The Central Dogma of Molecular Biology

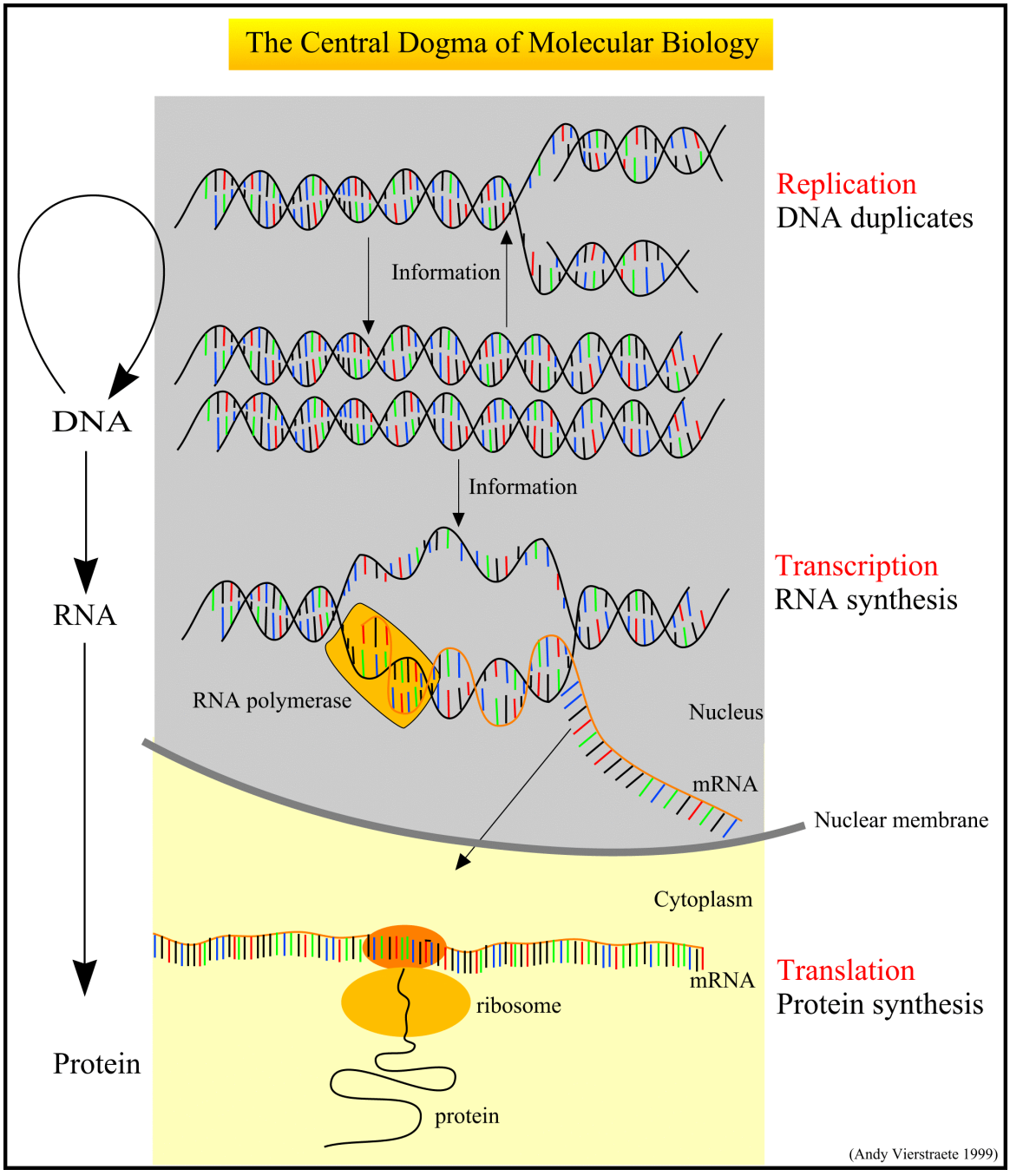


Figure 1 : The Central Dogma of molecular biology.

DNA contains the complete genetic information that defines the structure and function of an organism. Proteins are formed using the genetic code of the DNA. Three different processes are responsible for the inheritance of genetic information and for its conversion from one form to another :

1. **Replication** : a double stranded nucleic acid is duplicated to give identical copies. This process perpetuates the genetic information.

2. **Transcription** : a DNA segment that constitutes a gene is read and transcribed into a

single stranded sequence of RNA. The RNA moves from the nucleus into the cytoplasm.

3. **Translation** : the RNA sequence is translated into a sequence of amino acids as the protein is formed. During translation, the ribosome reads three bases (a codon) at a time from the RNA and translates them into one amino acid

In eucariotic cells, the second step (transcription) is necessary because the genetic material in the nucleus is physically separated from the site of protein synthesis in the cytoplasm in the cell. Therefore, it is not possible to translate DNA directly into protein, but an intermediary must be made to carry the information from one compartment to an other

What is DNA ?

1. Nucleotides are the building stones of DNA.

There are 4 different nucleotides :

- dATP : deoxyadenosine triphosphate
- dGTP : deoxyguanosine triphosphate
- dTTP : deoxythymidine triphosphate
- dCTP : deoxycytidine triphosphate

For convenience, these 4 nucleotides are called dNTP's (deoxynucleoside triphosphates). A nucleotide is made of three major parts : a **nitrogen base**, a **sugar** molecule and a **triphosphate**. Only the nitrogen base is different in the 4 nucleotides.

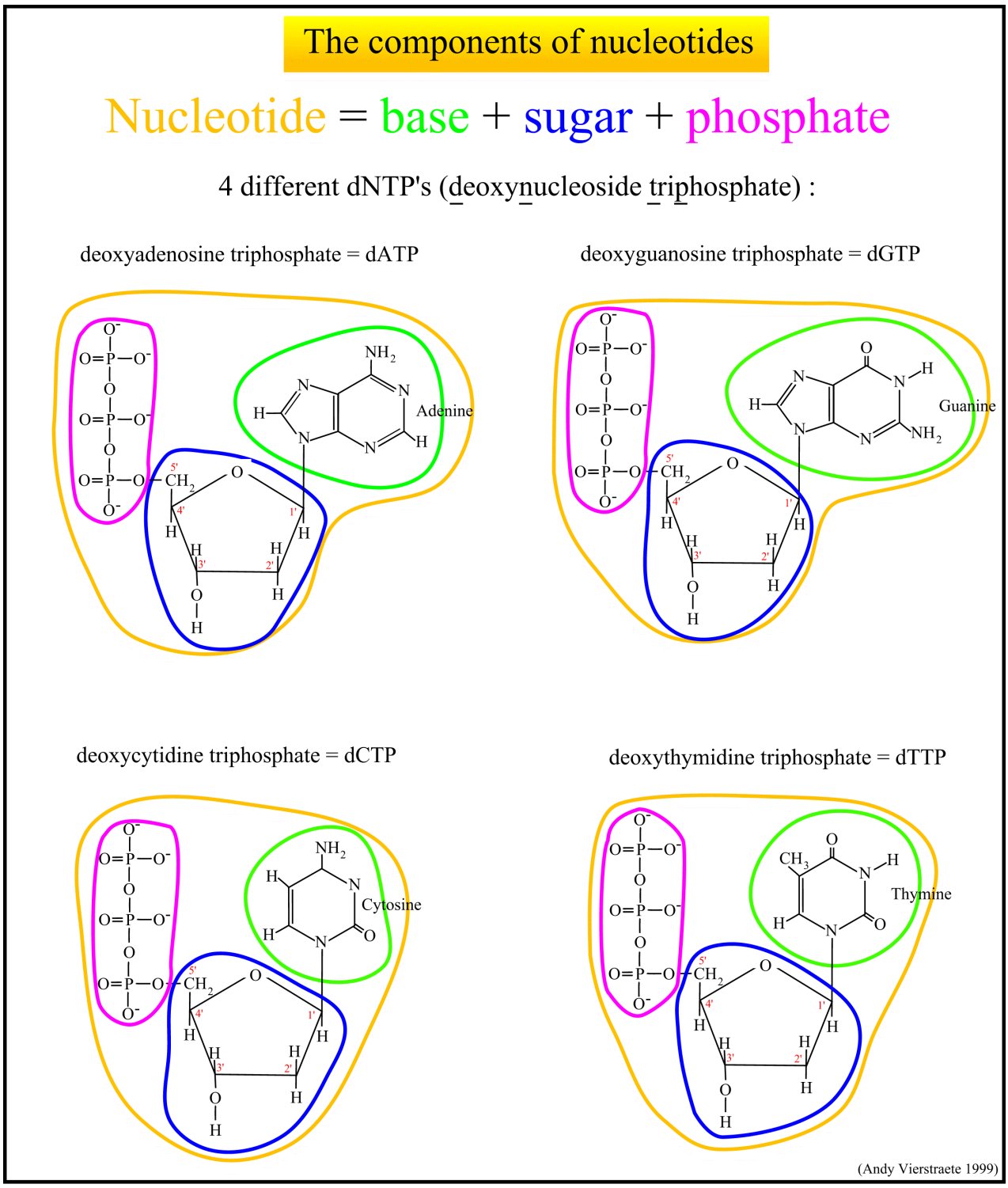


Figure 2 : The components of nucleotides.

2. How do the nucleotides form a DNA chain ?

DNA is formed by coupling the nucleotides between the phosphate group from a nucleotide (which is positioned on the **5th C-atom** of the sugar molecule) with the hydroxyl on **the 3rd C-atom** on the sugar molecule of the previous nucleotide. To accomplish this, a diphosphate molecule is split off (and releases energy). This means that new nucleotides are **always added on the 3' side** of the chain.

