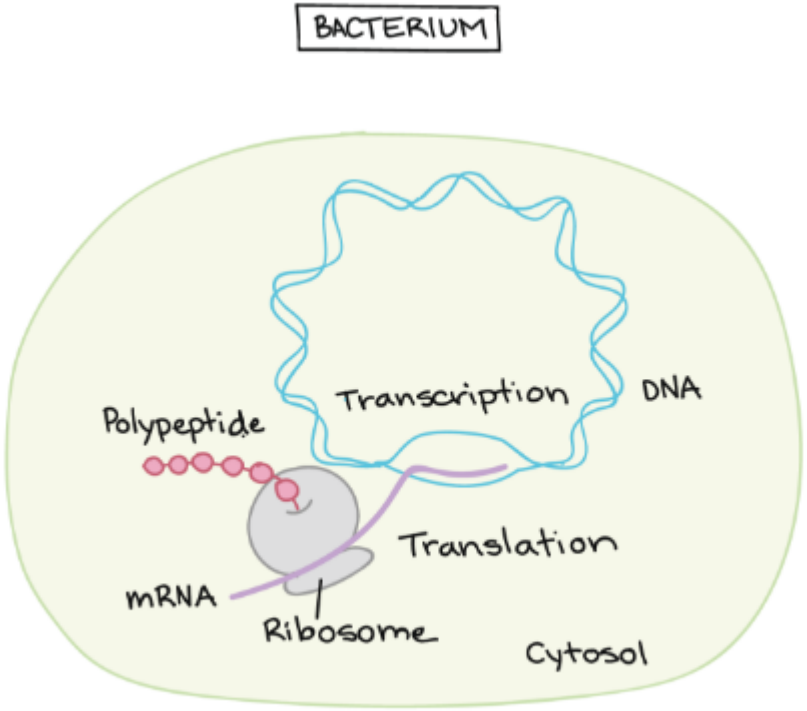
- Not all genes are transcribed all the time.
- Instead, transcription is controlled individually for each gene (or, in bacteria, for small groups of genes that are transcribed together).
- Cells carefully regulate transcription, transcribing just the genes whose products are needed at a particular moment.



Translation

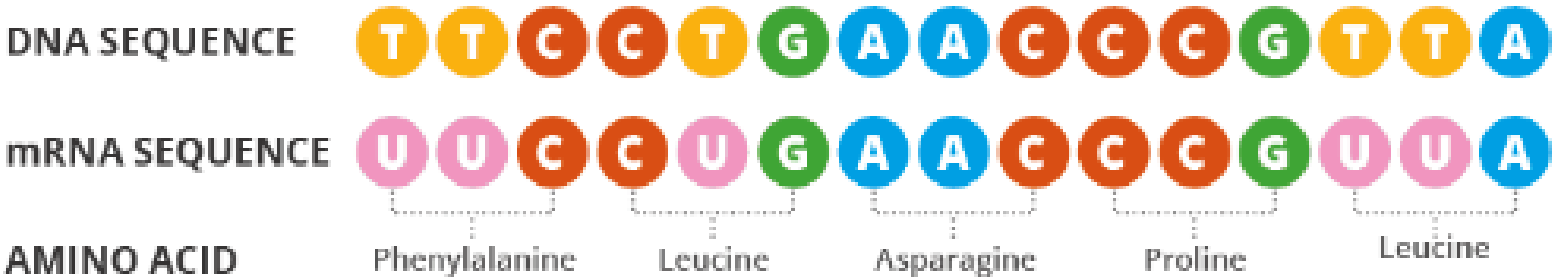
- Since proteins are constructed in the cytoplasm of the cell, mRNA must cross the nuclear membrane to reach the cytoplasm in eukaryotic cells.
- Once in the cytoplasm, ribosomes and another RNA molecule called transfer RNA work together to translate mRNA into a protein.
- This process is called translation.
- Proteins can be manufactured in large quantities because a single DNA sequence can be transcribed by many RNA polymerase molecules at once.





Translation

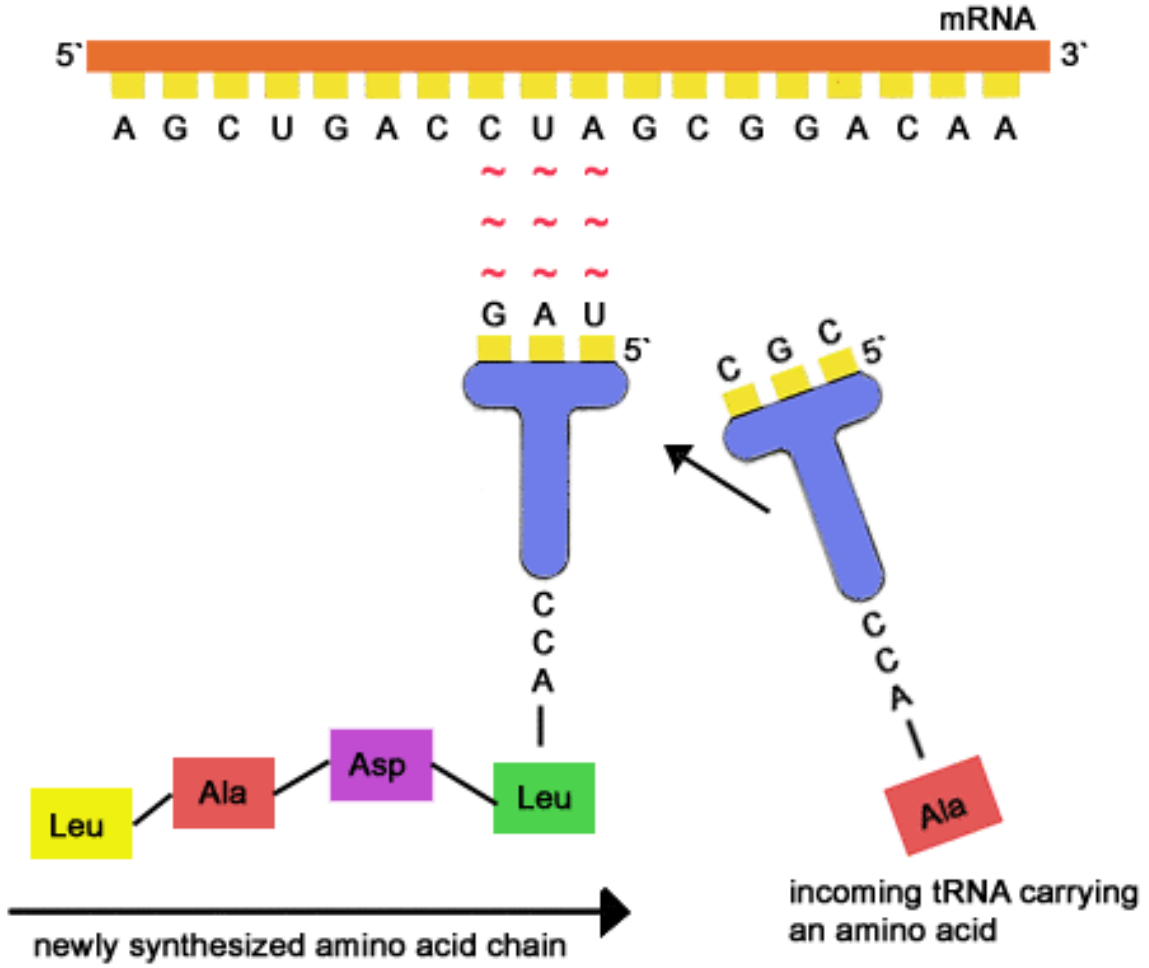
- Translation is the process of converting the mRNAs `code` into proteins by joining together amino acids in the order encoded in the mRNA.
- An amino acid is determined by 3 adjacent nucleotides (triplets) in the DNA.
- This is known as the triplet or genetic code.
- Each triplet is called a codon and codes for one amino acid.



RWTH Informatik 5 | Ahornstr. 55 D-52056 Aachen Germany |
Tel +49/241/8021501 | Fax +49/241/8022321 | http://dbis.rwth-aachen.de/cms

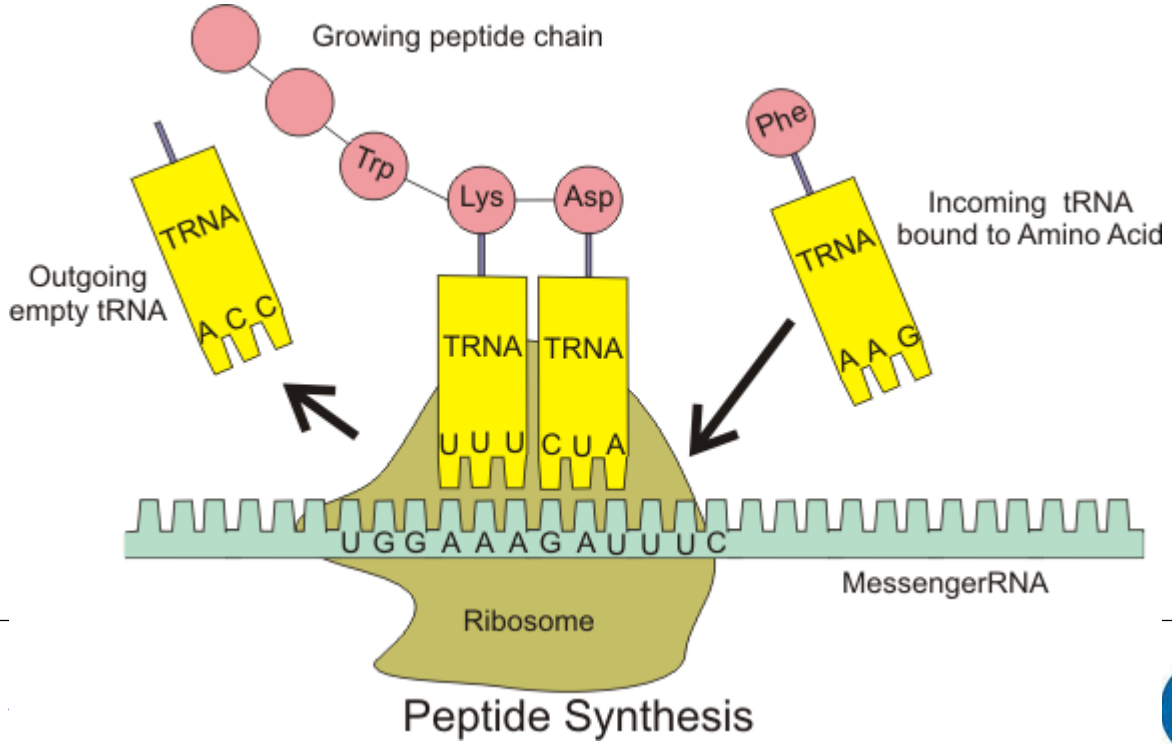
Translation

- tRNA builds proteins from the information in mRNA.
- The sequence of bases in the mRNA defines the order and sequence of amino acids.
- Each tRNA molecule carries an amino acid to the ribosomes, by matching its anticodon to a specific codon from the mRNA



Translation

- The amino acids are joined by a bond which is known as a peptide bond.
- For initiation, the ribosome binds to the mRNA at the start cordon (AUG) that is recognized only by the initiator tRNA.
- The ribosome moves from cordon to cordon along the mRNA. Amino acids are added one by one, translated into Polypeptide sequences dictated by DNA
- At the end, a release factor binds to the stop cordon, terminating translation and releasing the complete polypeptide from the ribosome



