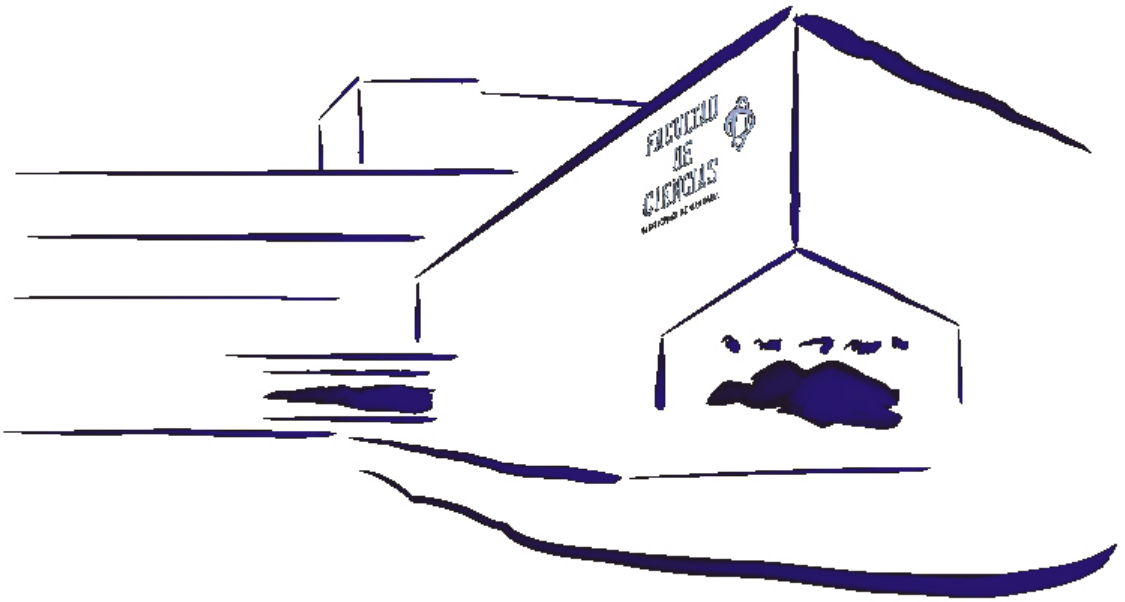



# MANUAL OF EXERCISES, LABORATORY AND BIOINFORMATICS FOR GENETICS II

BIOLOGY DEGREE

# MANUAL OF EXERCISES, LABORATORY AND BIOINFORMATICS FOR GENETICS II

This manual is part of the manuals "Problemas y Casos Prácticos de Genética" (ISBN: 978-84-15261-50-6) and "Manual de Prácticas de Genética" (ISBN: 978-84-15261-49-0) written by professors belonging to the Department of Genetics at the University of Granada, within the framework of a Teaching Innovation Project entitled "Nuevos recursos docentes para las prácticas del Departamento de Genética en el marco del EEES" (2010/2011) funded by The Vice Rectorate for Quality Guarantee - University of Granada.

During 2017/2018, this manual was revised within the framework of the Teaching Innovation Projects entitled "Actualización de material didáctico para la docencia práctica de la asignatura Genética II: de la secuencia a la función del Grado en Biología" and "Desarrollo de medios audiovisuales y virtualización de contenidos en asignaturas del área de Genética" funded by The Vice Rectorate for Quality Guarantee - University of Granada 2016/2018.

Next, the contents have been translated into English thanks to the Teaching Innovation Project entitled “Traducción del material docente de la asignatura Genética II: de la secuencia a la función (2º Grado en Biología) a lengua inglesa” funded equally by The Vice Rectorate for Quality Guarantee - University of Granada in 2022.

Authors of the English translation: Mohammed Bakkali, Francisco Javier Barrionuevo, Lara Mª Bossini Castillo, Miguel Burgos Poyatos, Míriam Cerván Martín, Roberto de la Herrán Moreno, Inmaculada López Flores, Tatiana López Pérez, Esther Viseras Alarcón, Federico Zurita Martínez.

Authors of the manual: Mohammed Bakkali, Francisco Javier Barrionuevo, Jiménez, Miguel Burgos Poyatos, Josefa Cabrero Hurtado, Roberto de la Herrán Moreno, Manuel Ángel Garrido Ramos, Michael Hackenberg, Rafael Jiménez Medina, María Dolores López León, Inmaculada López Flores, Ángel Martín Alganza, Rafael Navajas Pérez, Francisco Perfectti Álvarez, Francisca Robles Rodríguez, José Carmelo Ruiz Rejón, Esther Viseras Alarcón, Federico Zurita Martínez.

# INDEX

# EXERCISES

# MOLECULAR GENETICS

## 1. GUIDE TO SOLVING EXERCISES

### *Restriction maps*

A restriction map represents a linear sequence of the sites at which different restriction enzymes have targets on a particular DNA molecule. It consists of the arrangement of a series of restriction enzyme targets on a particular DNA molecule. The distances between these targets are plotted on the map and are measured in base pairs (or kilobases).

When a DNA molecule is cut with a restriction enzyme and the fragments generated are separated by electrophoresis on an agarose gel, the number of restriction sites and the distance between them can be determined from the number and position of the bands on the gel. Linear and circular DNA molecules must be distinguished for subsequent interpretation of the results:

**Linear DNA molecules**: Note that the number of fragments generated after a digestion is the number of targets present in the linear sequence for that enzyme plus one. The sum of the fragment sizes must coincide with the total size of the digested DNA. However, it should be noted that the number of fragments does not always coincide with the number of bands appearing on an agarose gel, as there may be fragments of the same size that migrate together.

**Circular DNA molecules**: The number of fragments generated after a digestion is the same as the number of targets present in the circular sequence for that enzyme. When an enzyme cuts only once, it reveals the size of the circular DNA. The sum of the size of the fragments should match the total size of the digested DNA. As described before, note that the number of fragments does not always coincide with the number of bands appearing on an agarose gel, as there may be fragments of equal size migrating together

In any case, the information obtained by electrophoresis does not reveal the order or location of the restriction targets. In order to make a map, a sample of the DNA to be mapped has to be cut with one restriction enzyme, a second sample of the same DNA with a different enzyme and a third sample of the same DNA with both enzymes simultaneously (double digestion). This third digestion gives us the key to determine the order of the targets for both restriction enzymes.

### *Molecular markers*

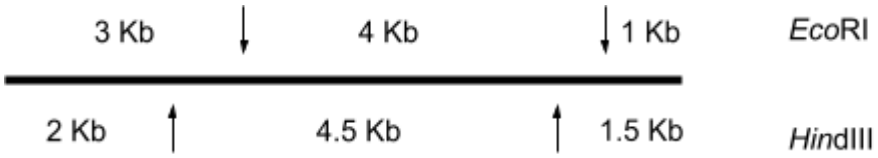
It is important to assign genotypes to individuals in the pedigree in order to elucidate the match between the alleles they carry and banding patterns.

The distance between restriction sites or targets must be taken into account, which will give the size of the observable bands, but special attention must also be paid to the region with which the probe hybridizes, as those fragments with which it does not hybridize cannot be detected when results are analyzed.

In the case of microsatellites, the different amplified sizes for a locus correspond to different alleles. Microsatellites present simple Mendelian inheritance and are codominant. For a microsatellite locus, one of the alleles present in an individual's genotype (amplified size) comes from the father and one comes from the mother.

## 2. SOLVED EXERCISES

**Exercise 1.** A cloned gene shows the following restriction map for the EcoRI and HindIII enzymes (↓ cut site):



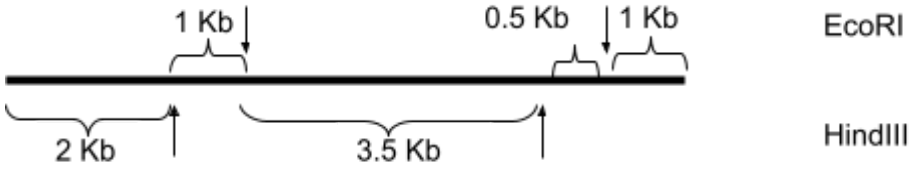

a) Draw the patterns of DNA fragments expected with each enzyme when separating the fragments by agarose gel electrophoresis. Do the same for the case of double digestion.

b) Draw the expected pattern for a mutant copy of the gene that has lost the first of the EcoRI cuts.