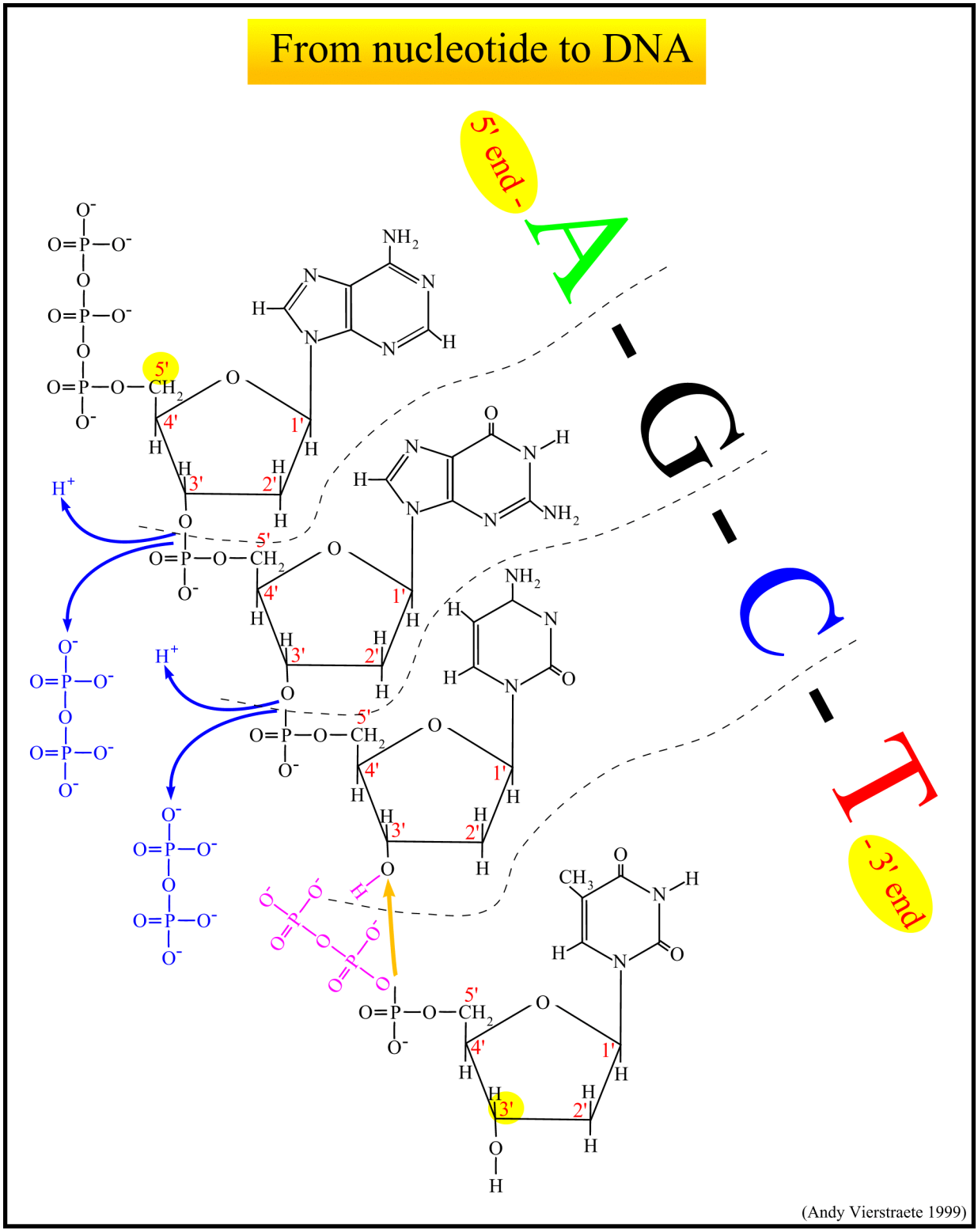


Figure 3 : From nucleotide to DNA.

DNA in a cell

1. Which organelles contain DNA ?

Eucariotic cells contain several organelles. The **nucleus** contains most of the DNA in a cell and this DNA is called the **chromosomal DNA**. It is separated from the rest of the cell (cytoplasm) by a double layer of membrane. The **mitochondria**, which have a role in the oxidative degradation of nutrient molecules, also contain DNA, called the **mitochondrial DNA**. Eucariotic cells that are capable of photosynthesis contain **chloroplasts** with **chloroplast DNA**.

Size of genetic material

Type of DNA	Organism	size in base pairs
chromosomal DNA	mammals	6×10^9
	plants	2×10^8 - 2×10^{11}
	fungi	2×10^7 - 2×10^8
mitochondrial DNA	animals	16×10^3 - 19×10^3
	higher plants	150×10^3 - 250×10^4
	fungi	17×10^3 - 78×10^3
	green alga	16×10^3
	protozoa	22×10^3 - 40×10^3
chloroplast DNA	higher plants	120×10^3 - 200×10^3
	green alga	180×10^3

To have an idea of the size of this : every million bases take up a linear space of 0,34 mm. So when you take one human cell, uncoil all the chromosomal DNA and put it on a line, you would have 204 cm of DNA (a human cell contains in total 6×10^9 nucleotide pairs). To store this information on paper, you need a few pages of A4 paper : one page stores 3100 letters on one side (font courier 12), so 6×10^9 nucleotides, printed on both sides of a sheet, would need 967.742 pages (you'll have a pile of 120,96 m paper !!!). You'd better start printing immediately...

2. There are three types of genes :

1. Protein-coding genes : these are transcribed into RNA and then translated into proteins.
2. RNA-specifying genes : these are only transcribed into RNA.
3. Regulatory genes : according to a narrow definition, these include only untranscribed sequences.

The first two types are also called 'structural genes'.

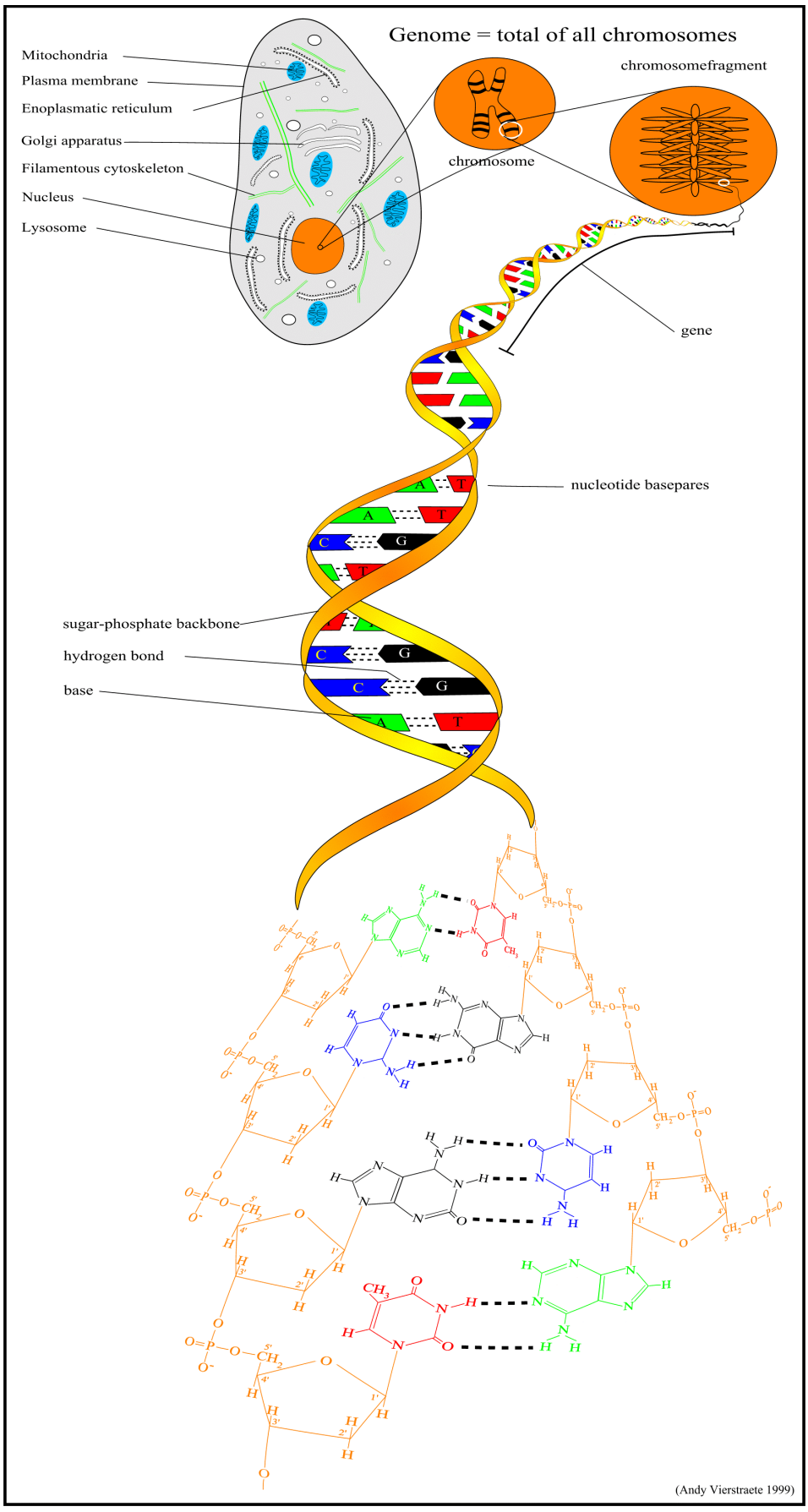


Figure 4 : The DNA in a cell.