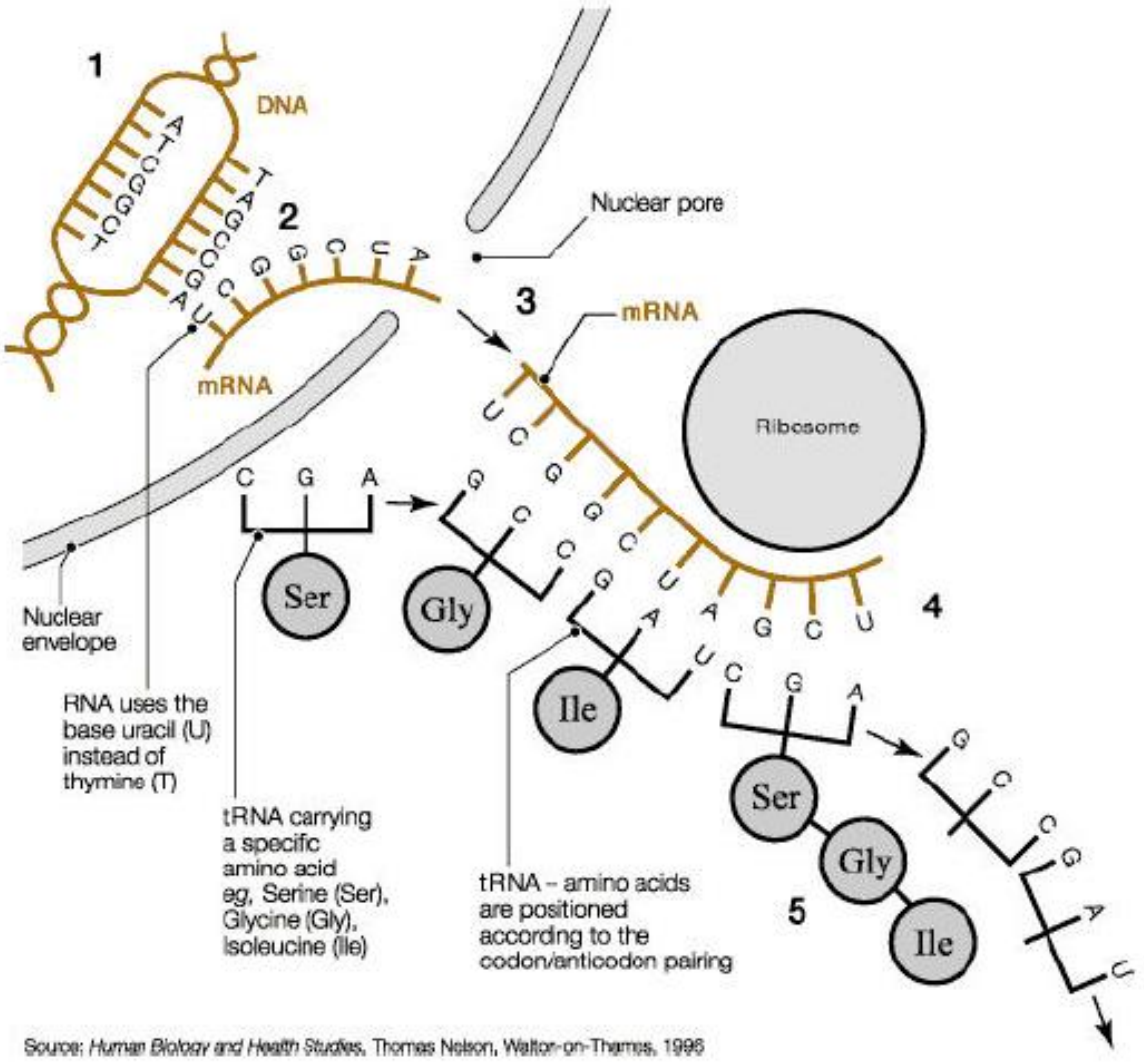
Translation

- As there are 64 codons and only 20 amino acids the code is redundant, for example histidine is encoded by CAT and CAC.

Second Position

First Position (5' end)	U	C	A	G	Third Position (3' end)
U	UUU Phe UUC UUA Leu UUG	UCU UCC UCA Ser UCG	UAU Tyr UAC UAA Stop UAG Stop	UGU Cys UGC UGA Stop UGG Trp	U C A G
C	CUU CUC CUA Leu CUG	CCU CCC CCA Pro CCG	CAU His CAC CAA Gln CAG	CGU CGC CGA Arg CGG	U C A G
A	AUU AUC Ile AUA AUG Met	ACU ACC ACA Thr ACG	AAU Asn AAC AAA Lys AAG	AGU Ser AGC AGA Arg AGG	U C A G
G	GUU GUC GUA Val GUG	GCU GCC GCA Ala GCG	GAU Asp GAC GAA Glu GAG	GGU GGC GGA Gly GGG	U C A G

The synthesis of proteins reects the central dogma



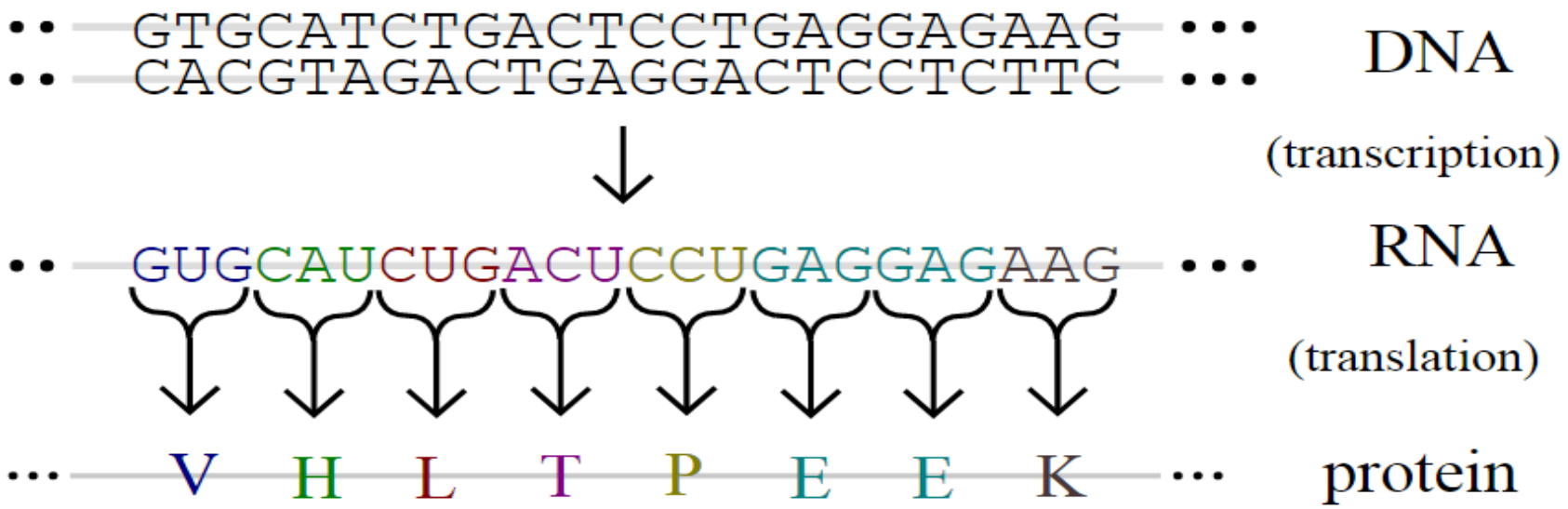
Source: *Human Biology and Health Studies*, Thomas Nelson, Walton-on-Thames, 1996

RWTH Informatik 5 | Ahornstr. 55 D-52056 Aachen Germany |
Tel +49/241/8021501 | Fax +49/241/8022321 | http://dbis.rwth-aachen.de/cms

Gene expression

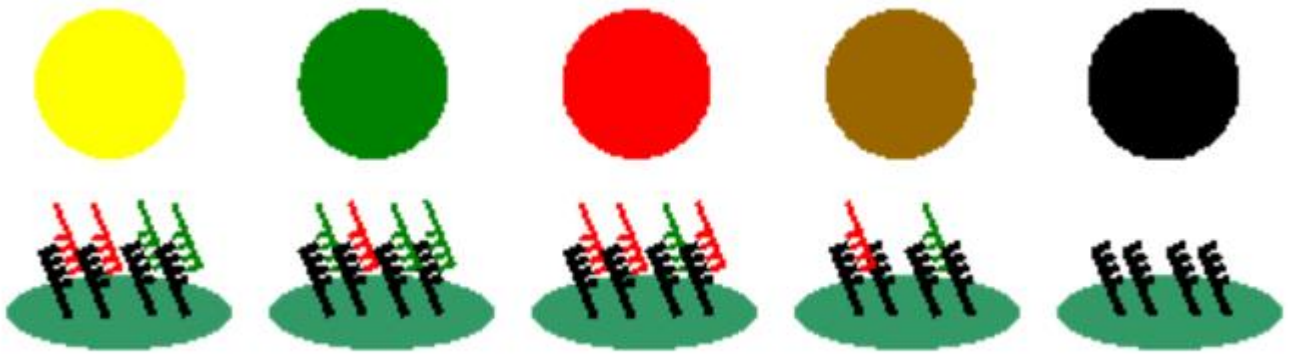
- Gene expression is the process by which information from a gene is used in the synthesis of a functional gene product.
  - often proteins,
  - non-protein coding genes such as
  - transfer RNA (tRNA)
  - small nuclear RNA (snRNA) genes

- In genetics, gene expression is the most fundamental level at which the genotype gives rise to the phenotype, i.e. observable trait.
- The genetic code stored in DNA is "interpreted" by gene expression, and the properties of the expression give rise to the organism's phenotype.



Microarray

- Microarray technology has two major applications: gene expression analysis and genetic variation analysis.
- Microarray is a hybridization of a nucleic acid sample (target) to a very large set of oligonucleotide probes, which are attached to a solid support.
- Aim is to determine sequence or to detect variations in a gene sequence or expression or for gene mapping

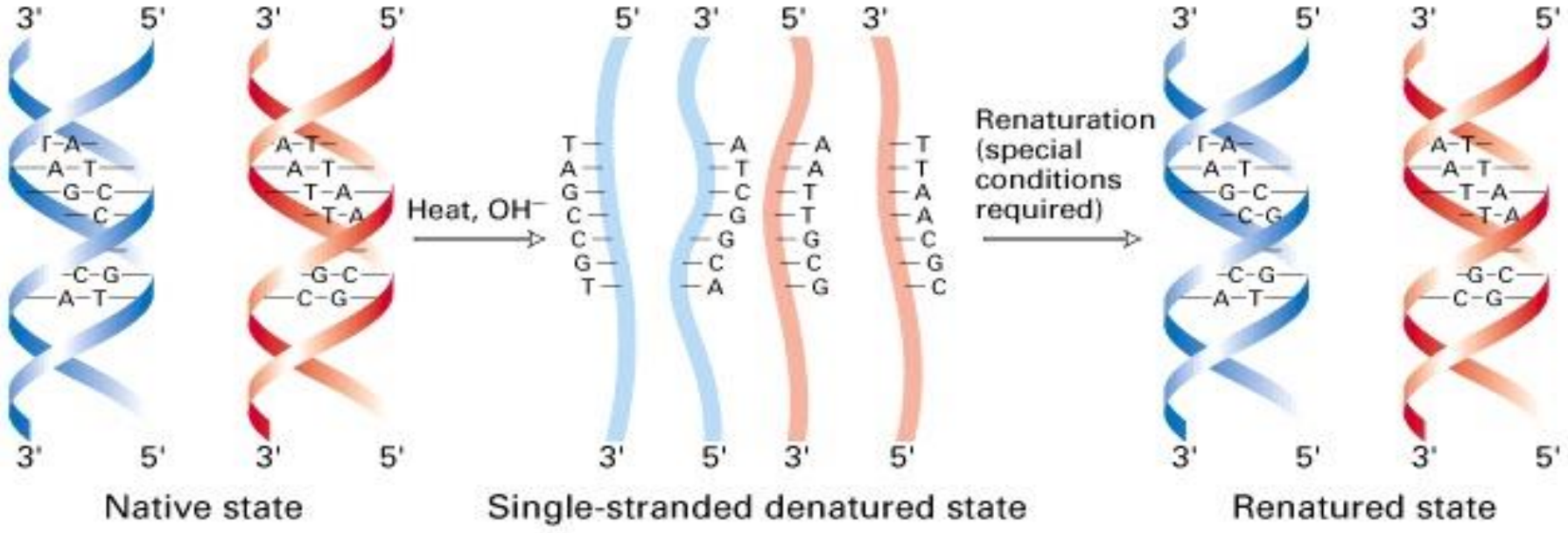


Nucleic acids hybridization

- Hybridization is the process by which two complementary, single-stranded nucleic acids combine into a single molecule.

- In molecular biology, **hybridization** (or **hybridisation**) is a phenomenon in which single-stranded deoxyribonucleic acid (DNA) or ribonucleic acid (RNA) molecules anneal to complementary DNA or RNA.

- Nucleotides bind to their complement (A with T and C with G) under normal conditions, so two perfectly complementary strands will bind to each other readily. This is called annealing.