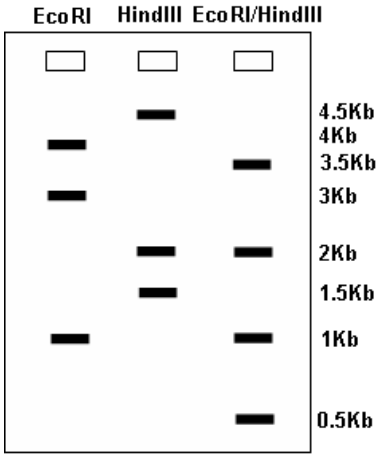
c) Draw the expected pattern for a mutant copy of the gene in which a new target for HindIII has appeared in the center of the 2Kb fragment.

*Answer*

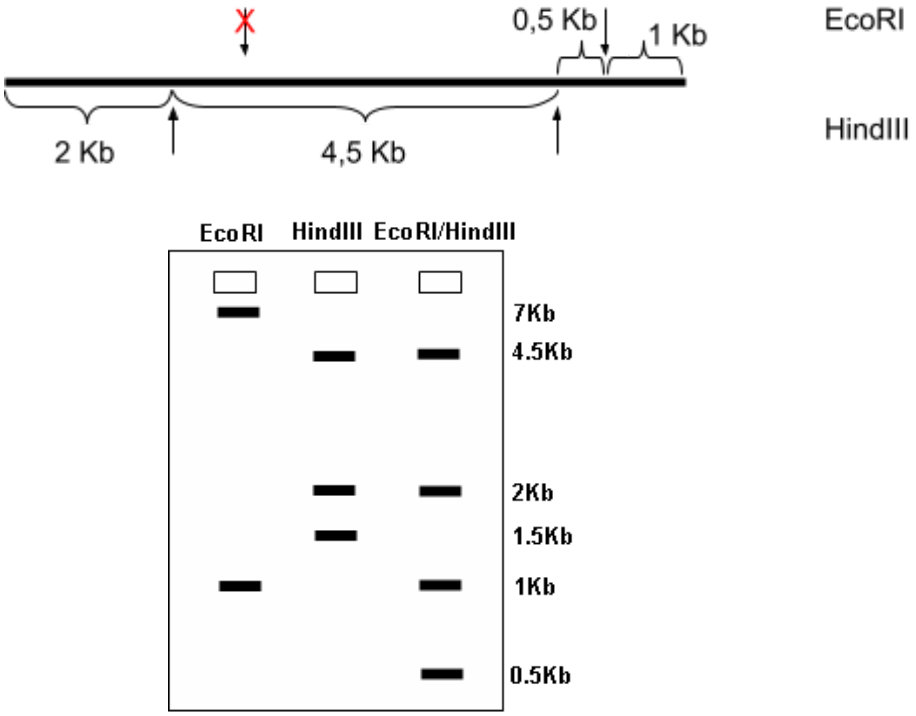
a) For EcoRI the gene has two targets, so it will be cut into three fragments of sizes 4Kb+3Kb+1Kb. For HindIII it also has two targets, but at different positions, so it will generate three fragments but of sizes 4.5Kb+2Kb+1.5Kb. When we use the two enzymes to digest the gene, we will obtain 5 different fragments (there are 4 cut-off points, generating fragments from target to target of both enzymes), although two of them have the same size (1Kb), so we will observe them as a single band in the agarose gel. The sizes will therefore be 3.5Kb+2Kb+1(x2)Kb+0.5Kb (see figure).



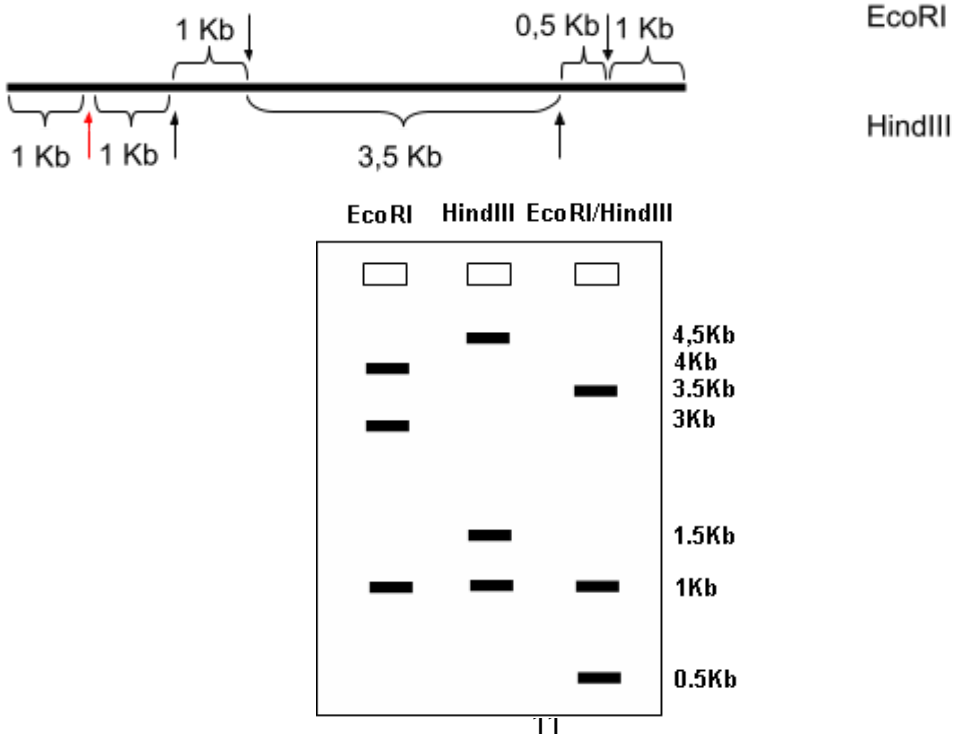

Thus, on an agarose gel, we observe the following banding patterns:



b) If the mutant copy of the gene loses a target for EcoRI, on cutting with this enzyme, we will obtain only two fragments, one of them being the sum of the two between which the target lost for EcoRI was found, 7Kb+1Kb. For the cut with HindIII the banding pattern would not be affected, but again for the double digestion, as there is one less cut, 4.5Kb+2Kb+1Kb+0.5Kb (see figure).




c) In this case, when we cut the gene with HindIII, as we have one more target (three cut points) we would obtain one more fragment. However, in the gel, 4 bands would not appear, since two fragments of equal size (1Kb) have been generated, so they will run in the same way. The fragments for HindIII would be 4.5Kb+1.5Kb+1Kb(x2). For EcoRI the pattern is not affected and for the double digestion the fragments generated would be 3.5Kb+1Kb(x4)+0.5Kb (see figure).



**Exercise 2.** A circular bacterial plasmid containing an ampicillin resistance gene has been cut with PstI. After electrophoresis, a band of 20 Kb is observed. What would you deduce from the following results?

a) With EcoRI, the plasmid is cut into two fragments: one of 11.5Kb and the other of 8.5Kb.

b) PstI+EcoRI digestion generates three fragments: 6Kb, 5.5Kb and 8.5Kb.

c) Plasmid DNA cut with PstI has been mixed and ligated with DNA fragments cut with PstI. All recombinant clones are ampicillin resistant.

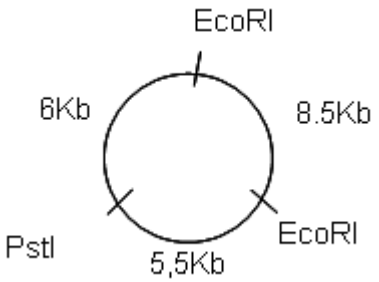
d) After cutting one of the recombinant clones with PstI, two fragments are obtained: 20 Kb and 6 Kb.

e) The previous clone is cut with EcoRI and 10 Kb, 8.5 Kb and 7.5 Kb are obtained.