Figure 5 shows the different parts of the genes that codes for ribosomal DNA (rDNA) (important in protein synthesis (see the Central Dogma figure)). In eucariotic cells, there are 50 – 5000 identical copies of the genes that specify **18S** (small sub unit (SSU)), **5.8S** and **28S** (large sub unit (LSU)) in the 10 million ribosomes. These genes are tandem wise arranged in large blocks on one or more chromosomes and are separated from each other by **non transcribed spacer** (NTS). These genes are transcribed into a single RNA precursor from which the mature rRNA molecules are released by cleavage. This process removes the **external transcribed spacer** (ETS), **internal transcribed spacer 1** (ITS1) and **internal transcribed spacer 2** (ITS2) out of the RNA precursor and results in 3 rRNA molecules : 18S rRNA, 5.8S rRNA and 28S rRNA.

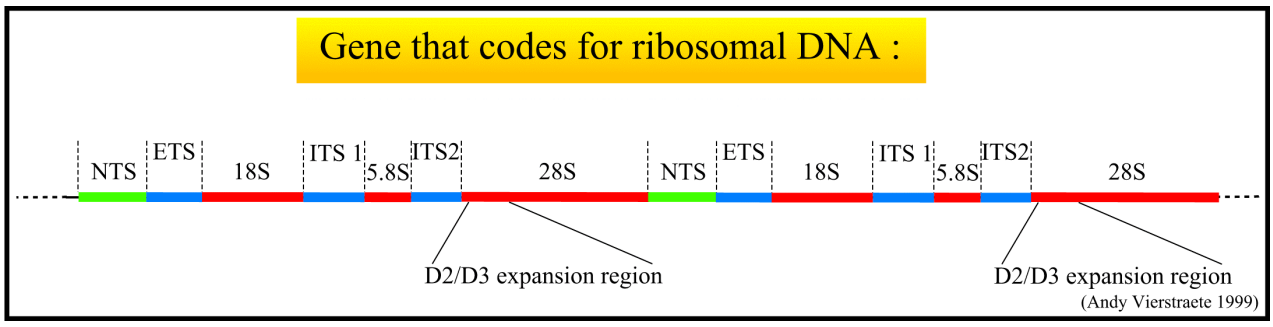


Figure 5 : The genes that codes for ribosomal DNA.

Preparation of gDNA

In eucariotic cells, the DNA is isolated in the nucleus, mitochondria (and chloroplasts). To extract the DNA, it is necessary to remove all the barriers around it. Mostly, **proteinase K** is used to dissolve the cell membrane and nuclear membrane, and it dissociates the proteins from the DNA. After this step, **phenol** is added which results in 3 phases in the tube :

1. The **aqueous phase** with the **DNA**, can be precipitated with ethanol.
2. The interface with denatured proteins.
3. The phenol phase with the dissolved proteins and fats.

When the DNA is prepared, it can be checked on gel to verify the concentration and quality (sometimes, the DNA is broken down by the enzyme DNAase when the organism isn't preserved correctly for genetic research.)

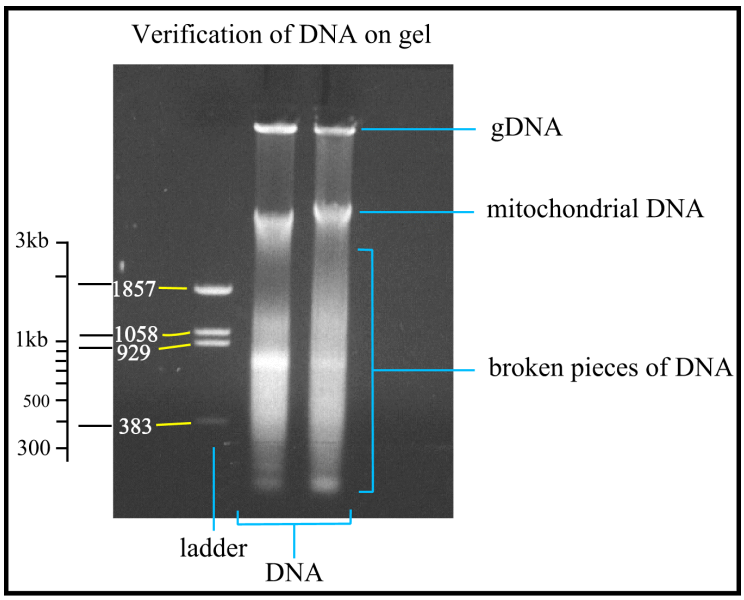


Figure 6 : Verification of the prepared DNA on gel.

Principle of the PCR

The purpose of a PCR (Polymerase Chain Reaction) is to make a huge number of copies of a gene. This is necessary to have enough starting template for sequencing.

1. **The cycling reactions :**

   There are three major steps in a PCR, which are repeated for 30 or 40 cycles. This is done on an automated cycler, which can heat and cool the tubes with the reaction mixture in a very short time.

   1. **Denaturation** at 94°C :

      During the denaturation, the double strand melts open to single stranded DNA, all enzymatic reactions stop (for example : the extension from a previous cycle).

   2. **Annealing** at 54°C :

      The primers are jiggling around, caused by the Brownian motion. Ionic bonds are constantly formed and broken between the single stranded primer and the single stranded template. The more stable bonds last a little bit longer (primers that fit exactly) and on that little piece of double stranded DNA (template and primer), the polymerase can attach and starts copying the template. Once there are a few bases built in, the ionic bond is so strong between the template and the primer, that it does not break anymore.

   3. **extension** at 72°C :

      This is the ideal working temperature for the polymerase. The primers, where there

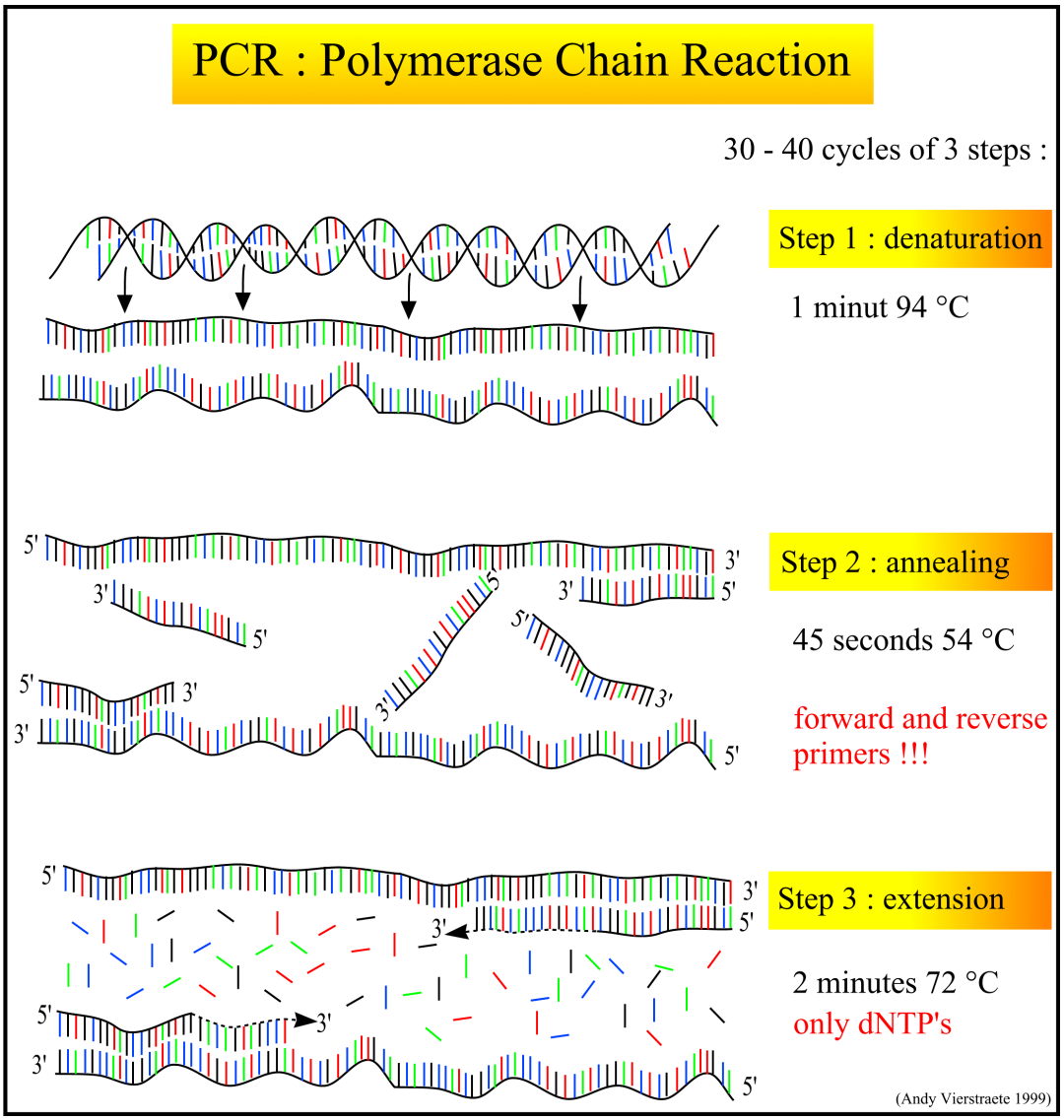


Figure 7 : The different steps in PCR.

are a few bases built in, already have a stronger ionic attraction to the template than the forces breaking these attractions. Primers that are on positions with no exact match, get loose again (because of the higher temperature) and don't give an extension of the fragment.
The bases (complementary to the template) are coupled to the primer on the 3' side (the polymerase adds dNTP's from 5' to 3', reading the template from 3' to 5' side, bases are added complementary to the template)

Because both strands are copied during PCR, there is an **exponential** increase of the number of copies of the gene. Suppose there is only one copy of the wanted gene before the cycling starts, after one cycle, there will be 2 copies, after two cycles, there will be 4 copies, three cycles will result in 8 copies and so on.