Nucleic acids hybridization

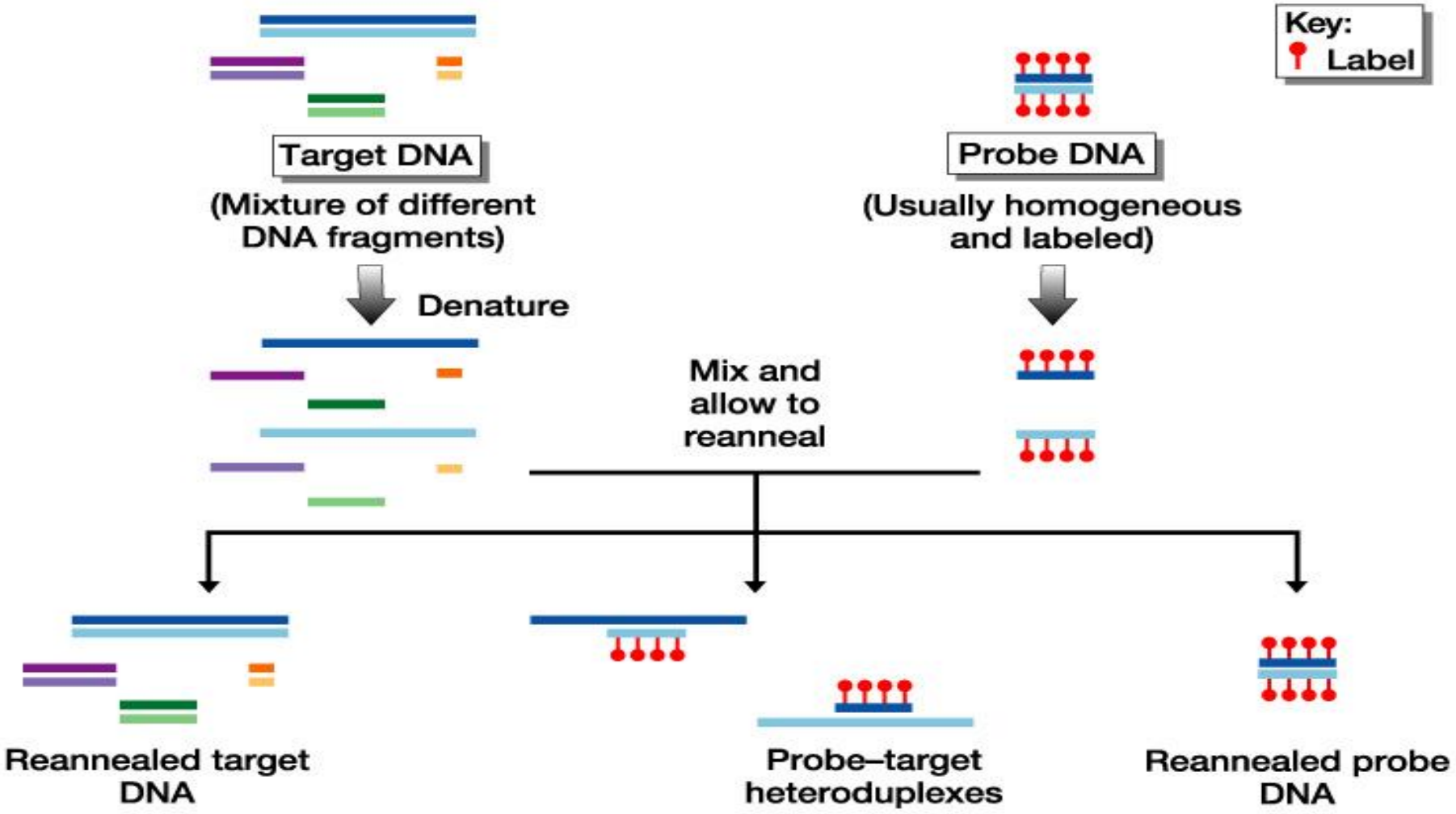
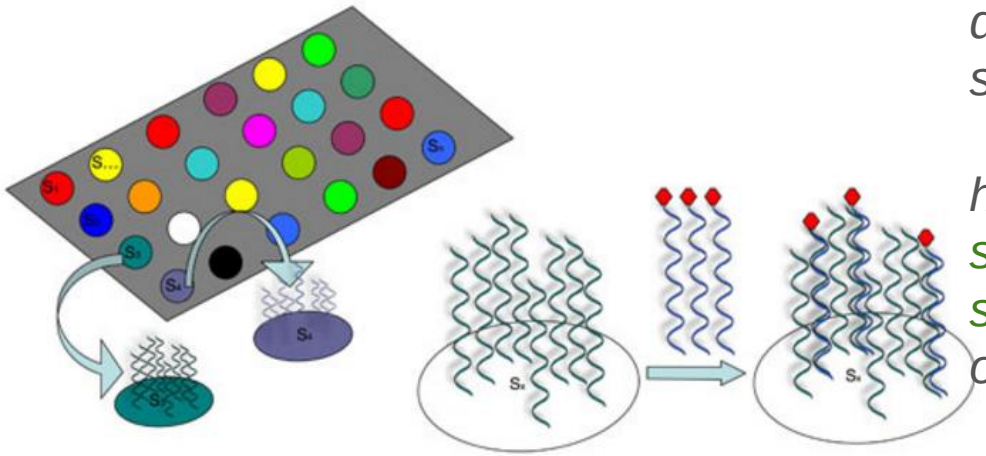


Figure 6-8 Human Molecular Genetics, 3/e. (© Garland Science 2004)

Microarray

- A microarray consists of a solid surface on which strands of polynucleotide called probes have been attached or synthesized in fixed positions.

- Two types of expression microarrays
- Spotted or cDNA microarrays take their name because probes are synthesized apart and printed mechanically on the slide. The term cDNA is used because the probe is a complimentary copy of the original sequence and each probe represents one gene.

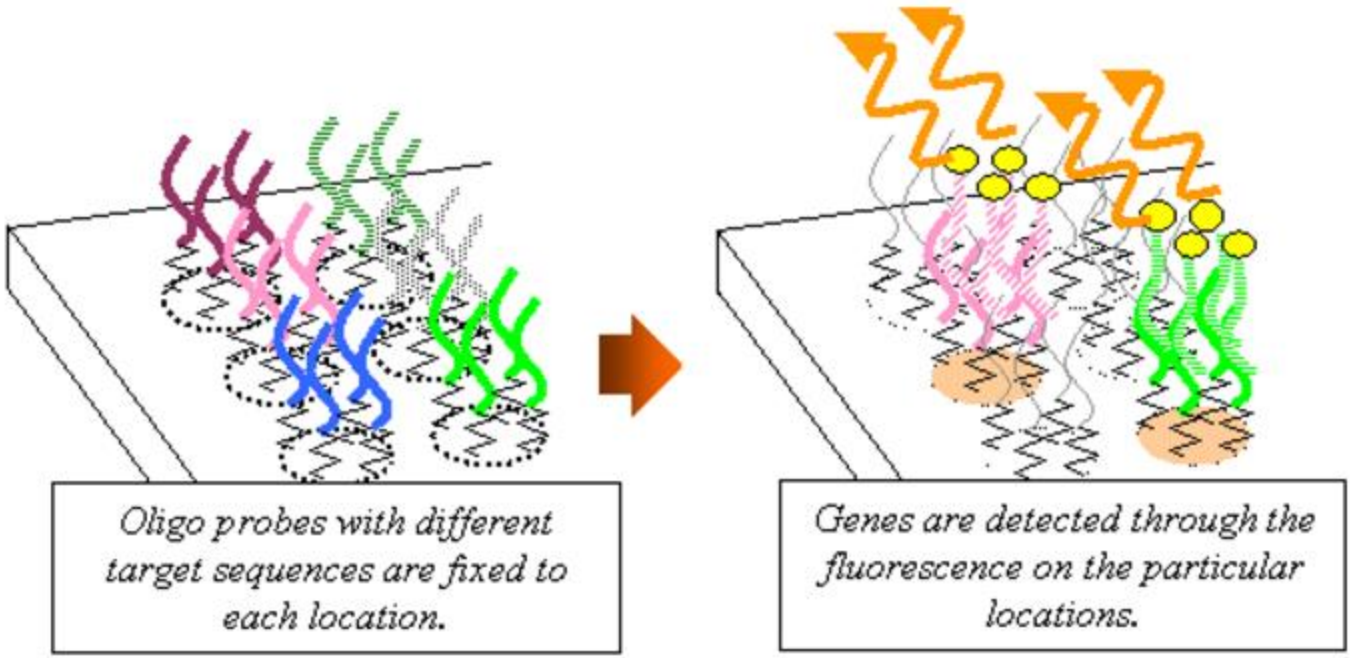


A DNA microarray is a collection of synthetic DNA probes attached to designated location, or spot, on a solid surface.

The resulting "grid" of probes can hybridize to complementary "target" sequences derived from experimental samples to determine the expression level of specific mRNAs in a sample.

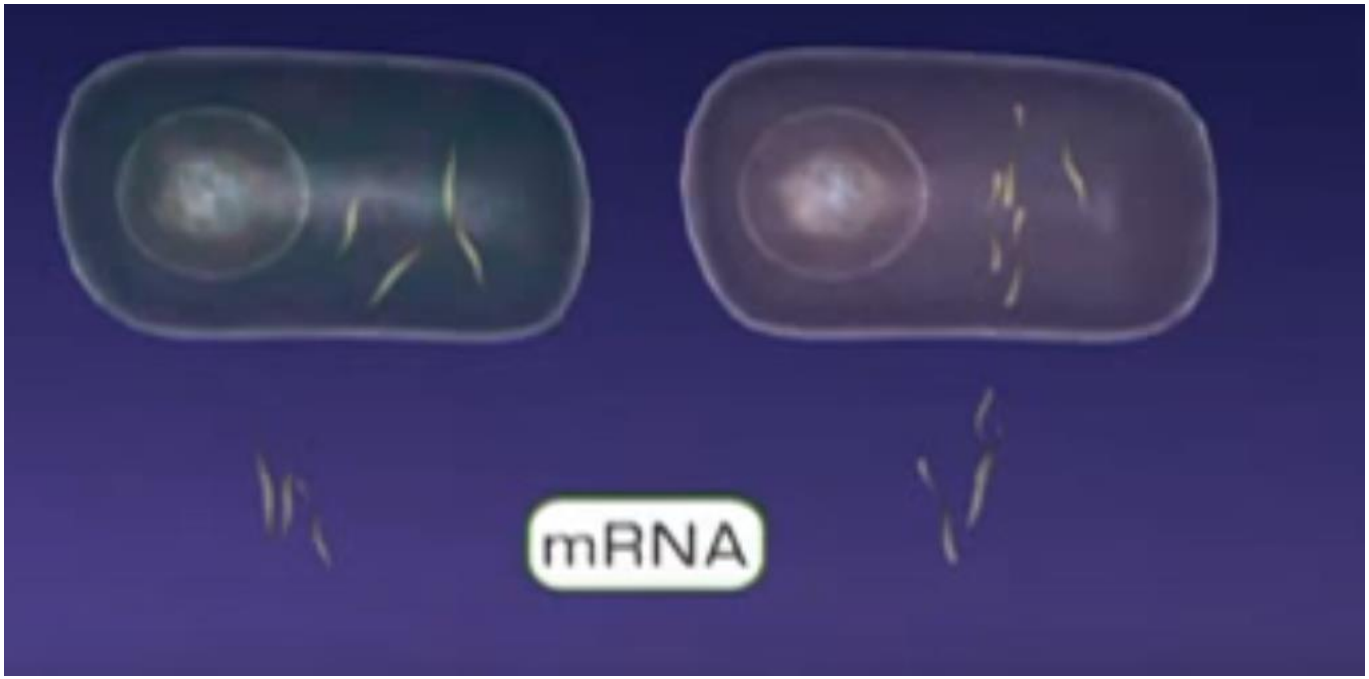
Microarray

- Oligonucleotide chips, where main representatives are Genechip or Aymetrix, the name of the commercial brand that manufactures them
- the probes are directly synthesized on the surface.
- The difference is that these chips contain the *oligos [*a polynucleotide whose molecules contain a relatively small number of nucleotides*]*, and not the targets.
- The term oligonucleotide refers to the fact that the synthesis process allows to create only small fragments so that a gene is not represented by one probe but by as a set of them (a probe set).



Micro Array

- To start a microarray experiment RNA is extracted from the subject cells.