*Answer*

a) Adding the fragments 11.5 Kb + 8.5 Kb gives a value of 20 Kb. This value coincides with the fragment generated with PstI, which means that the plasmid has a size of 20 Kb and that, therefore, PstI cuts it only once while EcoRI has two targets inside the circular molecule.

b) With the double digestion we can now obtain a restriction map of this circular molecule. We can deduce that the 11.5 Kb fragment generated by EcoRI is cut into two smaller fragments of 6Kb and 5.5Kb by the PstI enzyme:

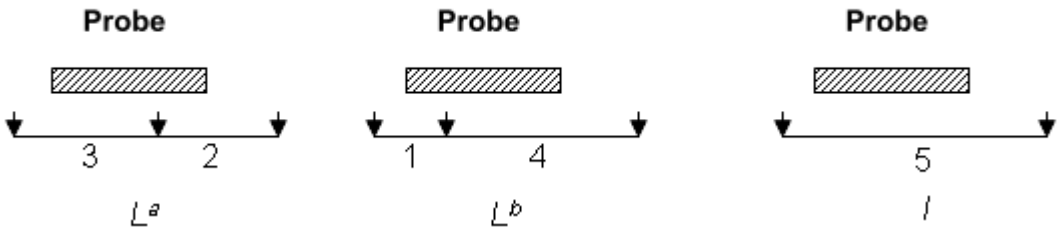



c) When cut with PstI, the plasmid remains in linear form with cohesive ends for that enzyme. When DNA fragments also cut with PstI are brought into contact with a ligase enzyme, the fragments, which have complementary ends, bind to the plasmid and the molecule recirculates with an insert inside it, yielding a recombinant plasmid. If the target for PstI were inside the ampicillin gene, the insert would "break" this gene, rendering it inactive and the bacteria sensitive to ampicillin. Therefore, we can deduce that the target for PstI is not inside the ampicillin resistance gene.

d) By cutting again with the PstI enzyme, what we are doing is separating the plasmid from the insert again, so we obtain a 20 Kb fragment corresponding to the plasmid and another 6 Kb fragment, which would be the size of the cloned fragment.

e) The appearance of a new fragment when we cut the recombinant plasmid with EcoRI, which did not appear in the initial bacterial plasmid, means that there is a new target for this enzyme. The difference between the initial bacterial plasmid and the recombinant plasmid is the presence of the 6 Kb insert. Therefore, we deduce that the cloned fragment has a target for EcoRI and that it is located between the two EcoRI targets separated by 11.5 Kb.

**Exercise 3.** An autosomal gene is discovered with three alleles (known as: $L^a$, $L^b$ and *I*) that differ in a target for the restriction enzyme PstI (↓ cleavage site):




Design an experiment to differentiate the genotypes of the different individuals that could exist in a population if we use the homologous DNA fragment indicated in the scheme as a probe.

*Answer*

The difference between the different alleles of the gene corresponds to differences in the nucleotide sequence. In this case, these nucleotide differences affect target sequences for the PstI enzyme. We will use this information to perform an experiment to detect an RFLP (restriction fragment length polymorphism) marker.

To do this, we must follow the following steps:

a) Digest the entire genomic DNA with the PstI enzyme, since we cannot initially isolate our gene from the rest of the genome.

b) In order to separate the fragments generated, according to their size, we must now perform an agarose gel electrophoresis.

c) The DNA fragments, as arranged in the agarose gel, must be transferred to a nylon membrane using the Southern-blot technique.

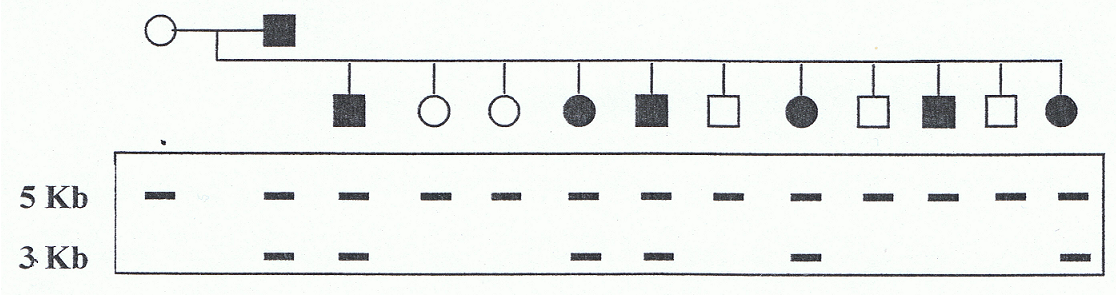
d) By means of a Southern-blot hybridization, using the probe indicated in the scheme, we can specifically locate the region corresponding to the gene studied, as it is complementary to this region.

e) The development of the hybridization will reveal the genomic DNA fragments with which the probe has hybridized.

Thus, we have different possible genotypes, which will coincide with banding patterns:

$L^aL^a$: 3Kb and 2Kb bands.
$L^aL^b$: bands of 3Kb and 2Kb for the first allele and 1Kb and 4Kb for the second allele
$L^aI$: bands of 3Kb and 2Kb for the first allele and 5Kb for the second allele
$L^bL^b$: 1Kb and 4Kb bands
$L^bI$: bands of 1Kb and 4Kb for the first allele and 5Kb for the second allele
*II*: 5Kb band

**Exercise 4.** The following pedigree represents a family with some of its members affected by an autosomal dominant disease. DNA from all individuals was digested with the PstI enzyme and subjected to agarose gel electrophoresis. This DNA was analyzed by Southern hybridization with a probe corresponding to a fragment of human DNA cloned into a bacterial plasmid. The results of the hybridization development are shown together with the pedigree.

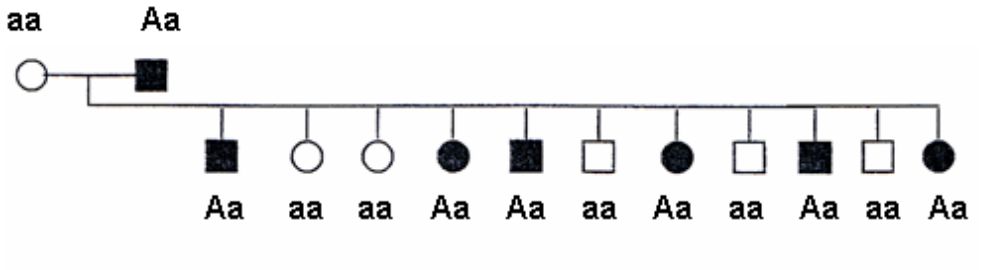



a) Explain the protocol followed and the results obtained in the different individuals.
b) Can we use the probe for diagnostic purposes for this disease?

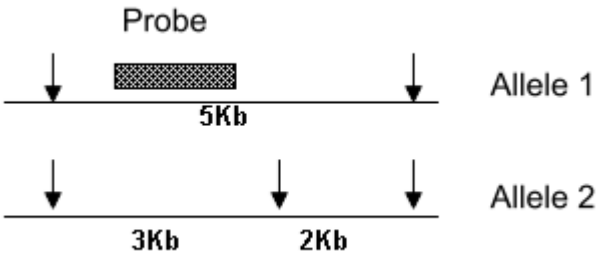
Answer

a) The molecular marker used in this analysis corresponds to an RFLP. Genomic DNA has been digested with the enzyme PstI and the fragments generated have been separated by agarose gel electrophoresis. These fragments (as migrated in the gel) are then transferred to a nylon membrane, to which they are fixed. A hybridization (Southern hybridization) with a labeled DNA fragment (probe) is performed on this membrane. When this hybridization is revealed, bands of different molecular weights appear, indicating sizes of genomic DNA fragments that are homologous to the probe.