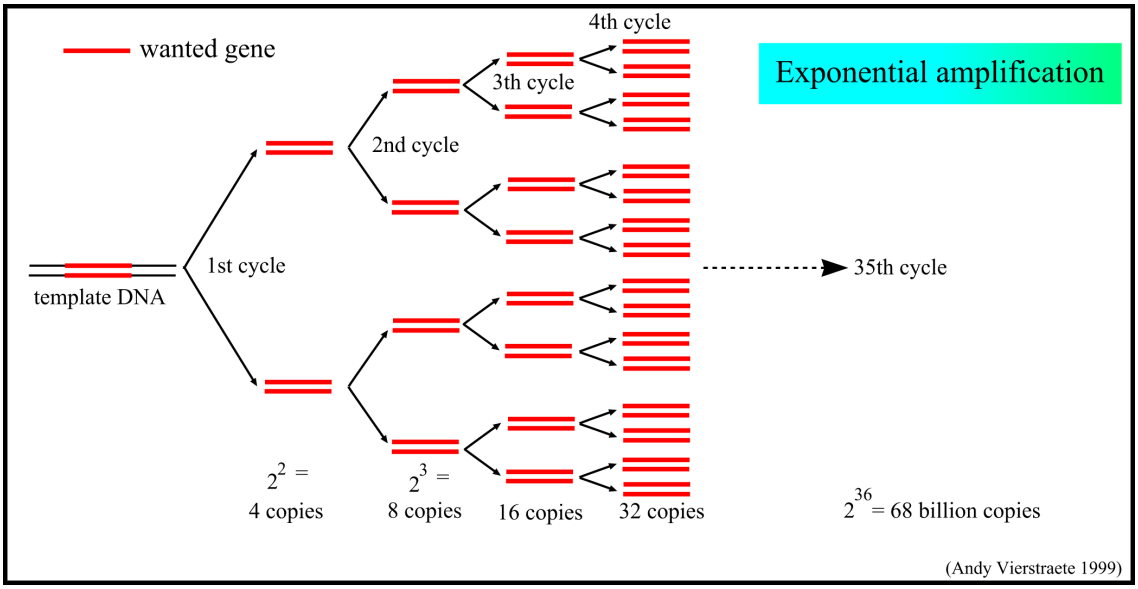


Figure 8 : The exponential amplification of the gene in PCR.

2. Is there a gene 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. There is a product formed.
   Though biochemistry is an exact science, not every PCR is successful. There is for example a possibility that the quality of the DNA is poor, that one of the primers doesn't fit, or that there is too much starting template

2. The product is of the right size
   It is possible that there is a product, for example a band of 500 bases, but the expected gene should be 1800 bases long. In that case, one of the primers probably fits on a part of the gene closer to the other primer. It is also possible that both primers fit on a totally different gene.

3. Only one band is formed.
   As in the description above, it is possible that the primers fit on the desired locations, and also on other locations. In that case, you can have different bands in one lane on a gel.

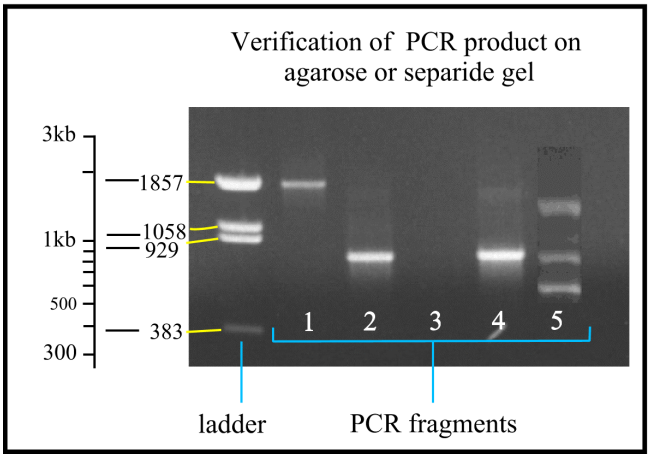


Figure 9 : Verification of the PCR product on gel.

The ladder is a mixture of fragments with known size to compare with the PCR fragments. Notice that the distance between the different fragments of the ladder is logarithmic. Lane 1 : PCR fragment is approximately 1850 bases long. Lane 2 and 4 : the fragments are approximately 800 bases long. Lane 3 : no product is formed, so the PCR failed. Lane 5 : multiple bands are formed because one of the primers fits on different places.

Principle of sequencing

(This is only an explanation of the method used for sequencing on an automated sequencer ABI 377)

The purpose of sequencing is to determine the order of the nucleotides of a gene. For sequencing, we don't start from gDNA (like in PCR) but mostly from PCR fragments or cloned genes.

1. **The sequencing reaction :**

   There are three major steps in a sequencing reaction (like in PCR), which are repeated for 30 or 40 cycles.

   1. **Denaturation** at 94°C :

      During the denaturation, the double strand melts open to single stranded DNA, all enzymatic reactions stop (for example : the extension from a previous cycle).

   2. **Annealing** at 50°C :

      In sequencing reactions, only one primer is used, so there is only one strand copied (in PCR : two primers are used, so two strands are copied). The primer is jiggling around, caused by the Brownian motion. Ionic bonds are constantly formed and broken between the single stranded primer and the single stranded template. The more stable bonds last a little bit longer (primers that fit exactly) and on that little piece of double stranded DNA (template and primer), the polymerase can attach and starts copying the template. Once there are a few bases built in, the ionic bond

is so strong between the template and the primer, that it does not break anymore.

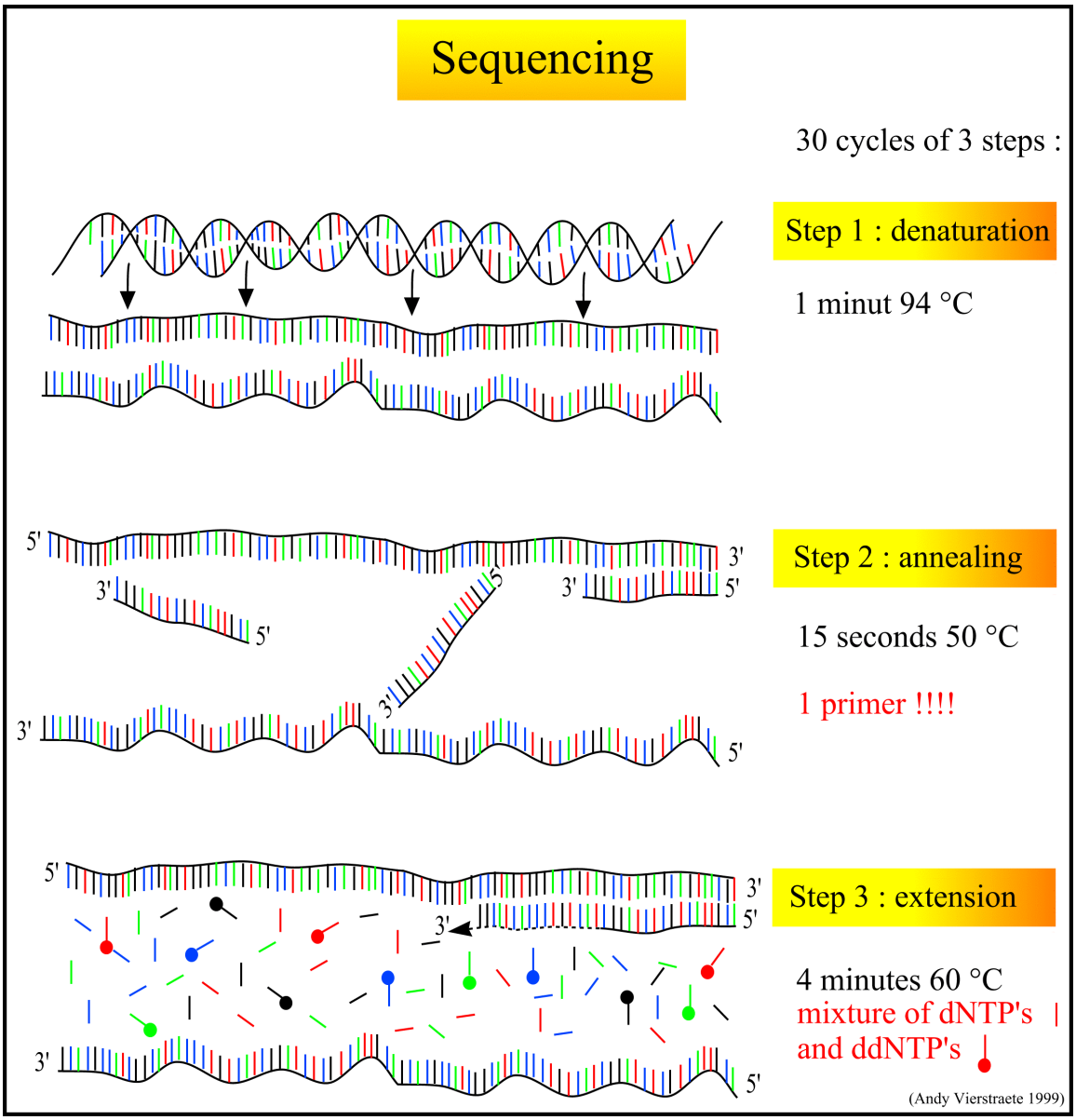


Figure 10 : The different steps in sequencing.