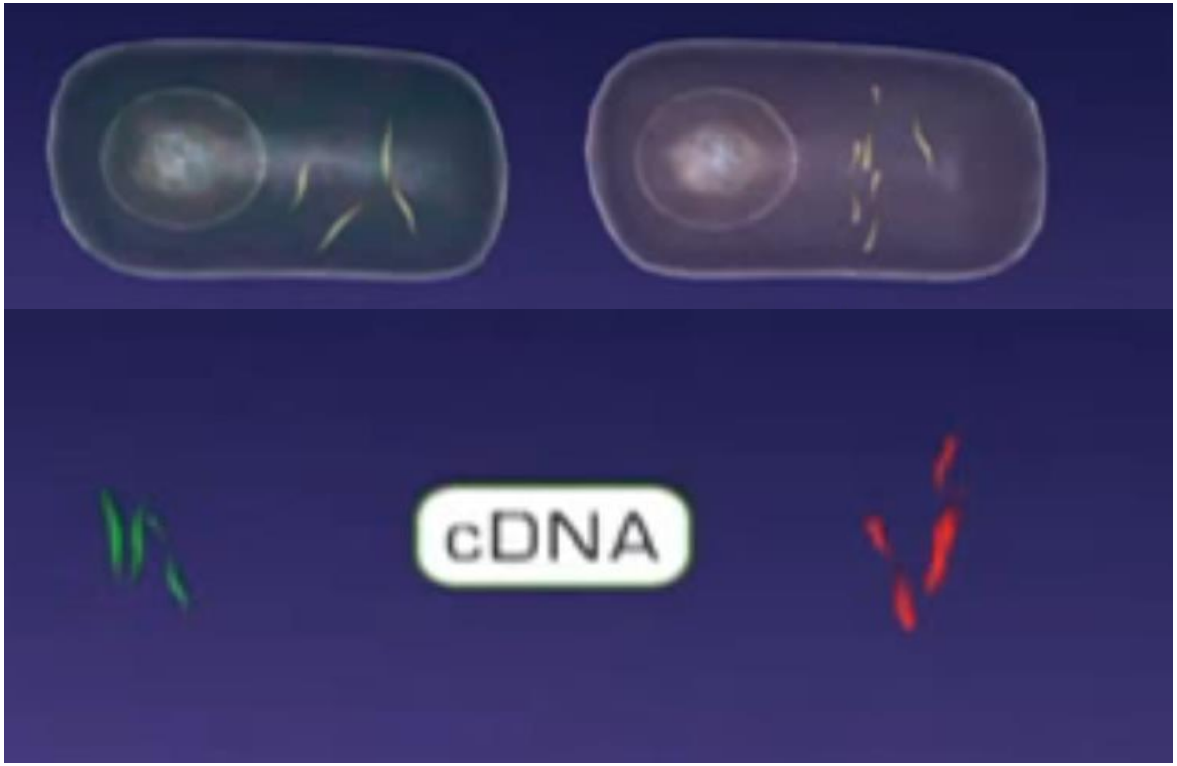


Than transfer into more stable complementary cDNA

https://www.youtube.com/watch?v=VNsThMNjKhM

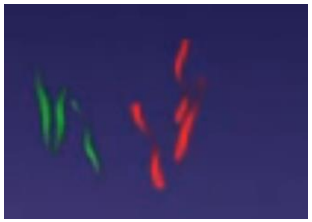
Micro Array

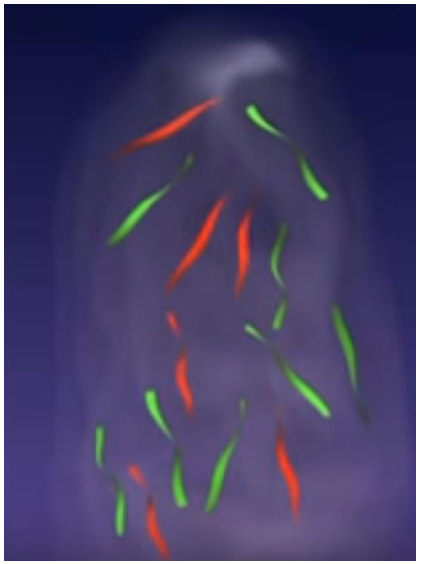
- After this, some of its molecules are substituted by others containing a fluorescent dye.
- The resulting labelled transcripts are called targets.

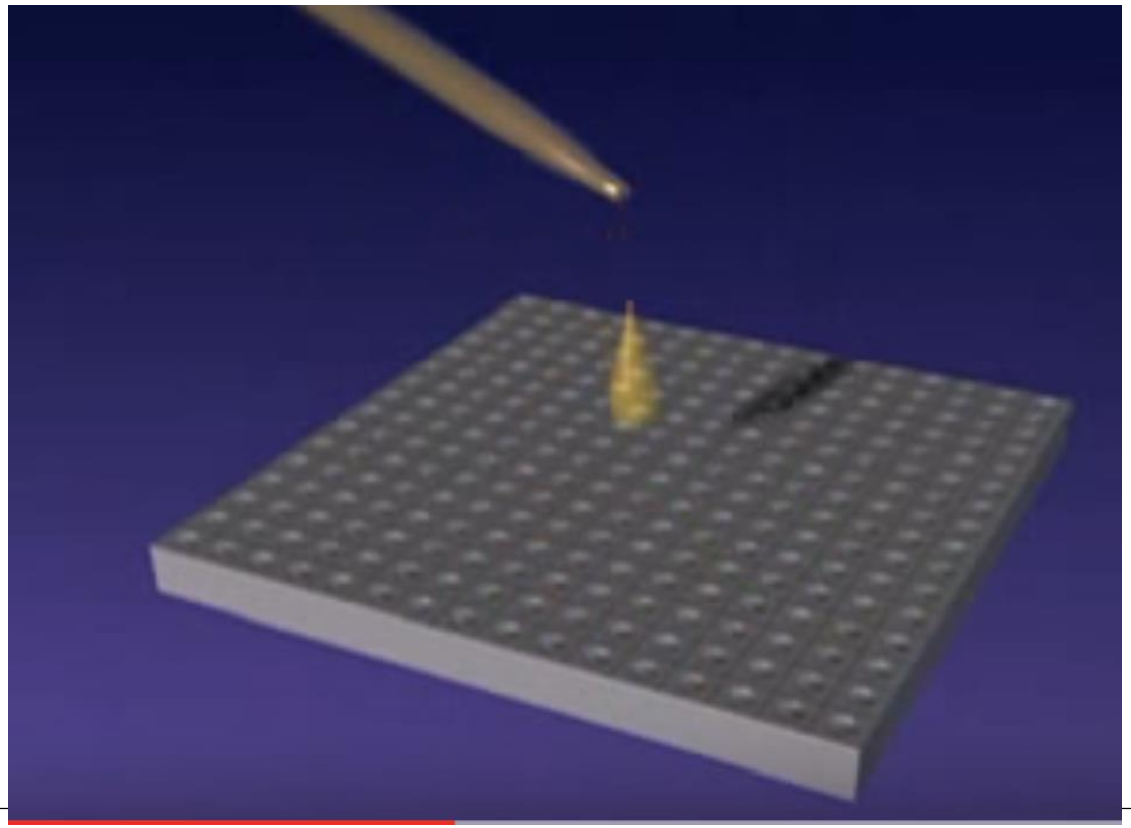


Micro Array

- Once the samples are prepared and combined they are deposited over the array and left inside a hybridization chamber for some hours.
- The labelled targets bind by hybridization to the probes on the array with which they share sufficient sequence complementarity.