The first step is to assign the genotypes to the individuals in the pedigree, taking into account that the disease is autosomal dominant:

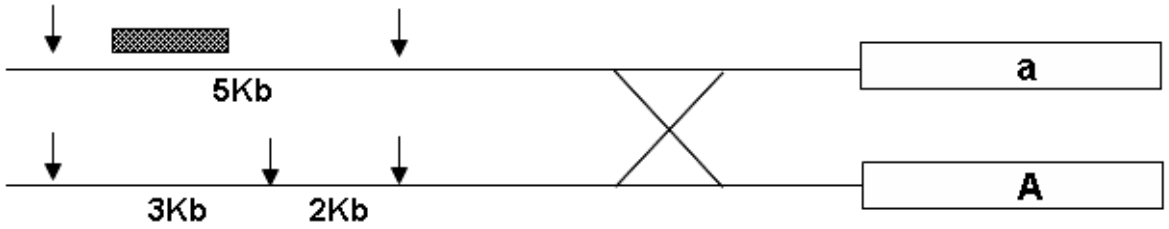



When comparing the results of the molecular marker with the phenotypes (affected/not affected by the disease) we can see how there is coincidence between the number and sizes of the RFLP bands and the development or not of the disease. Thus, with the exception of one individual (II-9), all affected individuals present two bands of 5Kb and 3Kb and those unaffected present a single band of 5Kb. Bearing in mind that, in diploid species, there are pairs of homologous regions (homologous chromosomes), we must "identify" two alleles. Thus, the difference in sequence between these alleles could be detected if it affected a target for the PstI enzyme, as shown in the scheme:




If the probe hybridizes in the indicated region, and taking into account that the 3Kb band is exclusive to those affected, we can deduce that allele 1 in the above diagram corresponds to allele *a* of the pedigree, while allele 2 corresponds to allele *A*, which is the cause of the disease. Heterozygous individuals (Aa), then, present two bands, 5Kb of allele *a* and 3Kb of allele A (since the 2Kb fragment of allele *A* is not detected by the probe). Healthy homozygous individuals (aa) have a single 5Kb band. The hypothetical homozygous individuals (AA) would have a single 3Kb band.

b) When establishing the relationship between the genotypes and the banding pattern, we observe that individual II-9 does not present this correspondence. This could be explained by the fact that the observable differences in the banding pattern are not due to changes in the sequence of the disease-causing gene itself, but to regions close to it. In other words, the RFLP that we detected does not correspond to differences in the sequence of the alleles of the gene, but is found in regions linked to it, as shown in the following diagram:




This diagram illustrates the case of the affected parent (individual I-2; genotype Aa). If, during the formation of the gametes of this individual, there were a crossover between the RFLP and the gene causing the disease (as indicated in the figure), a gamete with genotype A but with an RFLP marker of 5Kb would be generated. This is what happens to individual II-9, which has an *a* allele (5Kb) inherited from the mother and an *A* allele (5Kb recombinant) inherited from the father.

Therefore, the probe could be used as a diagnostic, but we must take into account that there is a percentage of error due to the possibility of recombination between the RFLP marker and the disease-causing gene.

**Exercise 5.** In an analysis with 4 microsatellite markers, the following results were obtained for 5 individuals (numbers indicate amplified fragment sizes in bp):

| | Individual 1 | | Individual 2 | | Individual 3 | | Individual 4 | | Individual 5 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Allele 1 | Allele 2 | Allele 1 | Allele 2 | Allele 1 | Allele 2 | Allele 1 | Allele 2 | Allele 1 | Allele 2 |
| **Locus 1** | 130 | 134 | 134 | 134 | 136 | 138 | 128 | 134 | 128 | 136 |
| **Locus 2** | 250 | 256 | 256 | 260 | 258 | 260 | 252 | 260 | 250 | 258 |
| **Locus 3** | 140 | 140 | 140 | 144 | 146 | 148 | 138 | 144 | 140 | 150 |
| **Locus 4** | 187 | 193 | 185 | 187 | 183 | 189 | 185 | 191 | 181 | 189 |

a) What is the difference between the different alleles of the same microsatellite locus?

b) How many alleles do the different microsatellite loci analyzed in this study have?

c) Is it possible to know the number of repeats for each of them, and the reason for repetition?

d) If individual 1 is the mother of individual 2, which of the other three individuals can be ruled out as possible parents?

*Answer*

a) Among the different alleles of the same microsatellite, the existing differences correspond to a variable number of repeats of a motif (usually dinucleotide, trinucleotide or tetranucleotide).

b) With this sample we cannot know the total number of alleles in the population, as there may be more alleles that are not represented in these individuals. In the individuals analyzed we have 5 alleles at locus 1, 5 alleles at locus 2, 6 alleles at locus 3 and 7 alleles at locus 4.

c) When amplifying the microsatellite repeats, the flanking regions are used to design the primers. The distance in base pairs between the repeated motif and the regions where the primers are designed are variable for each microsatellite, so the amplified fragment includes the repeated motif and part of the flanking regions whose size, in this case, we do not know. Therefore, we cannot know the number of repeats in each allele. Nor do we know the repeat motif, as we have no sequence information. What we do know is that in the four microsatellites, this motif corresponds to two nucleotides, as the alleles have variations of two base pairs between them.

d) We can rule out individuals 3 and 5, as individual 2 must for each microsatellite have one allele from the mother and one from the father.

## 3. EXERCISES TO BE SOLVED

**Exercise 1.** A DNA fragment was cut with *Pst*I and *Hind*III separately. Subsequently, a mixture of both enzymes was used to obtain the fragments shown below:

*Pst*I: 3Kb and 4Kb
*Hind*III: 2Kb and 5Kb
*Pst*I+*Hind*III: 1Kb, 2Kb and 4Kb

Draw the restriction map of this DNA segment.

**Exercise 2.** A cloned DNA fragment was digested with the restriction enzymes *Hind*III and *Sma*I, and with a mixture of both enzymes. The following was obtained:

*Hind*III: 2,5Kb and 5Kb
*Sma*I: 2Kb and 5.5Kb
*Hind*III+SmaI: 2Kb, 2,5Kb and 3Kb.

a) Draw the restriction map

b) When the mixture of fragments produced by the action of the two enzymes at the same time was also cut with the *Eco*RI enzyme, the 3Kb fragment disappeared and a 1.5Kb band was observed when results were analyzed by agarose gel electrophoresis. Indicate on your previous map the *Eco*RI cut point.

**Exercise 3**. A linear DNA fragment 11Kb in length was cut separately with the restriction enzymes *Eco*RI and *Hae*III and with a mixture of both enzymes. The following fragments were obtained: *Eco*RI: 6Kb, 3Kb and 2Kb; *Hae*III: 7Kb y 4Kb; *Eco*RI+*Hae*III: 5Kb,3Kb, 2Kb y 1Kb.

Draw the restriction map of this DNA segment.

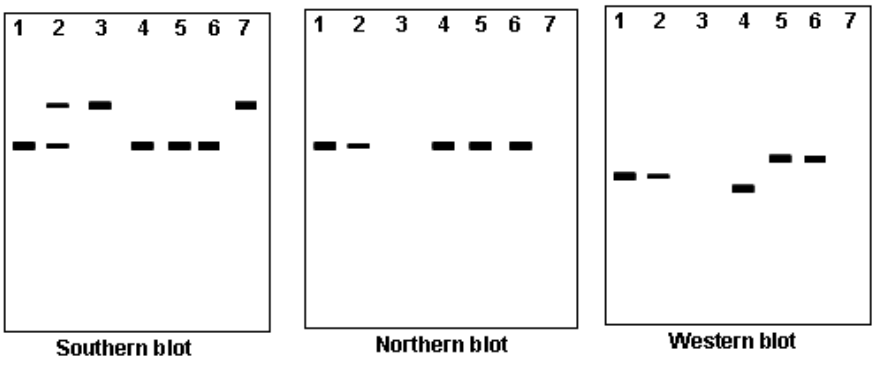
**Exercise 4**. The pAl21 plasmid was cut with different restriction enzymes and the following bands were observed after agarose gel electrophoresis analyses: *Bam*HI (3.7 Kb, 3.5Kb), *Pvu*II (7.2Kb), *Hin*dIII (7.2Kb), *Bam*HI+*Pvu*II (3.5Kb, 2.4Kb, 1.3Kb), *Pvu*II+*Hin*dIII (3.6Kb).

Draw the restriction map of the plasmid.

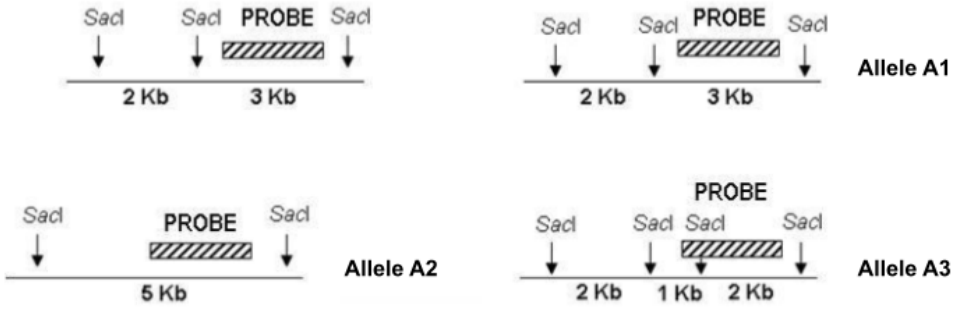
**Exercise 5.** A protein is encoded by a gene that has no introns. The *Sac*I restriction fragment containing the complete gene can be identified by Southern-blot hybridization with the cDNA of the radiolabelled gene. To determine the cause of an unknown disease, blood was obtained from patients and healthy controls. Their DNA was extracted, cut with the enzyme *Sac*I, transferred to a nylon membrane and hybridized with the labeled cDNA as a probe. Similarly, RNA was extracted, subjected to electrophoresis, transferred to a membrane (Northern-blot) and hybridized with the cDNA. In addition, Western-blotting was performed and the protein encoded by the gene was tested using a specific antibody.