3. **extension** at 60°C :

This is the ideal working temperature for the polymerase (normally it is 72 °C, but because it has to incorporate ddNTP's which are chemically modified with a fluorescent label, the temperature is lowered so it has time to incorporate the 'strange' molecules. The primers, where there are a few bases built in, already have a stronger ionic attraction to the template than the forces breaking these attractions. Primers that are on positions with no exact match, come loose again and don't give an extension of the fragment.
The bases (complementary to the template) are coupled to the primer on the 3'side (adding dNTP's or ddNTP's from 5' to 3', reading from the template from 3' to 5'

side, bases are added complementary to the template)
When a ddNTP is incorporated, the extension reaction stops because a ddNTP contains a H-atom on the 3rd carbon atom (dNTP's contain a OH-atom on that position). Since the ddNTP's are fluorescently labeled, it is possible to detect the color of the last base of this fragment on an automated sequencer.

Because only one primer is used, only one strand is copied during sequencing, there is a **linear** increase of the number of copies of one strand of the gene. Therefore, there has to be a large amount of copies of the gene in the starting mixture for sequencing. Suppose there are 1000 copies of the wanted gene before the cycling starts, after one cycle, there will be 2000 copies : the 1000 original templates and 1000 complementary strands with each one fluorescent label on the last base, after two cycles, there will be 2000 complementary strands, three cycles will result in 3000 complementary strands and so on.

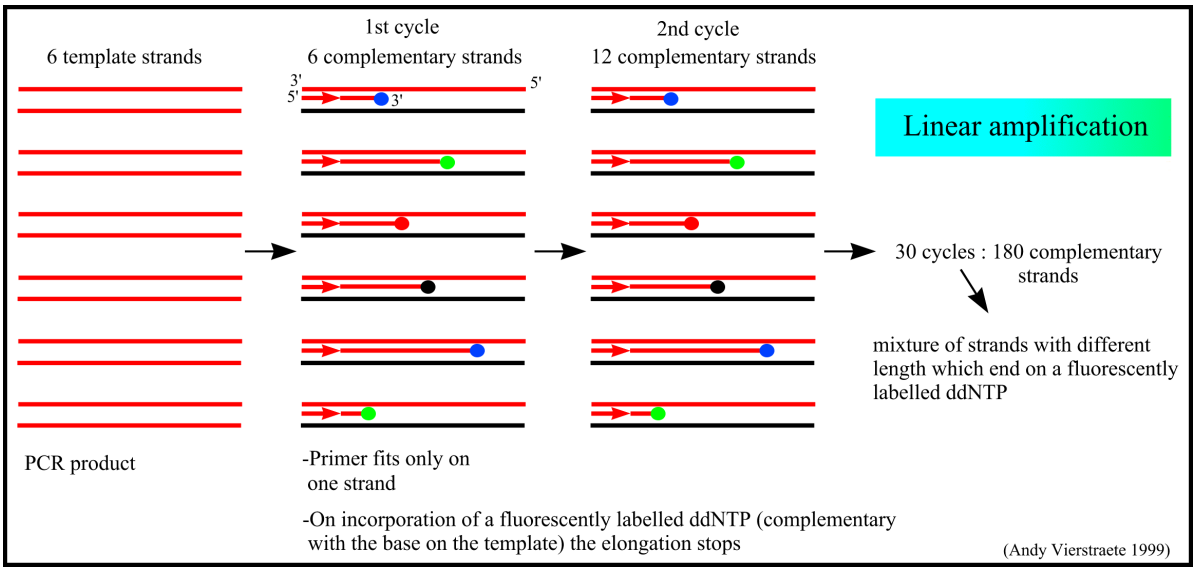


Figure 11 : The linear amplification of the gene in sequencing.

2. Separation of the molecules :

After the sequencing reactions, the mixture of strands, all of different length and all ending on a fluorescently labeled ddNTP have to be separated; This is done on an acrylamide gel, which is capable of separating a molecule of 30 bases from one of 31 bases, but also a molecule of 750 bases from one of 751 bases. All this is done with gel electrophoresis. DNA has a negative charge and migrates to the positive side. Smaller fragments migrate faster, so the DNA molecules are separated on their size.

3. Detection on an automated sequencer :

The fluorescently labeled fragments that migrate trough the gel, are passing a laser beam at the bottom of the gel. The laser excites the fluorescent molecule, which sends out light of a distinct color. That light is collected and focused by lenses into a spectrograph. Based on the wavelength, the spectrograph separates the light across a CCD camera (**c**harge **c**oupled **d**evice). Each base has its own color, so the sequencer can detect the order of the bases in the sequenced gene.

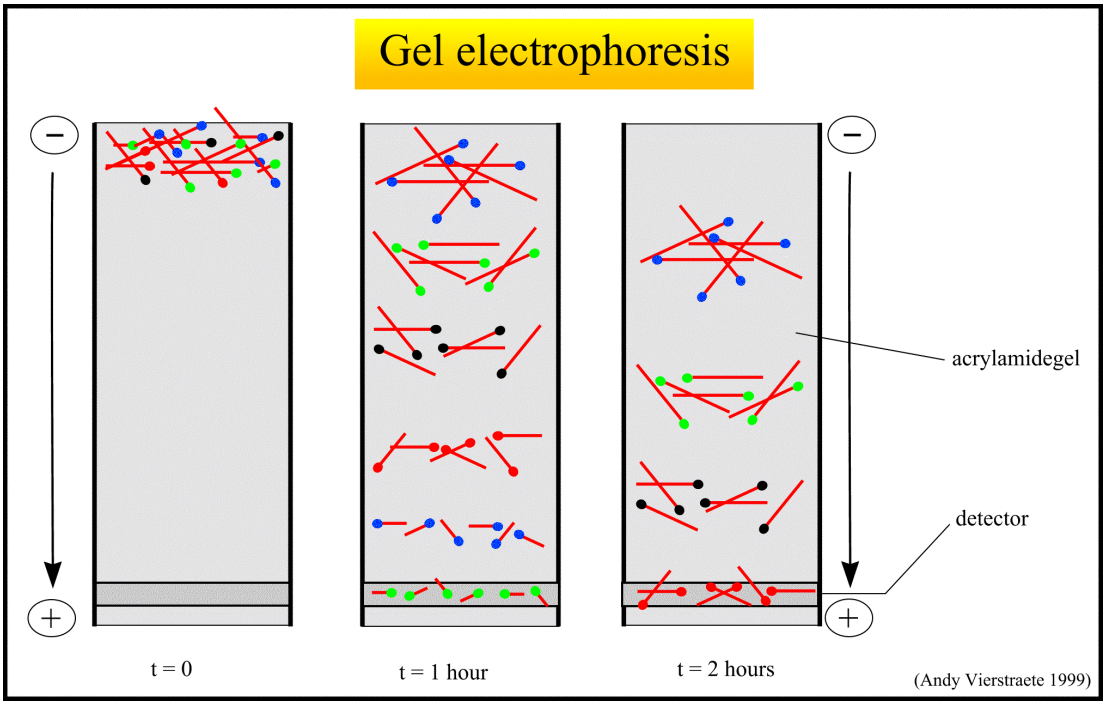


Figure 12 : The separation of the molecules with electrophoresis.

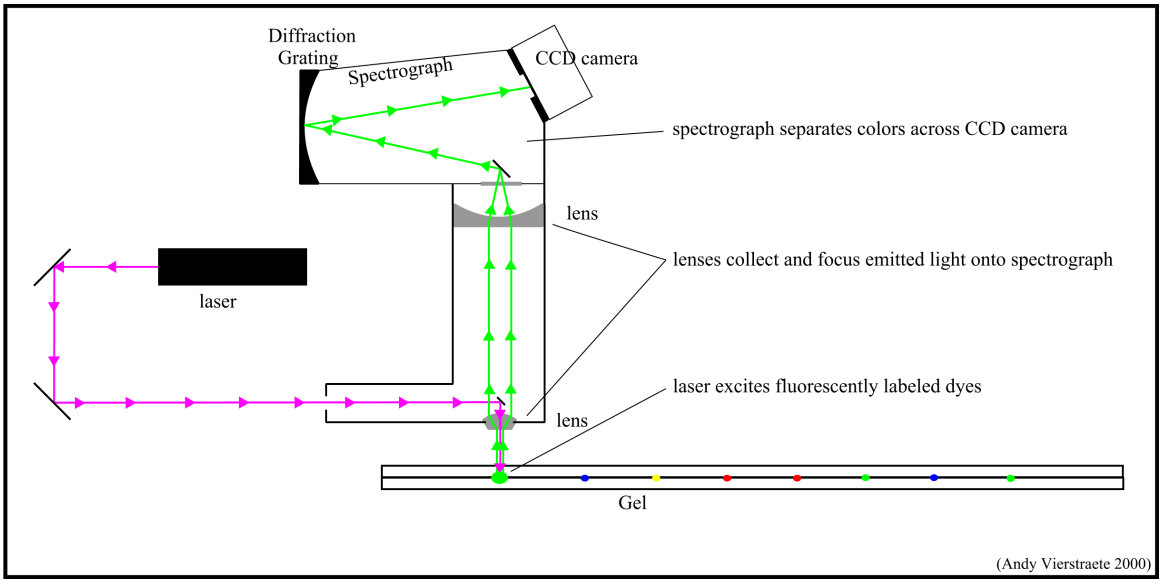


Figure 13 : The scanning and detection system on the ABI Prism 377 sequencer.