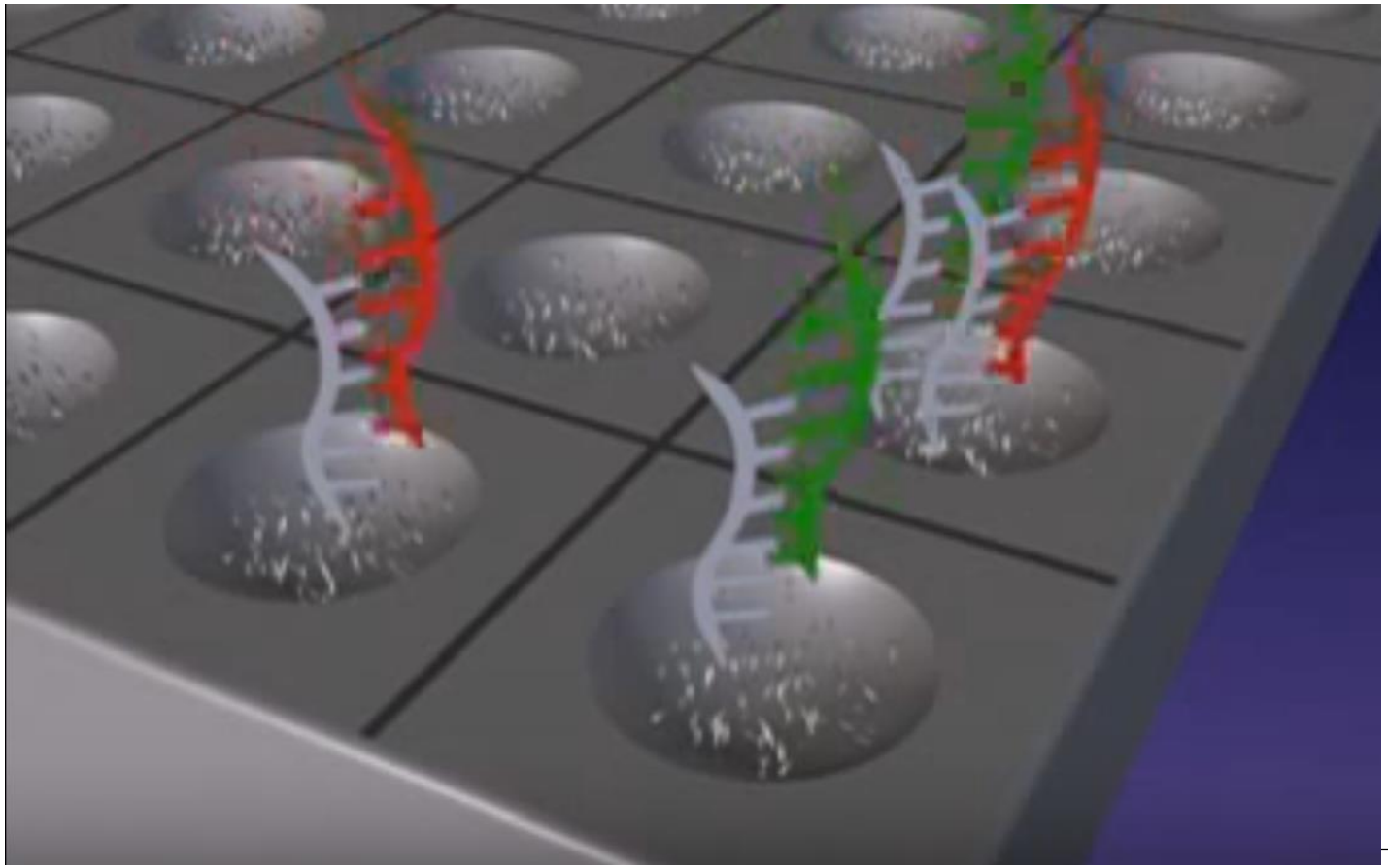




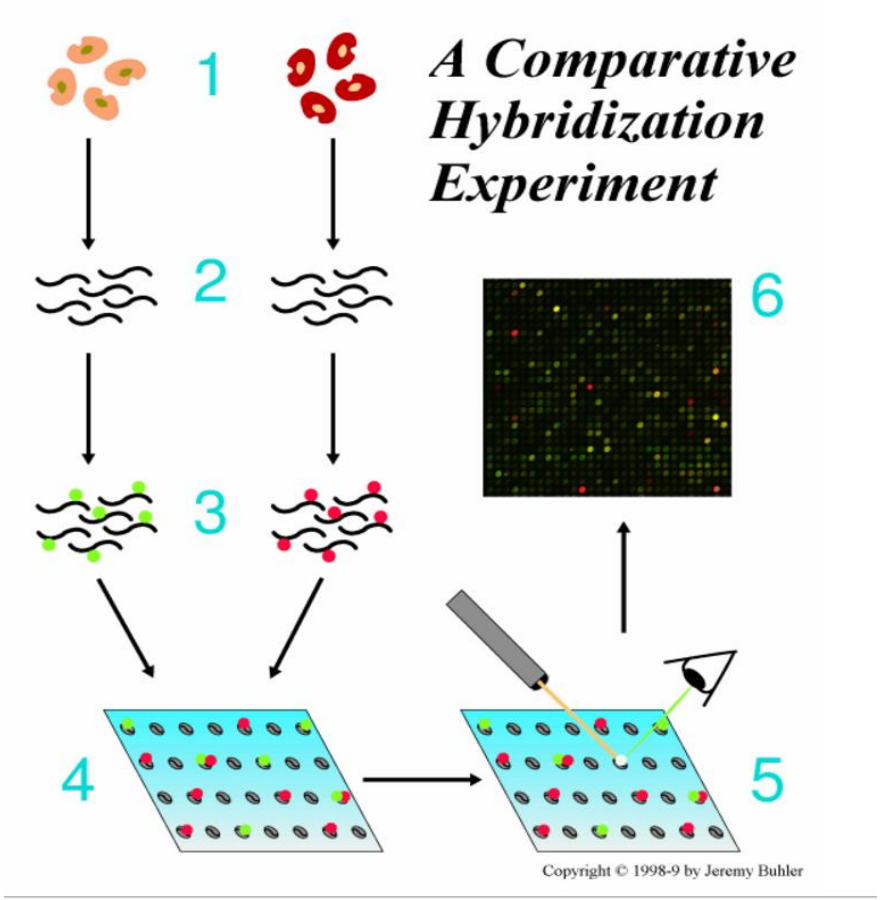


Micro Array

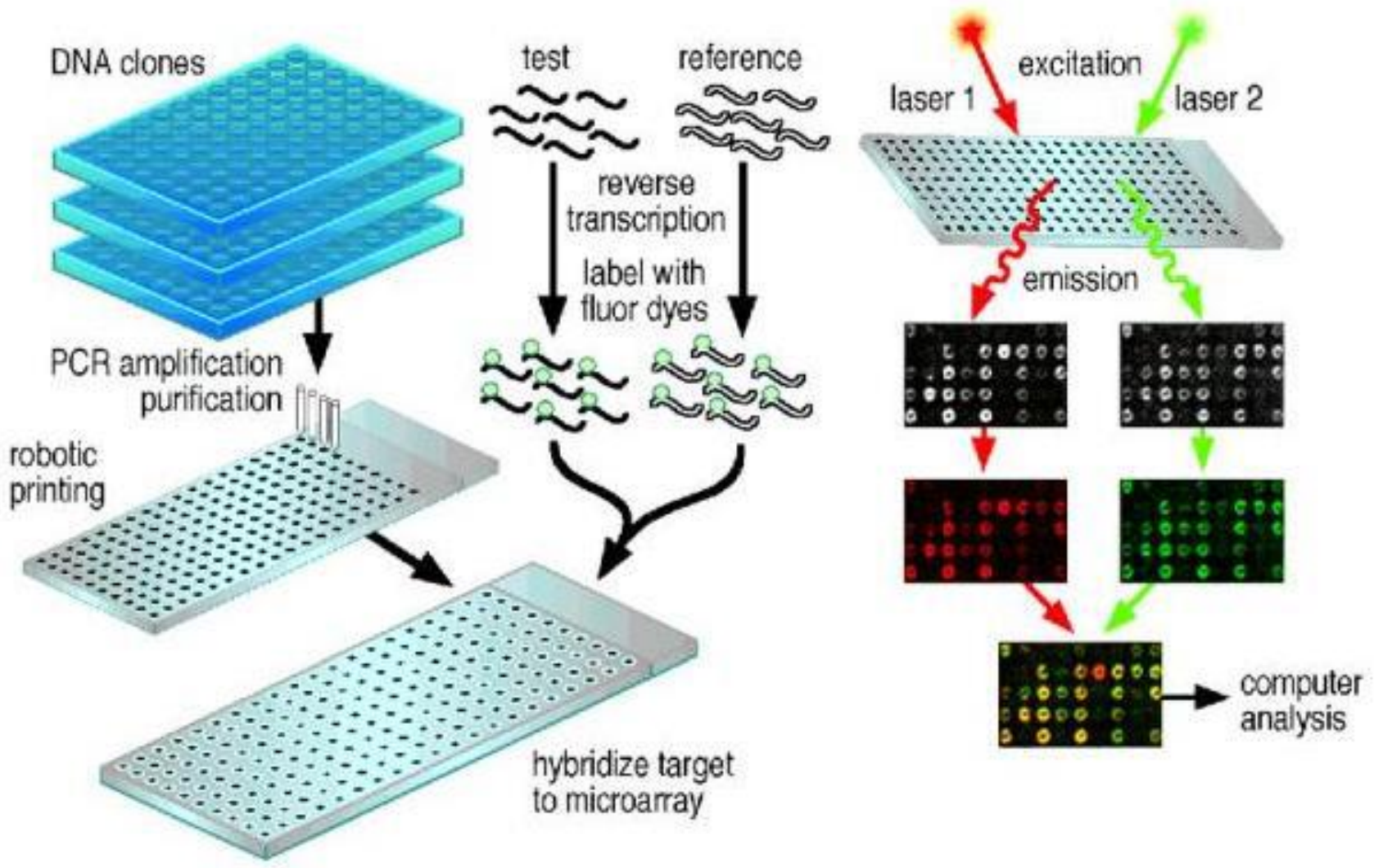
- After this time the array is washed which eliminates those targets which have not hybridized.
- cDNA binds to its complementary sequence on the chip



- In spotted microarrays cDNAs from two tissues of interest, labelled with fluorescent dyes of different color (usually red and green), are hybridized to a single chip.
- For obvious reasons spotted chips are also called two-color arrays.



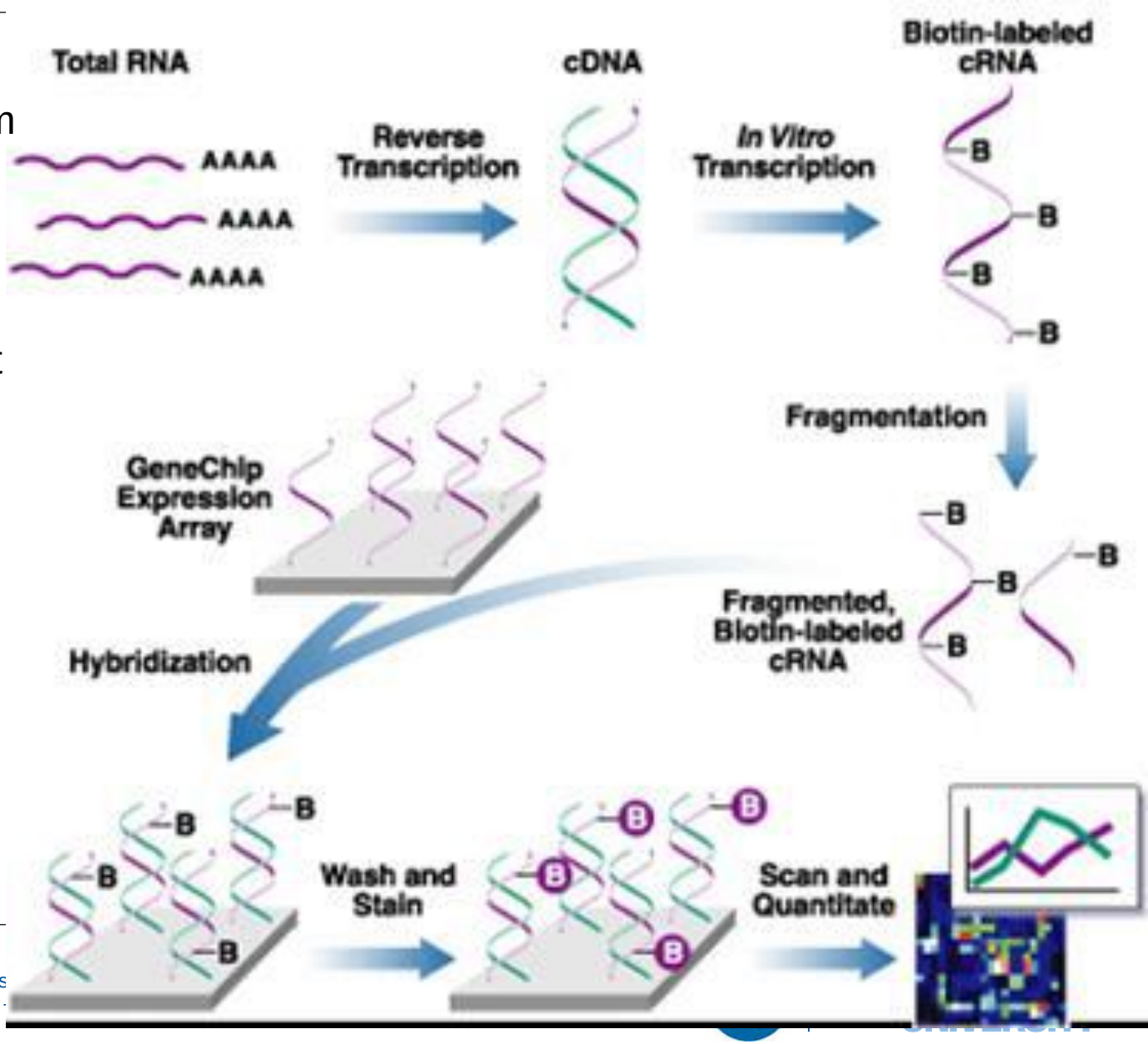
Micro Array: Two color cDNA chips



RWTH Informatik 5 | Ahornstr. 55 D-52056 Aachen Germany |
Tel +49/241/8021501 | Fax +49/241/8022321 | http://dbis.rwth-aachen.de/cms

Micro Array : One color aymetrix chips

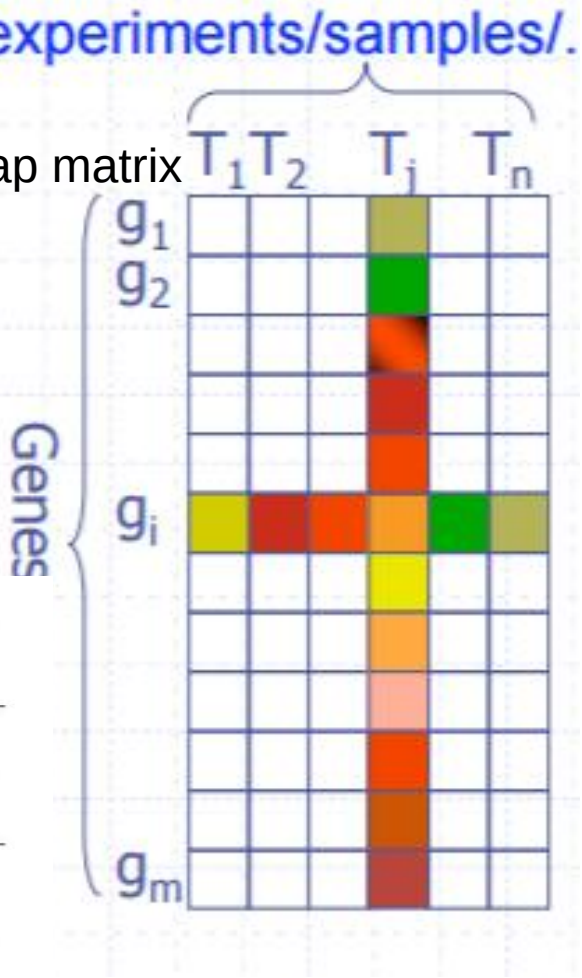
- The Affymetrix system hybridizes only one sample per chip
- This requires more slides per experiment and does not enjoy the advantage of using competitive hybridization
- However it simplifies experimental design and is based on a much more sensitive technology.

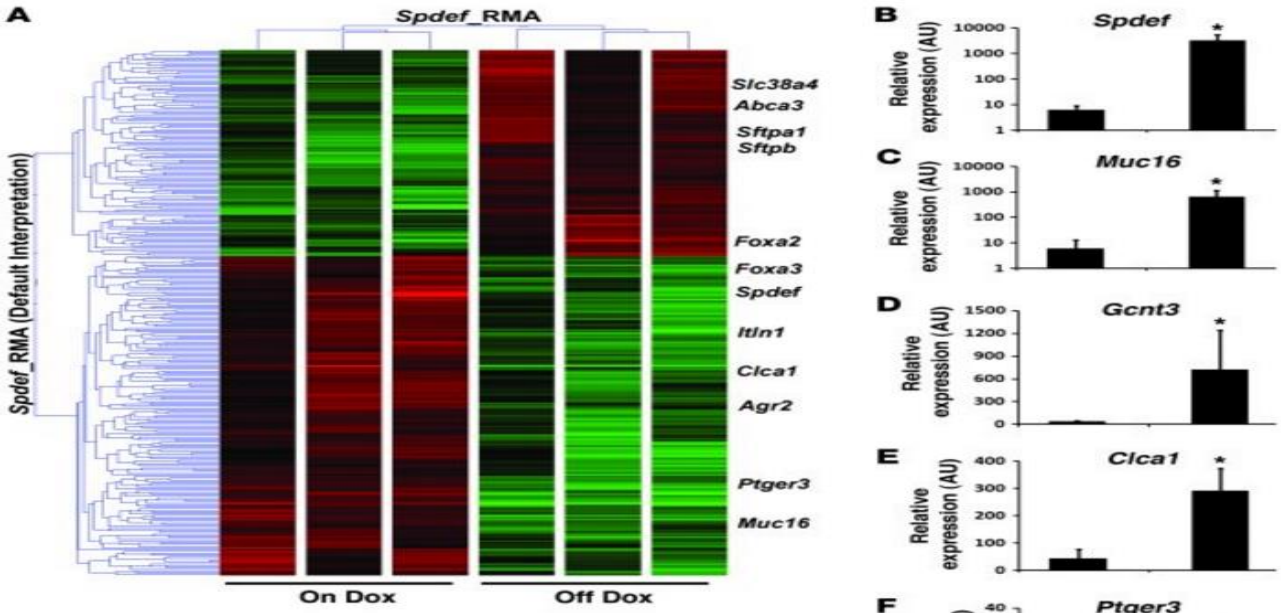


RWTH Informatik 5 | Ahorns
Tel +49/241/8021501 | Fax

Microarray Heat Map

- Microarray measurements may be organized in a heat-map matrix
- Row represent genes
- Columns represent tests
- Xij =expression level of gi under test Tj
- Expression level is visualized via colors
  - Green= under expressed (down regulated)
  - Red = over expressed (up regulated)