The results are shown below (persons 1 and 2 are healthy controls and persons 3, 4, 5, 6 and 7 have the disease)

What could be the cause of the disease in each of the diseased individuals?

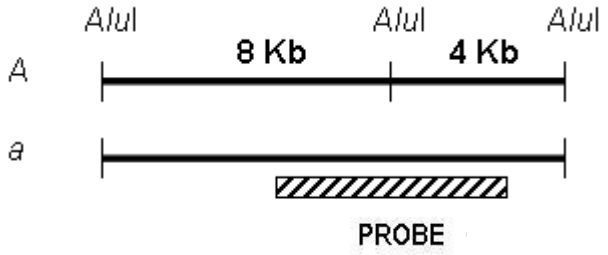



**Exercise 6.** Following a search for molecular markers, a probe was designed to the genomic DNA of the species under study. The figure shows the probe target region, which allows differentiation of three different alleles (A1, A2 and A3). DNA extracted from different individuals was cut with *Sac*I, electrophoresed and then transferred to a nylon membrane. This membrane was hybridized with the probe (radioactively labeled) and an autoradiography was performed.




Make a schematic drawing with the expected outcome for:
A homozygote for A1, a heterozygote A1A2, a homozygote A2 and a heterozygote A1A3.

**Exercise 7.** An autosomal gene has two alleles, A and a, which differ for the restriction enzyme *Alu*I as shown in the figure. Design an experiment to differentiate the genotypes in a population. Draw the possible results.

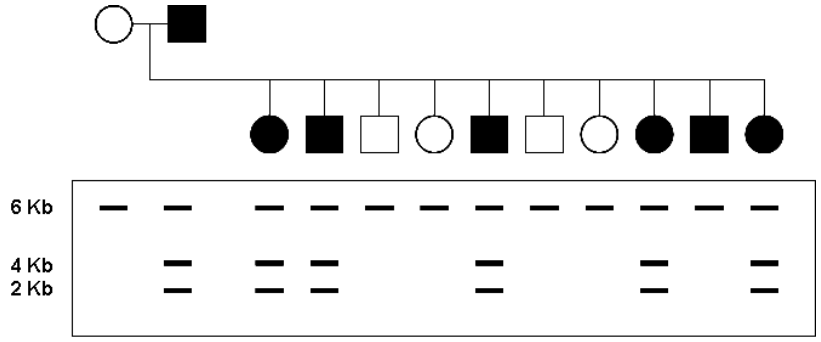



**Exercise 8.** Different DNA probes were tested by hybridizing with the genomic DNA of individuals from a large family in which some members are affected by an autosomal dominant disease of late onset (at approximately 40 years of age). On the Southern-blot

obtained with *Taq*I, one of the probes detects a restriction fragment length polymorphism (RFLP). The RFLP patterns of each individual in the pedigree are shown in the figure.

a) Explain the obtained results.

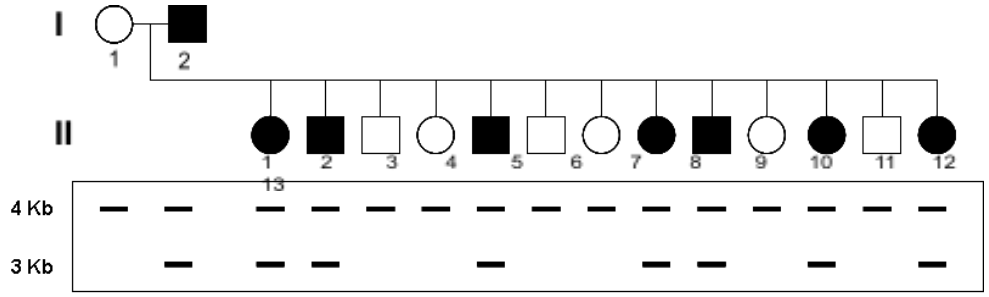
b) Are the RFLP and the disease-causing gene linked?



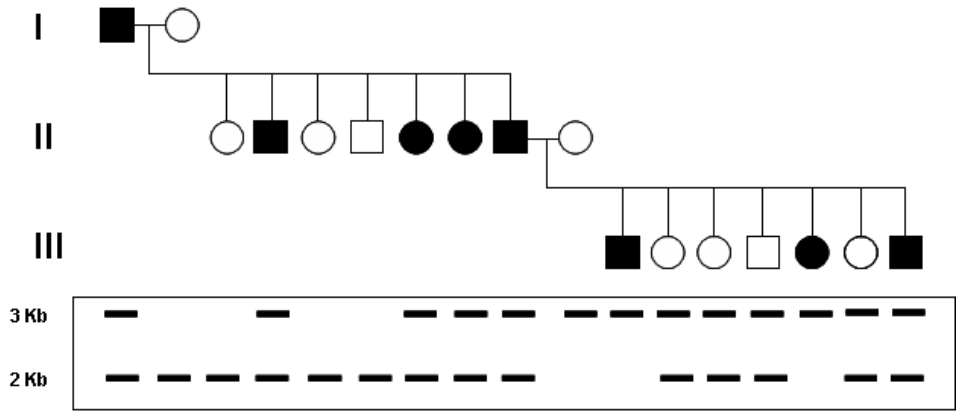

**Exercise 9.** Different DNA probes were tested by hybridizing with genomic DNA from individuals of a large family in which some members are affected by a mild autosomal dominant disease. On the Southern-blot obtained with *Eco*RI, one of the probes detected a restriction fragment length polymorphism (RFLP). The RFLP patterns of each individual in the pedigree are shown in the figure.

a) Explain the results.

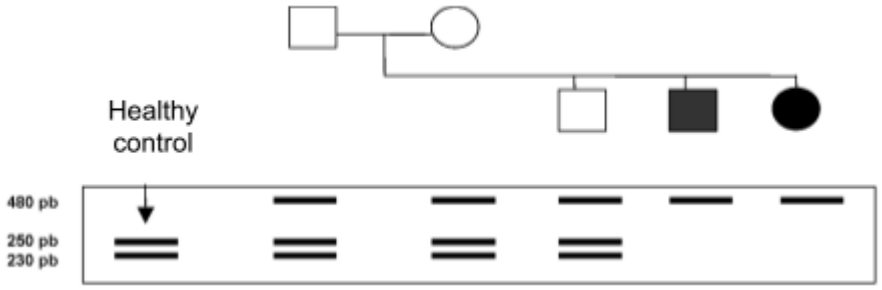
b) Are the RFLP and the disease-causing gene linked?




**Exercise 10.** Genomic DNA was extracted from members of a family affected by an autosomal dominant disease. The DNA was then digested with *Pvu*II and the fragments were separated by agarose gel electrophoresis. Southern-blot hybridization with a probe that detects an RFLP yields the results shown in the figure. Are the RFLP and the disease-causing gene linked?




**Exercise 11.** A disease is associated with the absence of activity of a particular enzyme. In each member of the family shown below, exon 2 of the gene encoding that enzyme was amplified by PCR and amplicons were then digested with the *Eco*RI restriction enzyme, yielding the following results:




a) Can this marker be used as a diagnostic method?

b) What type of disease is described in the pedigree?

c) Make a schematic drawing for explaining the results obtained at molecular level.