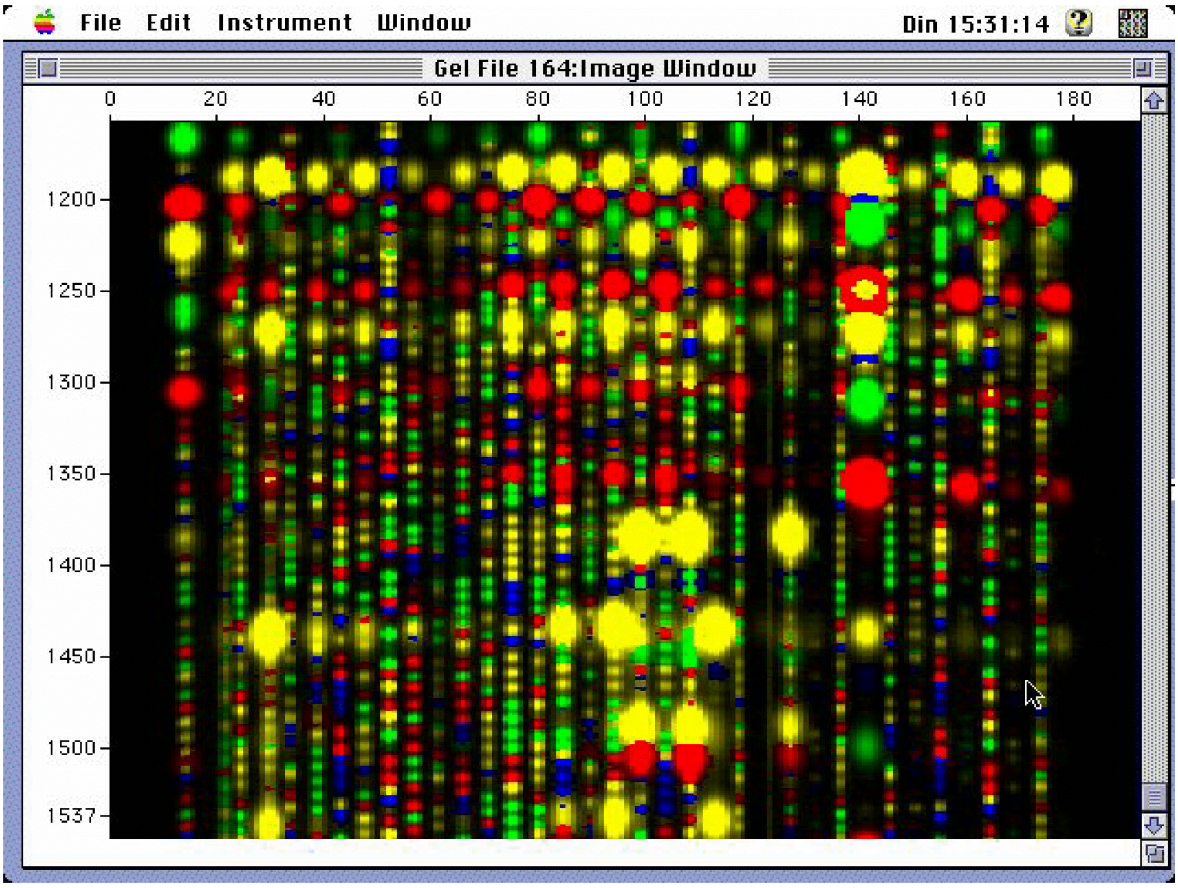


Figure 14 : A snapshot of the detection of the molecules on the sequencer.

4. Assembling of the sequenced parts of a gene :

For publication purposes, each sequence of a gene has to be confirmed in both directions. To accomplish this, the gene has to be sequenced with forward and reverse primers. Since it is only possible to sequence a part of 750 till 800 bases in one run, a gene of, for example 1800 bases, has to be sequenced with internal primers. When all these fragments are sequenced, a computer program tries to fit the different parts together and assembles the total gene sequence.

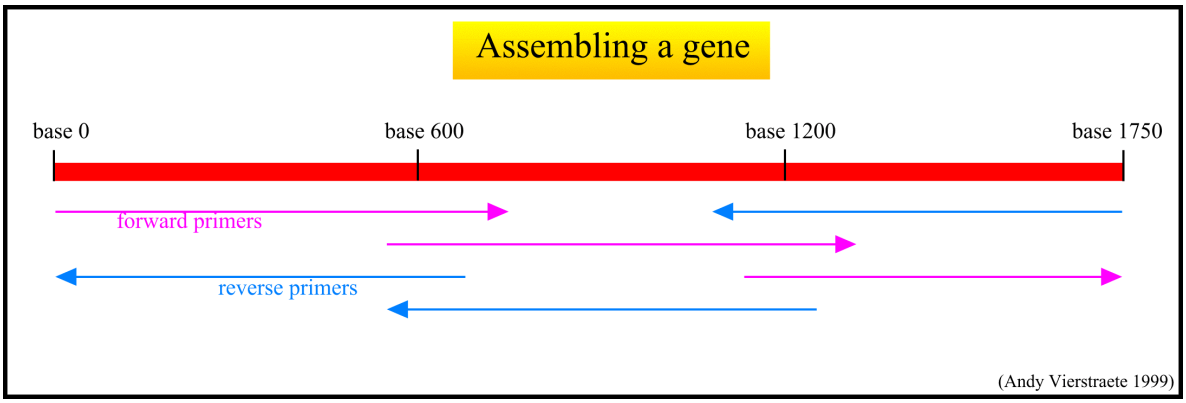


Figure 15 : The assemblage of the gene.

Sequence alignment

To compare two or more sequences, it is necessary to align the **conserved** and **unconserved** residues across all the sequences (identification of locations of insertions and deletions that have occurred since the divergence of a common ancestor). These **residues form a pattern** from which the relationship between sequences can be determined with phylogenetic programs. When the sequences are aligned, it is possible to identify locations of insertions or deletions since their divergence from their common ancestor. There are three possibilities :

- The bases match : this means that there is no change since their divergence.
- The bases mismatch : this means that there is a substitution since their divergence.
- There is a base in one sequence, no base in the other : there is an insertion or a deletion since their divergence.

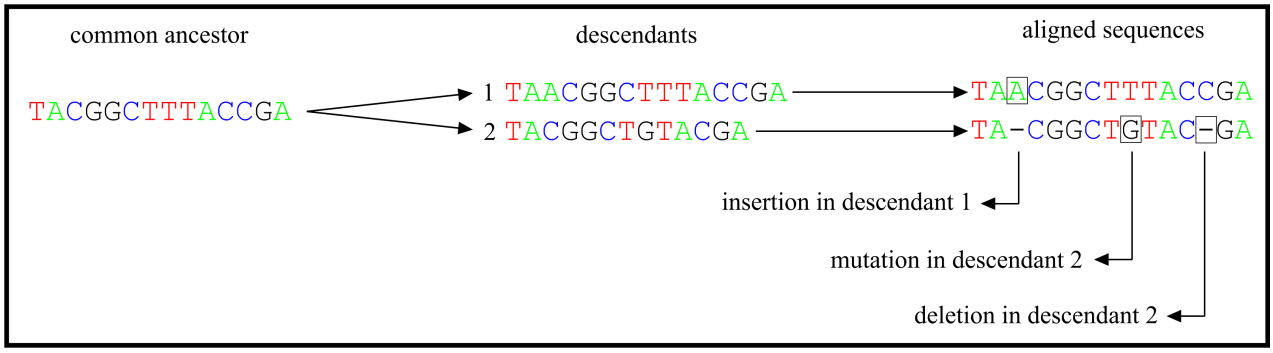


Figure 16 : The comparison of sequences.

A good alignment is important for the next step : the construction of phylogenetic trees. The alignment will affect the distances between 2 different species and this will influence the inferred phylogeny.

There are several programs available on the net for aligning sequences. These are all based on different mathematical models to compare two or more sequences with the most optimal score for matching bases with a minimum number of gaps inserted (because you can insert a huge amount of gaps, so every base will match an other).

Example : two sequences :

```
TCAGACGATTG
TCGGAGCTG
```

How can we get the best alignment ? There are several possibilities :

1. Reduce the number of mismatches :

```
TCAG-ACG-ATTG
||  |  |  |  |        0 mismatches  7 matches  6 gaps
TC-GGA-GC-T-G
```

2. Reduce the number of gaps :

```
TCAGACGATTG
|| ||               5 mismatches  4 matches  2 gaps
TCGGAGCTG--
```

3. Reduce neither the number of gaps nor the number of mismatches :

```
TCAG-ACGATTG
|| | | | |          2 mismatches  6 matches  4 gaps
TC-GGA-GCTG-
```

4. Same as 3. but one base (or gap) moved :

```
TCAG-ACGATTG
|| | | | | |        1 mismatch    7 matches  4 gaps
TC-GGA-GCT-G
```

Which of these is now the best alignment ??

There are several alignment algorithms to choose the best alignment. Let's use a simple one in this example :

$D = y + \text{sum}(w_k z_k)$

with :

D = distance
y : number of mismatches
w : penalty for gaps of length k
z : number of gaps of length k

Take gap penalty for gap length 1 = 2
Take gap penalty for gap length 2 = 6 (short gaps occur more frequent than long gaps)

in 1. : 0 + {(2 x 6) + (6 x 0)} = 12
in 2. : 5 + {(2 x 0) + (6 x 1)} = 11
in 3. : 2 + {(2 x 4) + (6 x 0)} = 10
in 4. : 1 + {(2 x 4) + (6 x 0)} = 9

We choose alignment 4 because it has the minimum distance.