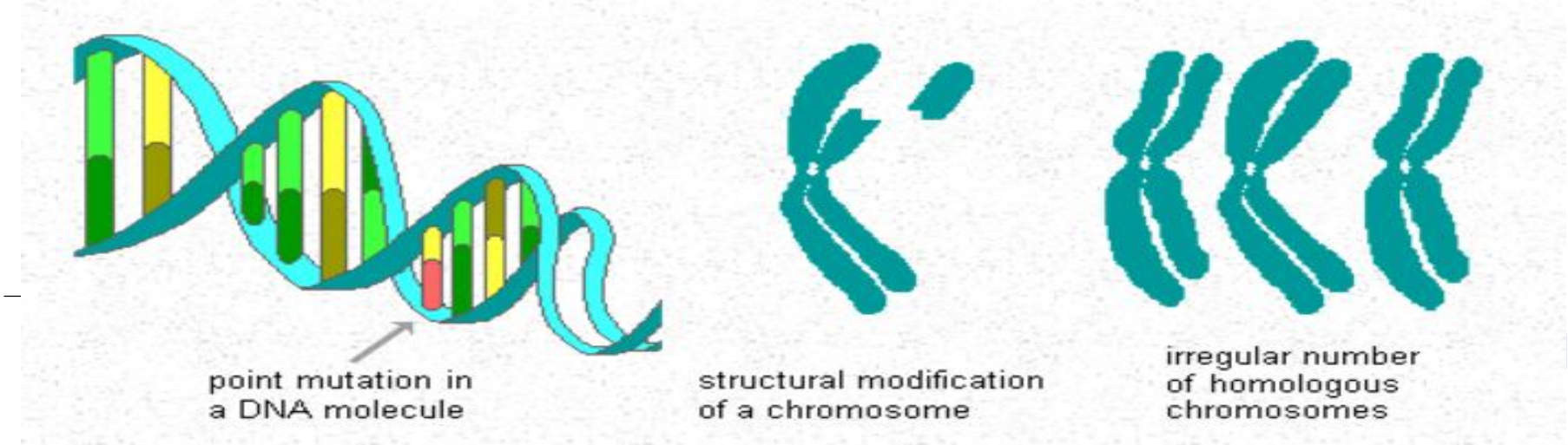
mRNA microarray analysis of bronchiolar epithelial cells: heat map and partial list of SPDEF-regulated genes.

Chen, Gang, et al. "SPDEF is required for mouse pulmonary goblet cell differentiation and regulates a network of genes associated with mucus production." *The Journal of clinical investigation* 119.10 (2009): 2914-2924.

RWTH AACHEN UNIVERSITY

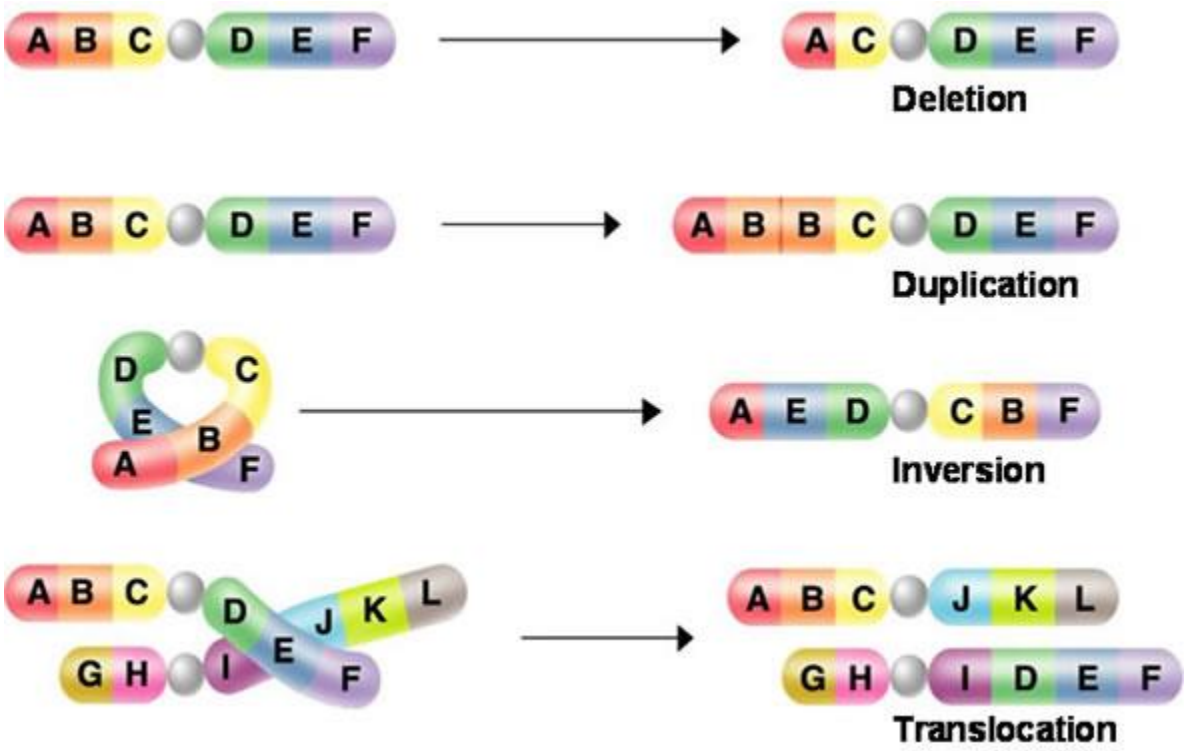
Mutations

- During **replication**, an organism's DNA can change or mutate. Changes to genes are called mutations.
- Gene mutations and chromosome mutations are the two basic types of mutations
- Gene mutation is a change in the nucleotide sequence, in a particular gene, whereas chromosomal mutation is a change in several genes, in the chromosome.
- Gene mutation is a small-scale alteration, but chromosome mutation can be considered as a serious alteration.
- Gene mutations can sometimes be corrected, but chromosomal mutations are hardly corrected.
- Gene mutation is only a slight structural alteration, whereas chromosomal mutations are either numerical or structural changes in the entire DNA strand.



Types of Chromosomal Mutations

- Deletions (whole or part deleted)
- Duplications (extra copies of parts)
- Inversion (reverses parts of chromosomes)
- Translocations (parts break off and relocate)



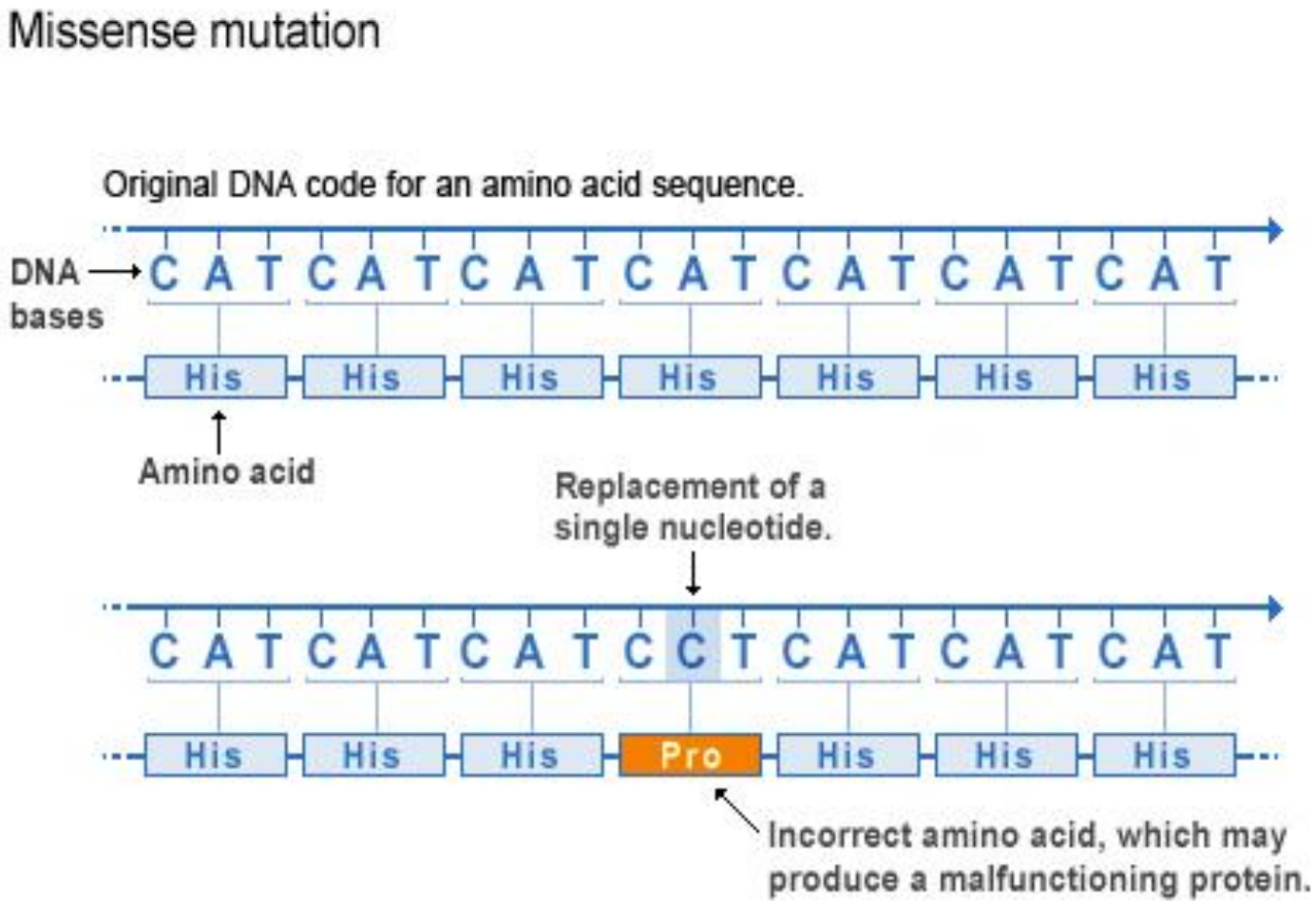
Some chromosomal mutations are common and harmless, such as different eye colors in humans.
Some result in a condition that affects health, such as Down's syndrome.
Sometimes chromosomal mutations are beneficial and can produce proteins that fight a particular disease.

Types of Gene Mutations

- The DNA sequence of a gene can be altered in a number of ways.
- Gene mutations have varying effects on health, depending on where they occur and whether they alter the function of essential proteins.
- Gene mutation is a small-scale alteration of the genetic material of an organism, which primarily is a change in the nucleotide sequence of a particular gene.
- These changes are of two types based on the way those take place: Point mutations and frame shift mutations are the two main types
- When a nucleotide of a particular gene is changed, the transcribe mRNA and the subsequent codons and synthesized amino acids are altered.
- Gene mutation may lead to alter the number or the structure of the entire chromosome, which could lead to chromosomal mutations.