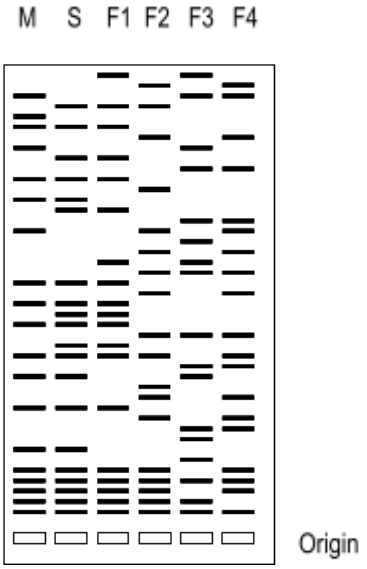
**Exercise 12.** A probe can detect an RFLP with two alternative alleles of 1.7Kb and 3.8 Kb from mouse DNA digested with *Hind*III. A mouse, heterozygous for a dominant allele determining a curved tail and with alleles for the 1.7Kb and 3.8Kb RFLP, was crossed with a wild-type mouse showing only the 3.8Kb fragment. Half of the offspring had curved tails. When analyzing these mice with curved tails for RFLP, we found that 20% of them are homozygous for the 3.8Kb allele and 80% are heterozygous for the 3.8Kb allele.

a) Are the locus that determines the curved tail and the RFLP linked?

b) If so, what is the distance between them?

c) Make a schematic drawing for explaining the results.

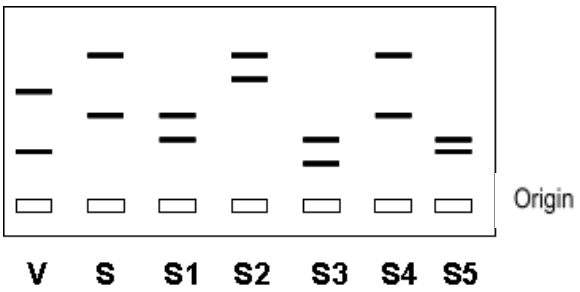
**Exercise 13.** Four men dispute the paternity of a child. The forensic experts decide to use the genetic fingerprinting method to solve the case, by analyzing the DNA of the mother (M), the son (S) and the four possible fathers (F1 to F4). The results obtained are shown in the figure:




a) Who is most likely to be the father?

b) Attribute as many bands as possible to the father and the mother.

**Exercise 14.** DNA is extracted from the blood of a rape victim (V), from semen taken from her body (S) and from samples taken from 5 suspects (S1, S2, S3, S4 and S5). A microsatellite study is carried out using a locus-specific primer pair. After amplification with the primer pair, the following results are obtained:

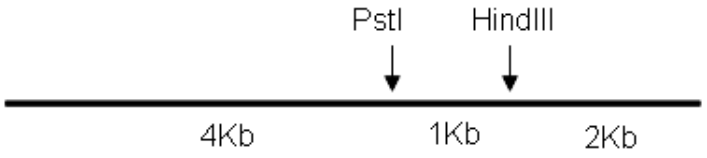



a) Explain the amplification patterns obtained.

b) Are there any suspects who appear to be guilty?

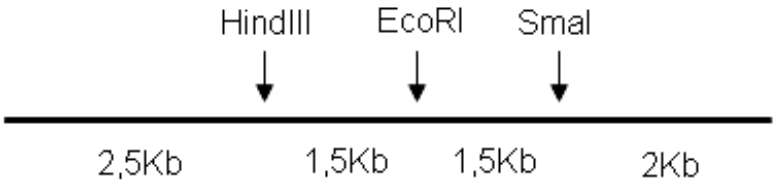
# 4. SOLUTIONS TO THE EXERCISES

**Exercise 1.**



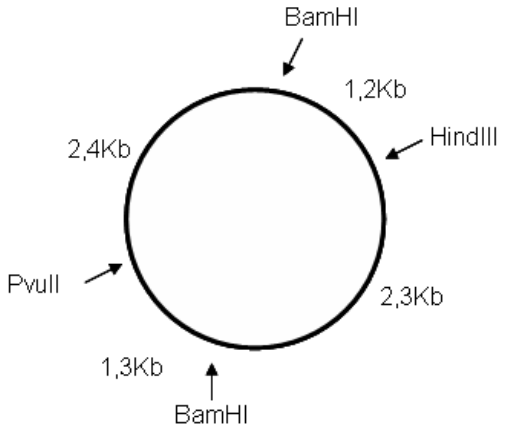

**Exercise 2**
**a**)


**b)**




**Exercise 3.**

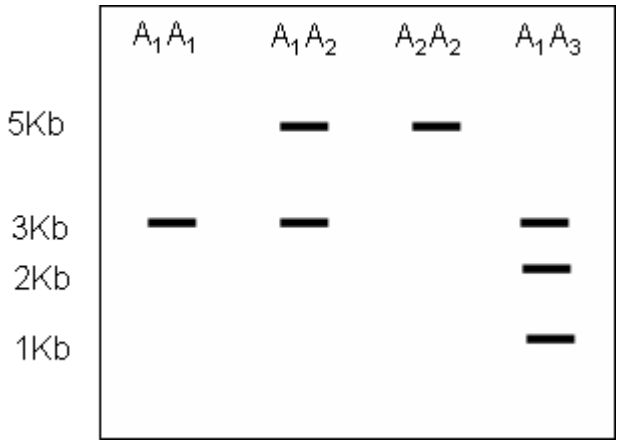

**Exercise 4.**



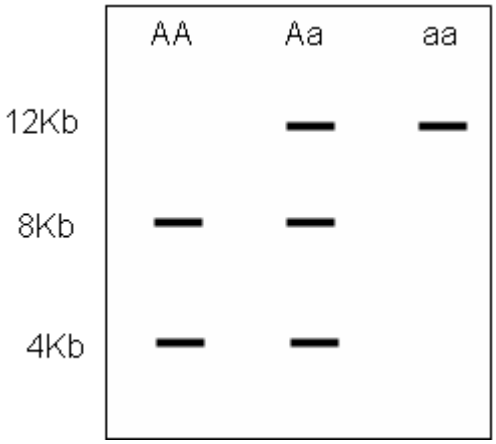
**Exercise 5.**

The gene is not transcribed (individuals 3 and 7) or the proteins produced are defective (individuals 4, 5 and 6). Individual 1 is homozygous for the normal allele and individual 2 is heterozygous for the allele that is not transcribed and produces half mRNA and protein than individual 1. Individuals 3 and 7 are homozygous for the latter allele and do not produce protein. Individuals 4, 5 and 6 are homozygous for the allele that transcribes and produces mRNA of the same length as the normal allele but have undergone some change that alters the protein sequence.

**Exercise 6.**

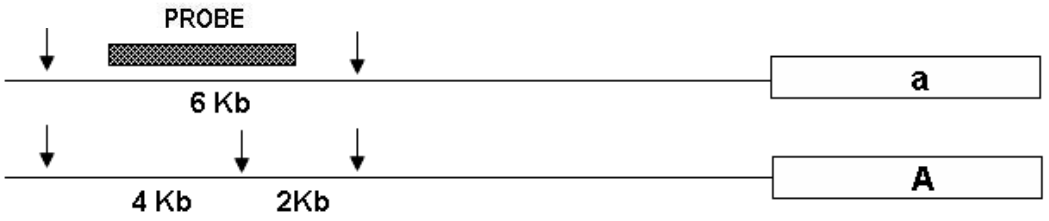



**Exercise 7.** The genomic DNA of each individual is digested with the enzyme *Alu*I. The cut DNA is then subjected to electrophoresis and the fragments are transferred to a nylon membrane using the Southern-blot technique. Hybridization is performed using the probe to detect homologous regions. The possible results are:



**Exercise 8.**

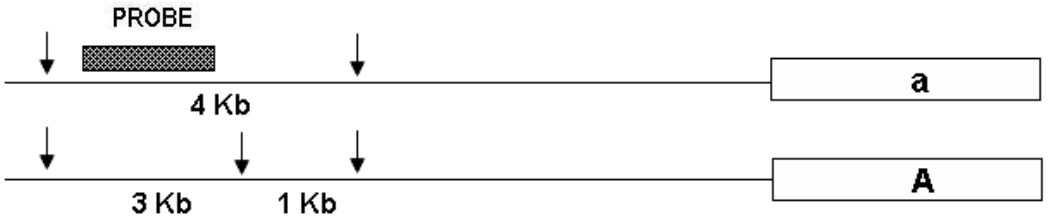
The RFLP and the disease-causing gene are linked, and a possible interpretation of the results is shown in the following diagram:




Affected, Aa (6Kb/4Kb/2Kb); Healthy, aa (6Kb). The genotype of the II-9 individual is the result of a crossover between the RFLP-containing region and the disease-causing gene in one of the parent's meioses.