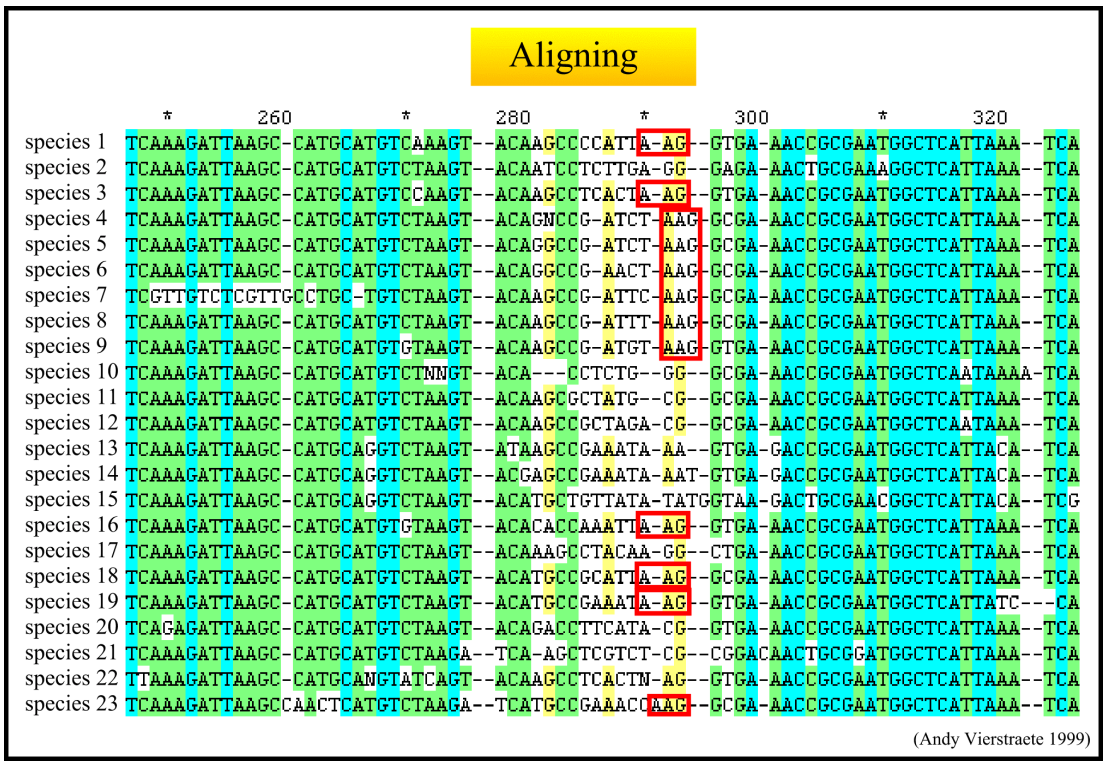


Figure 17 : The alignment of sequences. This is done with Clustalw 1.74, and as you can see, the more variable areas are not optimally aligned (indicated with red boxes). Therefore it is mostly necessary to improve the alignment by hand. In this case, it is obvious to improve the alignment, but in other cases it could be more difficult to make improvements.

Phylogenetics

1. Purpose of phylogenetics :

- With the aid of sequences, it should be possible to find the genealogical ties between organisms. Experience learns that closely related organisms have similar sequences, more distantly related organisms have more dissimilar sequences. One objective is to reconstruct the **evolutionary relationship** between species.
- An other objective is to estimate the **time of divergence** between two organisms since they last shared a common ancestor.

2. Disclaimers :

- The theory and practical applications of the different models are not universally accepted.
- With one dataset, different software packages can give different results. Changes in the dataset can also give different results. Therefore it is important to have a good alignment to start with.
- Trees based on an alignment of a gene represent the relationship between genes and this is not necessarily the same relationship as between the whole organisms. If trees are calculated based on different genes from organisms, it is possible that these trees result in

different relationships.

3. Terminology :

- **node :** a node represents a taxonomic unit. This can be a taxon (an existing species) or an ancestor (unknown species : represents the ancestor of 2 or more species).
- **branch :** defines the relationship between the taxa in terms of descent and ancestry.
- **topology :** is the branching pattern.
- **branch length :** often represents the number of changes that have occurred in that branch.
- **root :** is the common ancestor of all taxa.
- **distance scale :** scale which represents the number of differences between sequences (e.g. 0.1 means 10 % differences between two sequences)

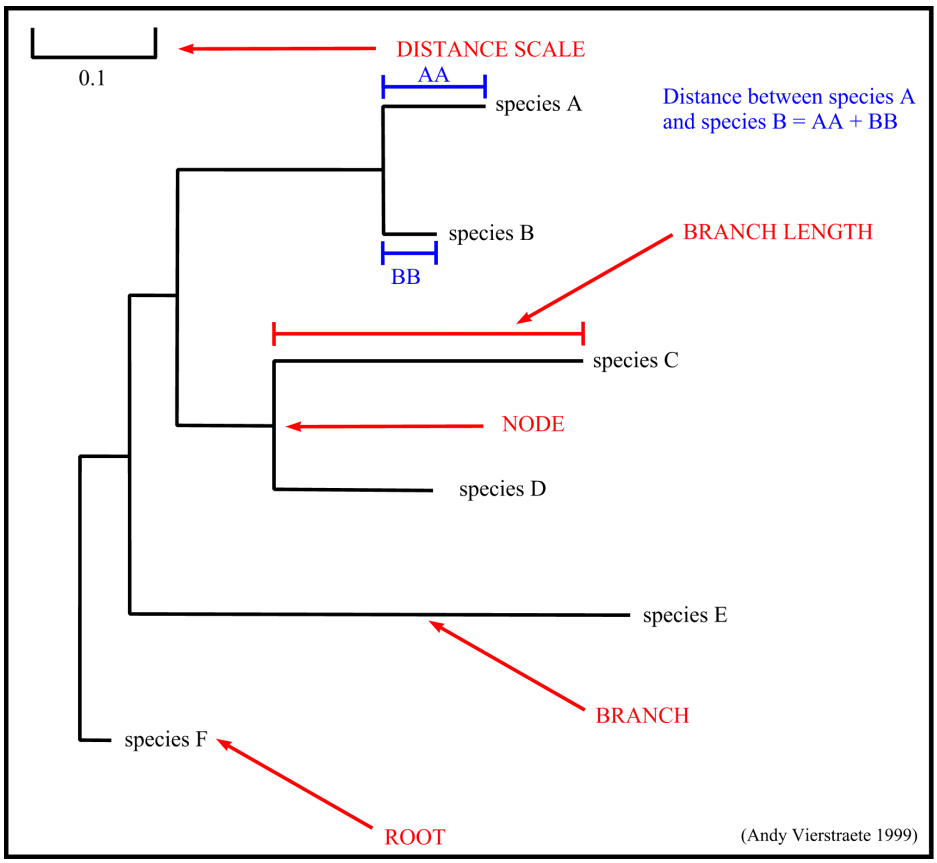


Figure 18 : The tree terminology.

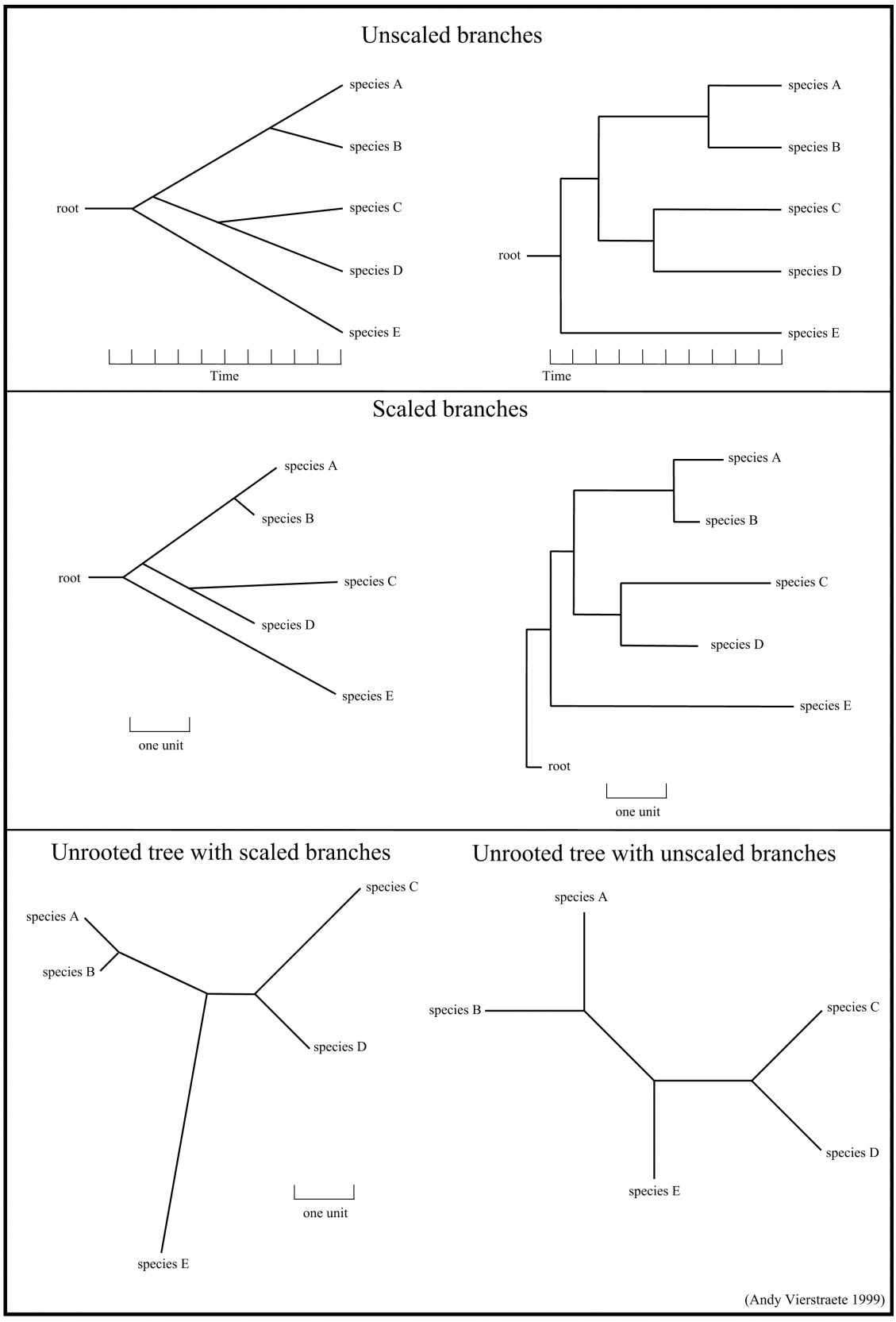


Figure 19 : Some possibilities for drawing a tree. (these are just a few examples, there are a lot of variations possible)

4. Possible ways of drawing a tree :

Trees can be drawn in different ways. There are trees with **unscaled branches** and with **scaled branches**.