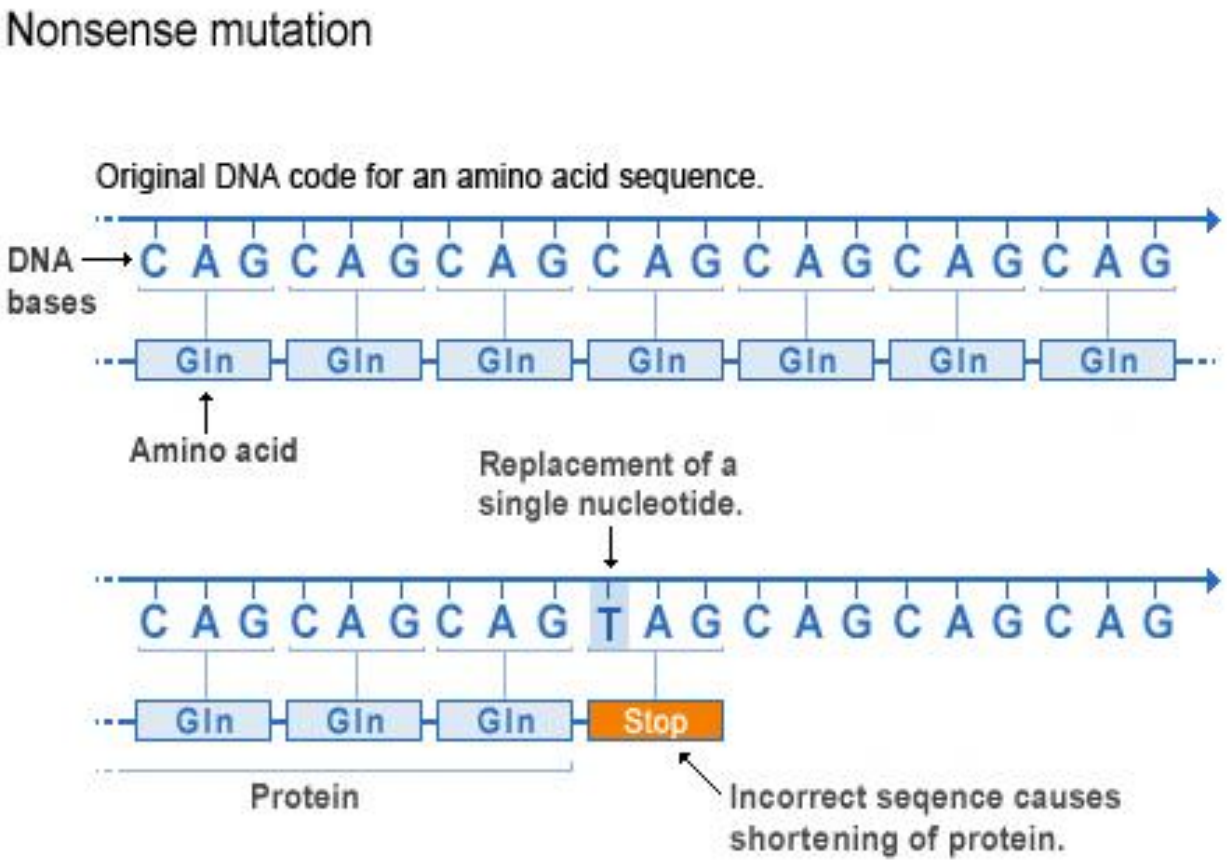
Types of Gene Mutations: Missense mutation

- This type of mutation is a change in one DNA base pair that results in the substitution of one amino acid for another in the protein made by a gene.



Types of Gene Mutations: Nonsense mutation

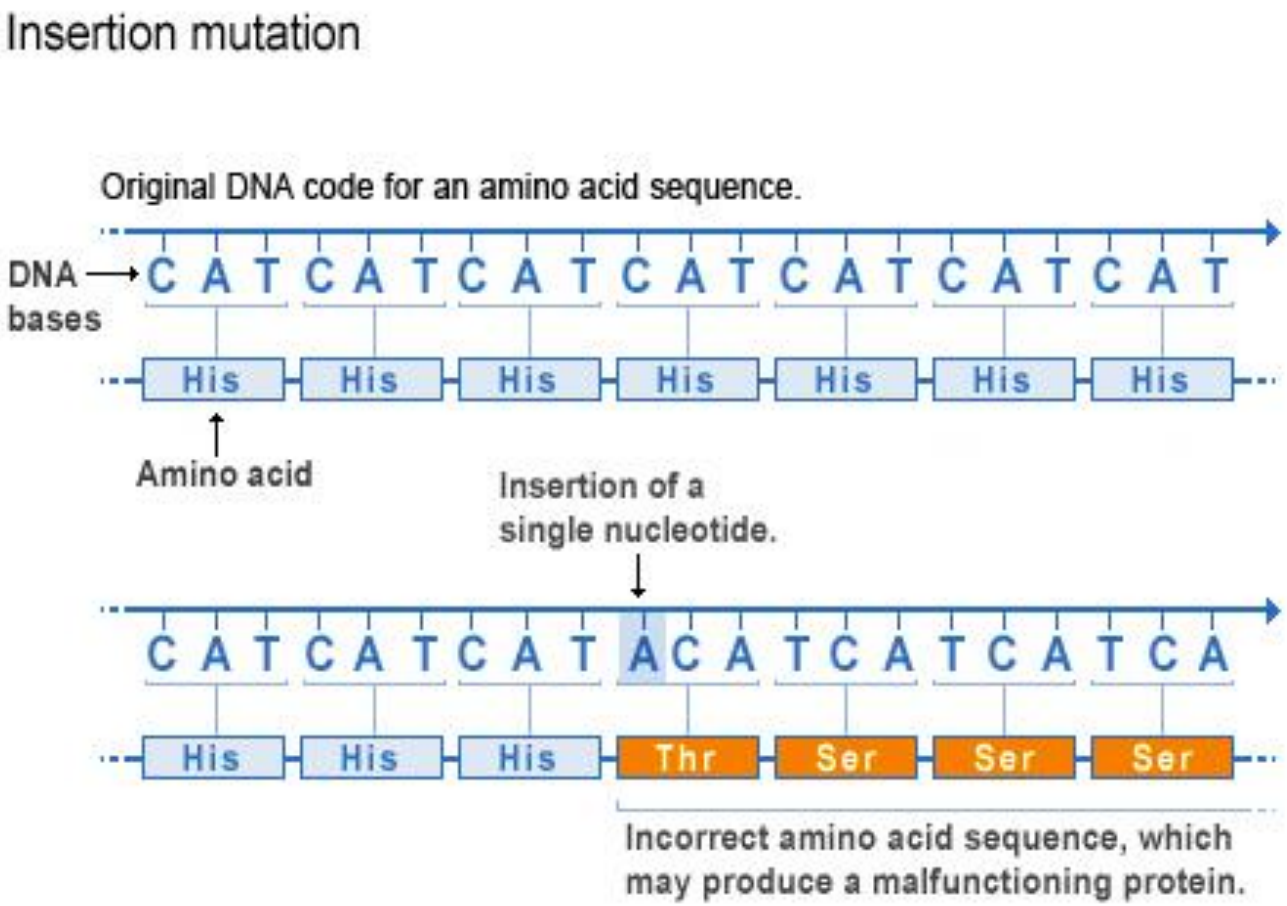
- A nonsense mutation is also a change in one DNA base pair. Instead of substituting one amino acid for another, however, the altered DNA sequence prematurely signals the cell to stop building a protein. This type of mutation results in a shortened protein that may function improperly or not at all.



U.S. National Library of Medicine

Types of Gene Mutations: Insertion

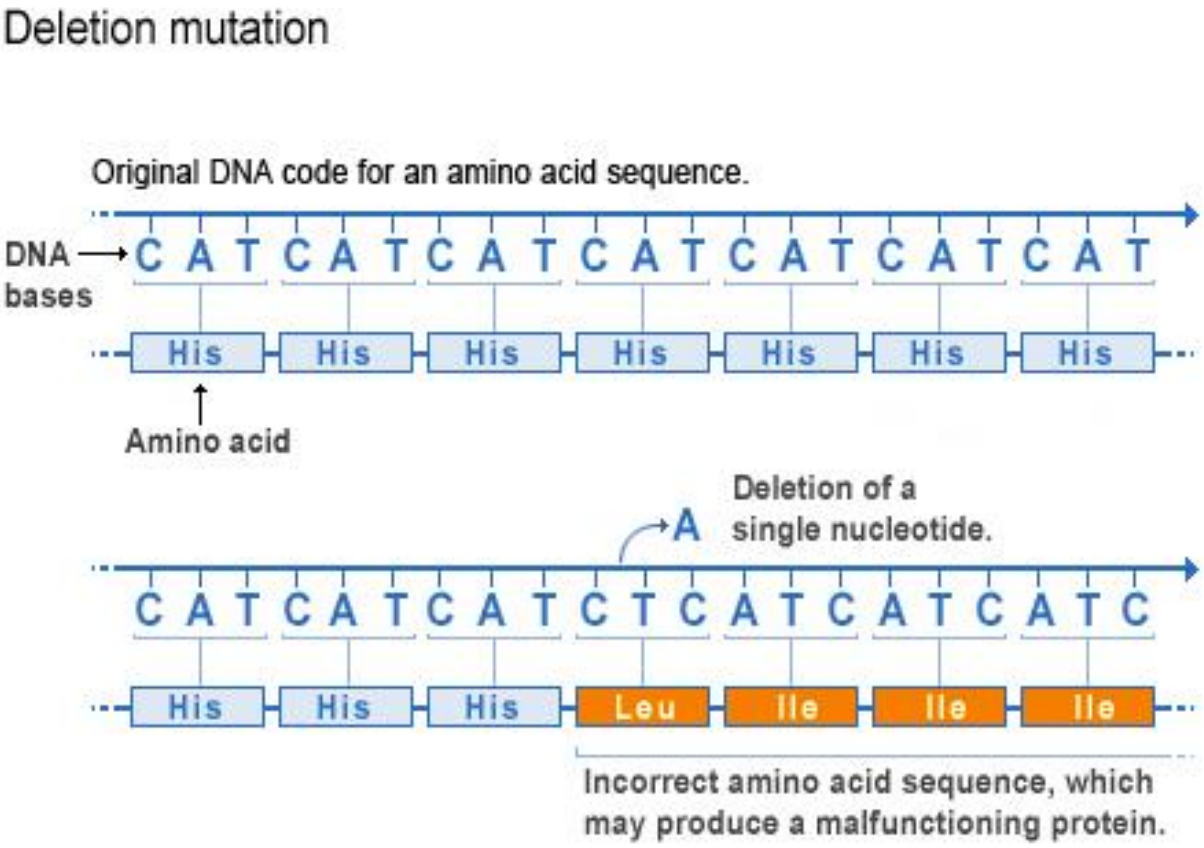
- An insertion changes the number of DNA bases in a gene by adding a piece of DNA. As a result, the protein made by the gene may not function properly.



U.S. National Library of Medicine

Types of Gene Mutations : Deletion

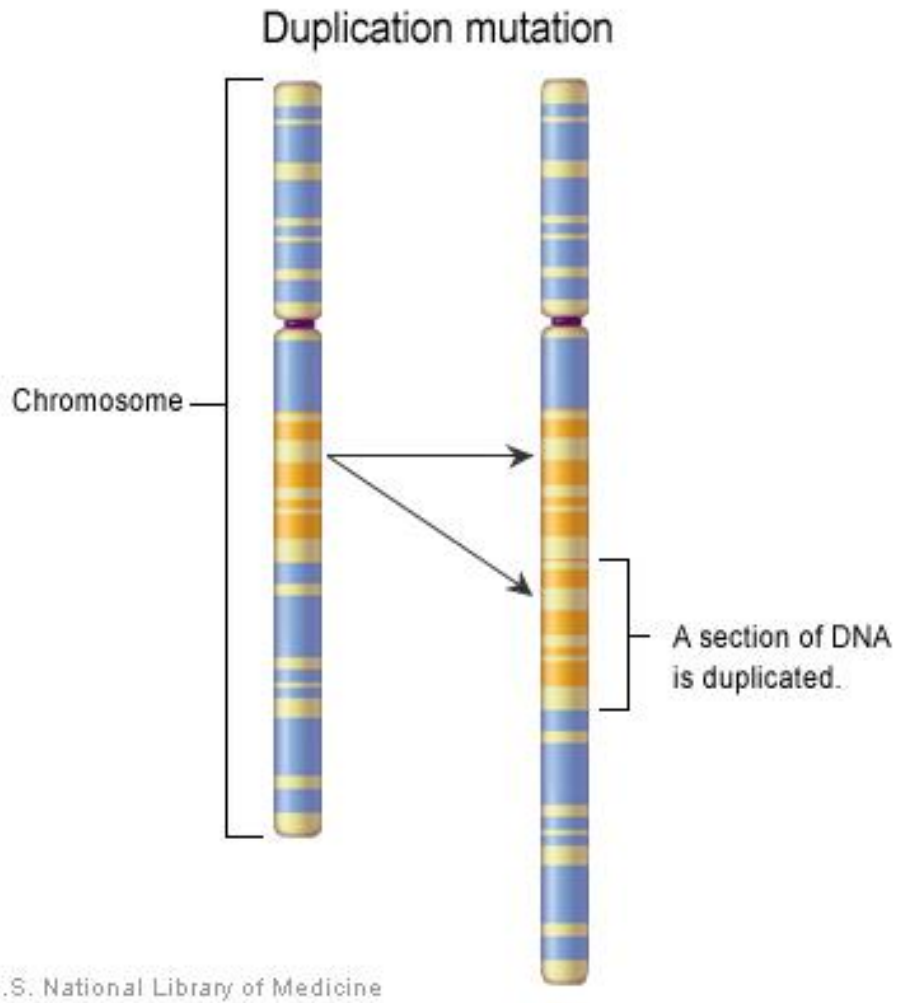
- A deletion changes the number of DNA bases by removing a piece of DNA.
- Small deletions may remove one or a few base pairs within a gene, while larger deletions can remove an entire gene or several neighboring genes.
- The deleted DNA may alter the function of the resulting protein(s).