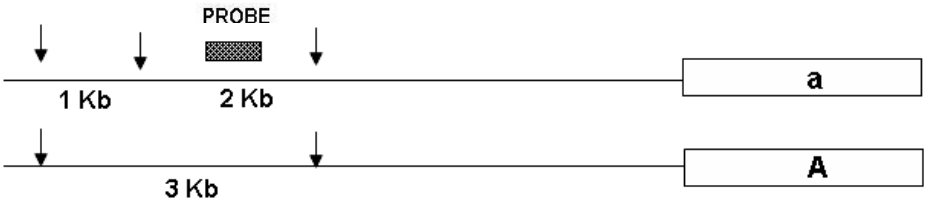
**Exercise 9.**

The RFLP and the disease-causing gene are linked, and a possible interpretation of the results is shown in the following diagram:




**Exercise 10.**

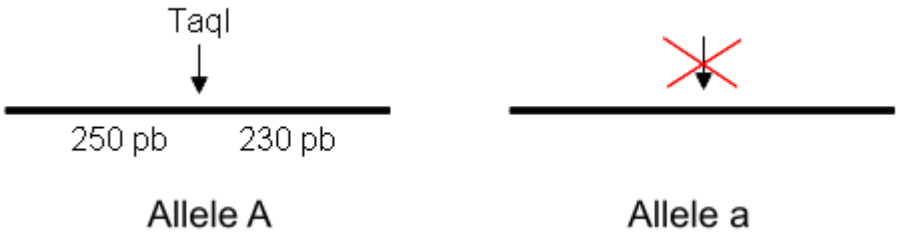
The RFLP and the disease-causing gene are linked, and a possible interpretation of the results is shown in the following diagram:




Affected, Aa (3Kb/2Kb); Healthy aa (2Kb). This would be for generations I and II. Individual II-8 is an individual from another family, in which the a allele is associated with the 3Kb band. This causes the banding pattern to change in generation III. Thus, the affected Aa (A from the father and a from the mother) are homozygous (3Kb/3Kb) and the healthy aa (a from the father and a from the mother) are heterozygous (3Kb/2Kb). The III-7 individual is the result of a cross between the RFLP and the disease-causing gene in the heterozygous father (II-7).

**Exercise 11.**

Yes, the marker can be used as a diagnostic method for this autosomal recessive disease.

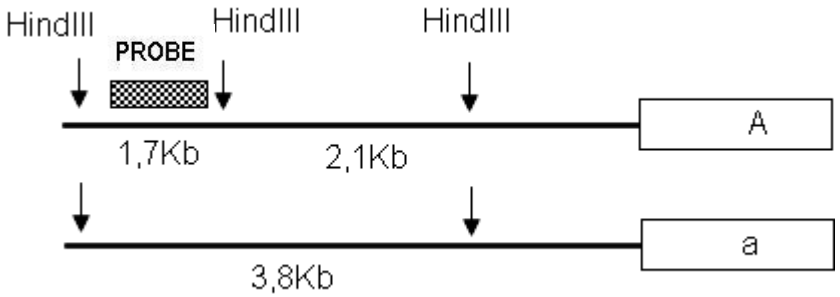



**Exercise 12.**

a) Yes**,** the locus that determines the curved tail and the RFLP are linked.

b) 20 m.u.

c)




**Exercise 13.**

**a)** Individual F1.

b) Half must be present in the mother and half must be present in the father.

**Exercise 14.**

a) The amplification patterns are due to the difference in the number of tandem repetitions.

b) The S4.

# LABORATORY AND BIOINFORMATICS

# I.- APPLICATION OF THE PCR TECHNIQUE TO GENETIC DIAGNOSIS: DETECTION OF PARASITES THAT INFECT MOLLUSCS

## 1.1. AIM

This practice is intended to verify the effectiveness of PCR in the genetic diagnosis of diseases and parasitic infections in bivalve molluscs. The specificity of the primers is used to amplify a specific region of the parasite genome when it is present in a sample.

## 1.2. THEORETICAL BASIS

PCR is a technique that allows the exponential amplification (multiplication) of a specific fragment of DNA in vitro. To carry it out requires a DNA template, a pair of primers that pair to the end of a 3' chain of DNA to be amplified (where the elongation takes place, always from 5' to 3'), a sufficient quantity of deoxyribonucleotide triphosphates (dATP, dCTP, dGTP and dTTP), a DNA polymerase (named Taq polymerase), its buffer, and the conditions for an efficient reaction.

The reaction is cyclical after the initial step of **denaturing** the template DNA (2 to 5 minutes at 96ºC), generally consisting of about 25 to 35 cycles. Each of the cycles is made up of a denaturation stage (about 30-60 seconds at 94ºC), an **annealing** of the primers to the DNA template (about 30-60 seconds), and an **extension** or polymerization step of the DNA to be amplified (whose duration depends on the size of the DNA). The duration of the extension step can be approximated as one minute per kilobase of DNA to be amplified. Following the denaturation, primers annealing to the DNA template, and extension cycles, the reaction is terminated with a final extension step that typically lasts five minutes.

Although a single DNA template molecule may suffice for PCR detection, the DNA must be of good quality (not degraded) and free of enzymatic activity inhibitors. The DNA region to be amplified (the "amplicon") should also be no larger than 3 or 4 kilobases, and preferably free of secondary structures that may block the progression of DNA polymerase during synthesis. The primers consist of the reverse and complementary strands of the starting sequence of the two strands of DNA to be amplified. These should align exclusively with the complementary region in the DNA fragment to be amplified, so that they cannot bind with any other DNA sequence. The primers are usually between 15 and 35 nucleotides in size, where more nucleotides allow for greater specificity. They should also include roughly equal shares of CGs (Cytosine and Guanine bases) and ATs (Adenine and Thymine bases). Above all, they should be free of secondary structures and either internal or external complementarities that could cause them to fold over or align with each other, forming dimers.

The DNA is first denatured by incubation at 94ºC. The temperature is then reduced to a level known as the "melting temperature" (Tm), at which the primers will bind to their reverse and complementary sequences. The Tm should be roughly the same for both primers, with a maximum difference of 5°C that depends on their composition and size. Several algorithms can be used to precisely determine its level.
A simple formula for estimating Tm is 4 x GC + 2 x AT, where GC is the number of GCs and AT is the number of ATs. Other simple formulas can be used to determine the potential for both internal and external secondary structure formation or complementarity between primers.

In principle, any DNA polymerase can be used to synthesize DNA *in vitro*. But because PCR requires high temperatures for both DNA template denaturation (94°C) and primer annealing (40-65°C or more depending on the primers), it calls for thermostable DNA polymerases such as those found in microorganisms that live in hot places. The DNA polymerase most commonly used for PCR, Taq polymerase, comes from the thermophilic bacterium *Thermus aquaticus,* and has an optimal DNA polymerization temperature of 72°C, or roughly the habitat temperature for this microorganism. The optimal DNA synthesis temperatures for other DNA polymerases will likewise tend to correspond to the habitat temperatures of their source microorganisms.

Like any biochemical reaction, PCR needs a buffer solution that consists of a mixture of salts and reagents, among which magnesium chloride stands out. Furthermore, cyclic repetitions at different temperatures throughout the PCR reaction are achieved by use of **thermocyclers**. These devices achieve precise temperatures, maintain them for a certain time and switch to another temperature evenly and quickly.

Thus, PCR consists of a series of steps: first, denaturation, which opens the double strand of the DNA template, meaning both strands will be separated at the end of this step. Secondly, annealing, which allows the anchoring of the primers to their corresponding reverse and complementary sequences; and last but not least, the extension of these primers which allows the synthesis of a single strand of DNA, starting from the last 3' nucleotide of the primer attached to its corresponding DNA template strand. During this step, the exponential increase in the number of fragments is verified.

As can be seen, each cycle results in the doubling of the number of molecules corresponding to the DNA to be amplified. So, after the second cycle there will be four times the number, after the third cycle there will be eight times the number of molecules, and so on. Upon finishing, there will be a theoretical quantity of 2n x C amplified DNA molecules where “n” is number of PCR cycles and C is number of initial template molecules. It is recommended not to exceed 35 PCR cycles since, on the one hand, DNA polymerase has a synthesis error rate (close to one per million nucleotides incorporated) and, on the other hand, the differential depletion of products in the reaction may result in more errors (e.g. if dATPs are depleted, Taq may insert a dTTP at a position corresponding to a dATP).