- Unscaled branches : the length is not proportional to the number of changes. Sometimes, the number of changes are indicated on the branches with numbers. The nodes represents the divergence event on a time scale.

- Scaled branches : the length of the branch is proportional to the number of changes. The distance between 2 species is the sum of the length of all branches connecting them.

It is also possible to draw these trees with or without a root. For **rooted trees**, the root is the common ancestor. For each species, there is a unique path that leads from the root to that species. The direction of each path corresponds to evolutionary time. An **unrooted tree** specifies the relationships among species and does not define the evolutionary path.

5. Methods of phylogenetic analysis :

There are two major groups of analyses to examine phylogenetic relationships between sequences :

1. Phenetic methods : trees are calculated by *similarities of sequences* and are based on **distance** methods. The resulting tree is called a **dendrogram** and does not necessarily reflect evolutionary relationships. Distance methods compress all of the individual differences between pairs of sequences into a single number.

2. Cladistic methods : trees are calculated by considering the *various possible pathways of evolution* and are based on **parsimony** or **likelihood** methods. The resulting tree is called a **cladogram**. Cladistic methods use each alignment position as evolutionary information to build a tree.

5.1. Phenetic methods based on <u>distances</u> :

1. Starting from an alignment, **pairwise distances** are calculated between DNA sequences as the sum of all base pair differences between two sequences (the most similar sequences are assumed to be closely related). This creates a **distance matrix**.

    - All base changes can be considered equally or a matrix of the possible replacements can be used.

    - Insertions and deletions are given a larger weight than replacements. Insertions or deletions of multiple bases at one position are given less weight than multiple independent insertions or deletions.

    - it is possible to correct for multiple substitutions at a single site.

2. From the obtained **distance matrix**, a phylogenetic tree is calculated with **clustering algorithms**. These cluster methods construct a tree by linking the least distant pair of taxa, followed by successively more distant taxa.

    - **UPGMA clustering** (<u>U</u>nweighted <u>P</u>air <u>G</u>roup <u>M</u>ethod using <u>A</u>rithmetic averages) : this is the simplest method

    - **Neighbor Joining** : this method tries to correct the UPGMA method for its assumption that the rate of evolution is the same in all taxa.

5.2. Cladistic methods based on Parsimony :

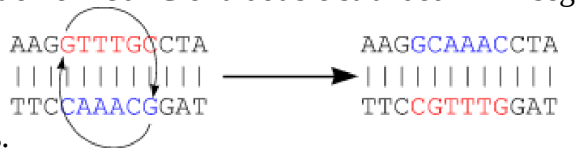
For **each position** in the alignment, **all possible trees** are evaluated and are given a score based on the number of evolutionary changes needed to produce the observed sequence changes. The **most parsimonious tree** is the one with the **fewest evolutionary changes** for all sequences to derive from a common ancestor. This is a more time-consuming method than the distance methods.

5.3. Cladistic methods based on Maximum Likelihood :

This method also uses **each position** in an alignment, evaluates all possible trees, and calculates the **likelihood for each tree** using an explicit **model of evolution** (Parsimony just looks for the fewest evolutionary changes). The likelihood's for each aligned position are then multiplied to provide a likelihood for each tree. The tree with the maximum likelihood is the most probable tree. This is the slowest method of all but seems to give the best result and the most information about the tree.

6. Theoretical problems with evolutionary changes between sequences

- Transitions : substitutions from A to G ; G to A ; C to T ; T to C.
- Transversions : substitutions from G to C ; C to G ; T to A ; A to T.
- Deletions : removal of one or more nucleotides.
- Insertion : addition of one or more nucleotides.
- Inversion : rotation of 180 °C of a double stranded DNA segment compromised of of 2 or more base pairs.



The next figure shows that there is a chance that many more mutations occur than visible at a certain time. Even the best evolutionary models can't solve this problem...

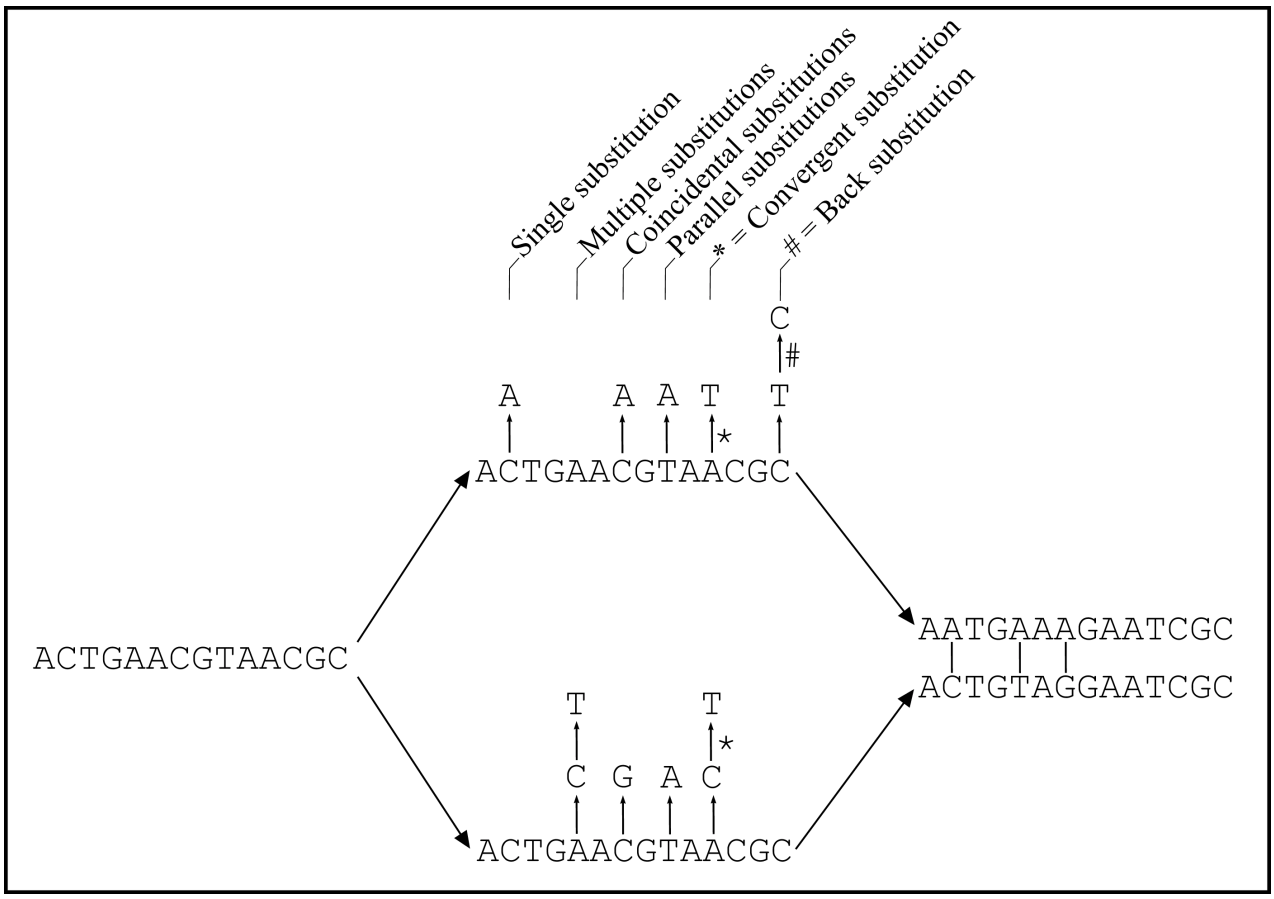


Figure 20: Two homologous DNA sequences which descended from an ancestral sequence and accumulated mutations since their divergence from each other. Note that although 12 mutations have accumulated, differences can be detected at only three nucleotide sites. (from Fundamentals of Molecular Evolution, Wen-Hsiung Li and Dan Graur, 1991)

Written by *Andy Vierstraete* for those who are interested in these techniques. This is free available and may be copied in huge amounts, as long as you don't charge anything for it and leave my name on this document...

Andy Vierstraete
Department of Biology
University of Ghent
K. L. Ledeganckstraat 35
B-9000 Gent
Belgium
http://allserv.rug.ac.be/~avierstr/index.html