U.S. National Library of Medicine

Types of Gene Mutations : Duplication

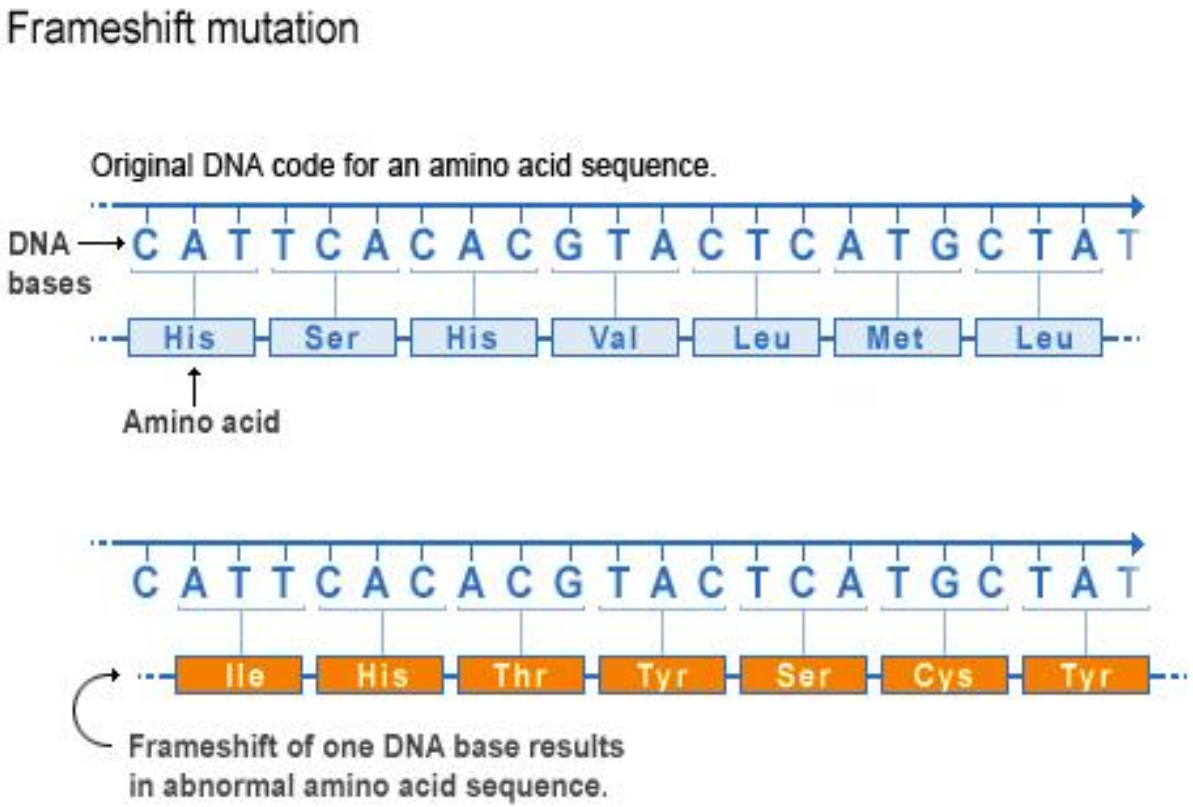
- A duplication consists of a piece of DNA that is abnormally copied one or more times. This type of mutation may alter the function of the resulting protein.



U.S. National Library of Medicine

Types of Gene Mutations : Frameshift mutation

- This type of mutation occurs when the addition or loss of DNA bases changes a gene's reading frame.
- A reading frame consists of groups of 3 bases that each code for one amino acid. A frameshift mutation shifts the grouping of these bases and changes the code for amino acids.

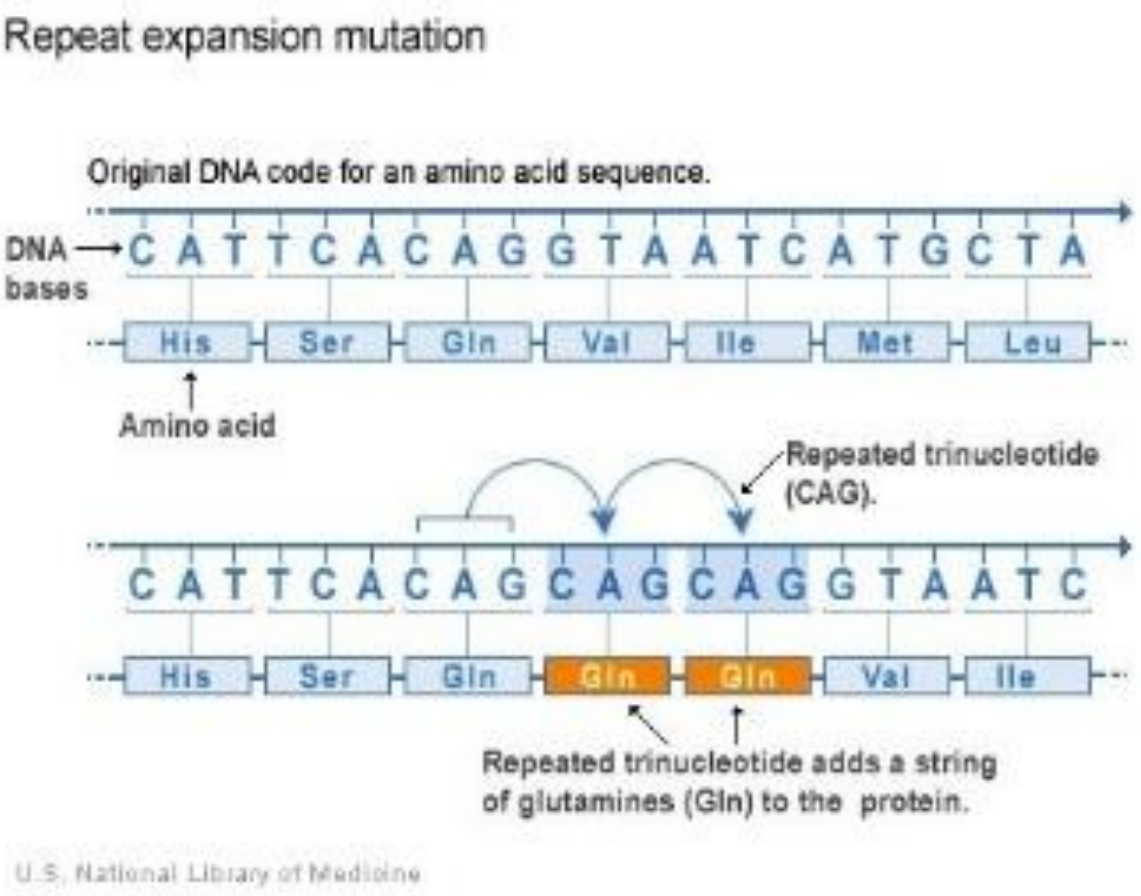


The resulting protein is usually nonfunctional.

Insertions, deletions, and duplications can all be frameshift mutations.

Types of Gene Mutations : Repeat expansion

- Nucleotide repeats are short DNA sequences that are repeated a number of times in a row.
- For example, a trinucleotide repeat is made up of 3-base-pair sequences, and a tetranucleotide repeat is made up of 4-base-pair sequences.



A repeat expansion is a mutation that increases the number of times that the short DNA sequence is repeated.

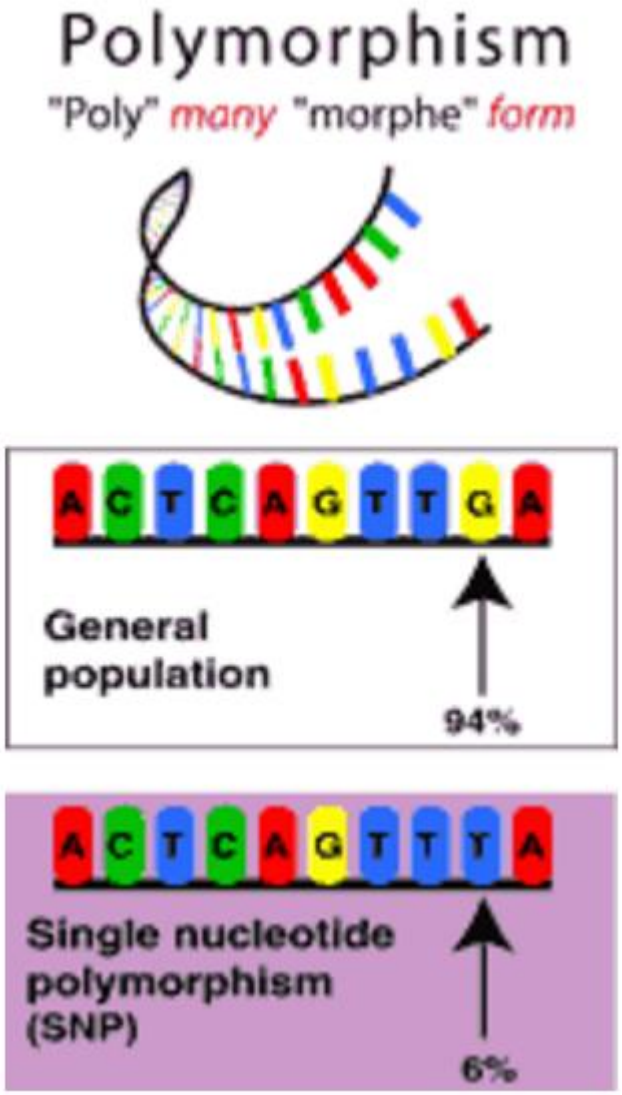
This type of mutation can cause the resulting protein to function improperly.

Variations

- Individuals of a species have similar characteristics but they are rarely identical, the difference between them is called variation.
- Genetic variation is a result of subtle differences in our DNA
- Mutation vs variation
  - A mutation is a change that occurs in the genome of an individual organism
  - Variation is something that occurs within a species
- Genetic variation results in different forms, or alleles, of genes.
- For example, if we look at eye colour, people with blue eyes have one allele of the gene for eye colour, whereas people with brown eyes will have a different allele of the gene.
- Eye colour, skin tone and face shape are all determined by our genes so any variation that occurs will be due to the genes inherited from our parents.

Variations

- Genome variations include mutations and polymorphisms
- Polymorphism (a term that comes from the Greek words "poly," or "many," and "morphe," or "form")
- It is a DNA variation in which each possible sequence is present in at least 1 percent of people.
- For example, a place in the genome where 93 percent of people have a T and the remaining 7 percent have an A is a polymorphism.



RWTH Informatik 5 | Ahornstr. 55 D-52056 Aachen Germany |
Tel +49/241/8021501 | Fax +49/241/8022321 | http://dbis.rwth-aachen.de/cms

Variation in the human genome: Single base changes

- The most common type of variation present in the human genome is the single nucleotide polymorphism (SNP)
- A single nucleotide in the genome differs between individuals (or paired chromosomes), such that there are (at least) two alleles.
- Numerous common SNPs from many individuals of different ethnic backgrounds were typed by the International HapMap Project, and form a valuable database of common variation and a resource for identifying suitable SNPs for genome-wide association studies.
- SNPs that occur in <1% of the population are classified as rare variants or mutations, and may have a profound phenotypic effect.
- Single bases may also be added (insertions) or removed (deletions), which can have a substantial effect in coding regions where they may result in a ‘frameshift’ in the downstream genetic code.

Variation in the human genome: Single base changes

- In human beings, 99.9 percent bases are same.
- Remaining 0.1 percent makes a person unique.
- These variations can be:
  - Harmless (change in phenotype)
  - Harmful (diabetes, cancer, heart disease, Huntington's disease, and hemophilia)
  - Latent (variations found in coding and regulatory regions, are not harmful on their own, and the change in each gene only becomes apparent under certain conditions e.g. susceptibility to lung cancer)

case

…. GCCGTTGAC…. .
…. GCCATTGAC…. .

control

…. GCCATTGAC…. .
…. GCCATTGAC…. .

RWTH Informatik 5 | Ahornstr. 55 D-52056 Aachen Germany |
Tel +49/241/8021501 | Fax +49/241/8022321 | http://dbis.rwth-aachen.de/cms

Variation in the human genome: Multiple base changes

- Genomic variation can also be caused by multiple base changes,
- The most common of which is in the form of INDELs - INsertions and DELetions (which may be co-localised) that range in size from 1-1000 base pairs.
- Larger insertions or deletions are referred to as copy number variants (CNVs), and include both common and rare variants, though CNVs greater than 5Mb are rare.
- Translocation, inversion and duplication events also contribute towards CNVs, which can result in large structural changes affecting many genes that may be visible under microscope.