Once the PCR reaction is complete, the reaction products are visualized using the agarose gel electrophoresis technique. By loading the PCR products on the agarose gel and subjecting the latter to a direct electric field, the negative electric charge of the DNA is used to make it migrate differentially from the negative pole to the positive pole of the direct electric field (positive and negative poles). The porosity of the agarose (which depends on the percentage of agarose) gel will cause the DNA molecules to separate based on their size as they migrate from the negative pole to the positive pole, with the shorter fragments migrating faster through (and therefore going further down) the gel. For reference, DNA molecules are also separated from a mixture of fragments of known sizes and relative amounts (molecular weight markers), referred to as a “ladder”. Simultaneous but separate division in different wells of the PCR products and the molecular weight marker on the same gel allow us to determine the molecular sizes of the PCR products that should match those expected.

The presence of *Perkinsus spp* is known to exist in the majority of warm waters of the world and has historically been associated with mass mortalities of bivalve molluscs. The presence of *Perkinsus olseni* in clams off the European coast has been known since 1987. This parasite has been detected, for example, in the grooved carpet shell (*Ruditapes decussatus*); in the Japanese Littleneck (*R. philippinarum*), in the pullet carpet shell (*Venerupis pullastra*), in the golden carpet shell (*V. aurea*) and in the carpet banded shell (*V. rhomboides*). *Perkinsus*

*olseni* can currently be considered the main pathological problem for the development of clam culture on the European coast. Until recently, techniques were used for their diagnosis that took three to five days and whose development and efficacy ranged between 60-90%. The implementation of new, more sensitive and faster techniques has been a very important step forward in the control, management and protection of bivalve mollusc populations and cultures. Since its implementation, the application of the technique of DNA amplification by Polymerase Chain Reaction (PCR) has revolutionized the diagnosis of infectious diseases not only in aquaculture but also in humans, as brought to light by COVID-19. Sensitivity and speed are the most notable qualities of these techniques.

In this practical lesson, our aim is to determine the presence of parasites in different samples of bivalve molluscs via PCR amplification of a DNA fragment whose sequence is specific to the parasite. This specific sequence is a fragment of the intergenic spacer of ribosomal genes (Figure 1). The genes coding for three of the four RNAs that form part of the ribosome (named 18S, 5.8S and 28S ribosomal RNAs) are arranged to form a transcription unit composed of the ETS sequence (external spacer transcribed upstream of the 18S gene), the 18S gene, ITS-1 (internal spacer between the 18S and 5. 8S gene), the 5.8S gene, ITS-2 (internal spacer between the 5.8S and 28S gene), 28S and another ETS (external spacer that is transcribed downstream of the 28S gene). At a ribosomal locus, several hundred of these transcription units are repeated in tandem and separated by an NTS (non-transcribed spacer) sequence. Together, the NTS and the ETSs constitute the so-called IGS (intergenic spacer). Ribosomal genes are characterized by a high degree of conservation through evolution, that is, individuals belonging to evolutionarily distant species show a very high level of similarity of base sequences for these genes. In contrast, this is not the case for the spacers between these genes, which, since they do not encode any gene product, are not subject to selective pressure, and therefore their sequence is highly variable between species. We will use this characteristic, which confers a high diagnostic value, in order to identify those cultures that are contaminated by *Perkinsus sps* and those that are not.
Thus, we will use genomic DNA as a template to amplify a DNA fragment (a 760 bp fragment of the NTS of *P. olseni*) that has a specific sequence of the parasite.


**Figure 1.** Organization of ribosomal genes in eukaryotic genomes. Arrows indicate where a specific pair of primers hybridizes (by hydrogen bonding between base pairs) the DNA template.

## 1.3. METHODOLOGY

***Amplification Reaction (PCR)***

In a 200µl microtube add, following the indicated order, the following reagents for a final volume of 25µl:

- Sterile water 16 µl
- PCR buffer (10x) 2.5 µl
- 2mM of each dNTPs (10 mM) 1 µl
- Primer PkI (0.2 µM) 2 µl
- PkII Primer (0.2 µM) 2 µl
- Clam DNA 1 µl
- Taq polymerase (2U) 0.5 µl

The microtubes are then placed in the thermocycler and the thermocycler is programmed for 35 cycles according to the following program:

| | | |
|---|---|---|
| Denaturation: | 94ºC | 30 sec. |
| Annealing: | 58ºC | 30 sec. |
| Extension: | 72ºC | 30 sec. |

DNA samples from different clams will be analyzed to determine whether or not they are infected with the parasite.
Once the PCR is finished, the samples will be loaded on an agarose gel and this will be subjected to an electric field in order to identify the specific bands of the parasite's DNA that will allow us to undertake a diagnosis.

***Preparation of the agarose gel***

In a 250 ml flask, add 40 ml of TBE buffer (0.04 M Tris-acetate; 0.04 boric acid; 0.01 M EDTA) and 0.4 g of agarose (1% agarose).
Heat it using the microwave until the agarose has melted. Allow to cool to approximately 50°C and add 4 µl of SYBR® Safe DNA Stain Solution (10,000x).
While the agarose is cooling, place the mold in which the gel will be prepared in the adapter. Leave the mold in the adapter on a horizontal surface and place the comb that will carve the wells a few centimeters from the edge.
Once the agarose has cooled, the solution is added to the mold taking care to remove any bubbles that form. Allow the agarose to gel until it acquires a translucent appearance. Remove the comb and mold from the adapter.

Place the gel in the electrophoresis cuvette and cover it with electrophoresis buffer (TBE 1X). The buffer in the electrophoresis cuvette must be the same buffer that was used to dissolve the agarose (in our case, TBE 1X).

### *Electrophoresis*

Taking care not to break the wells, load the different samples corresponding to each of the amplification reactions on the gel. To do this, add 5 µl of loading buffer to the tubes in which the PCR was developed and carefully drop several times into the same tube. Once mixed with the DNA load the mixture into a well of the gel with a micropipette (one sample per well). Load 4 µl of the already prepared mixture of molecular weight marker (ladder) into another well, which will serve as a reference to determine the size of the fragments that we want to characterize.
Connect the power supply to the gel for 30 minutes at 50 volts/cm.
Analyze the results by observation in a transilluminator.

### *Diagnosis of individuals*

In those individuals where we observe an amplification corresponding to 760 bp, the parasite will be present, and therefore, we can diagnose them as positive for this disease. A result like the one in the figure is expected:

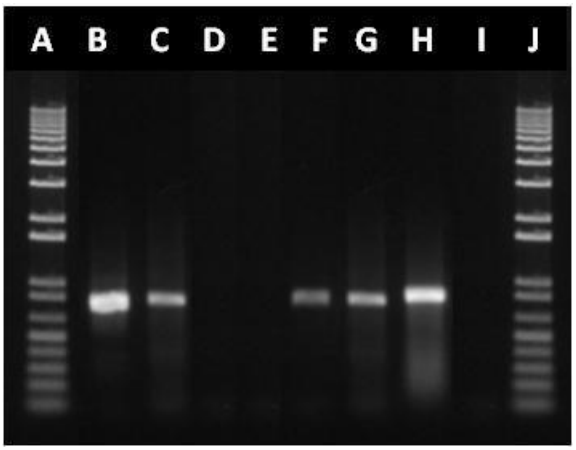



**Figure 2**. The result of the diagnostic test for *Perkinsus* is shown. In an electrophoresis gel, different samples were loaded with the content of the amplification reaction: A and J, DNA marker to determine the size of the amplified fragment. B, amplification of product from a sample used as a positive control (*Perkinsus* DNA). C, amplification of product from a sample used as positive control (DNA from infected animal). D and E, absence of product amplification in clam tissue samples from uninfected cultures. F-H, amplification of the product in clam tissue samples from infected cultures. I, absence of amplification of the product in samples devoid of biological material (negative control, without DNA in the amplification reaction). In this Figure, the arrow points to the amplified DNA fragments (760 bp).The efficacy of the diagnostic method for *Perkinsus olseni* in clam cultures is demonstrated by using the primers PkI and PkII, to detect the presence of the parasite in infected clams. These primers did not cause false positives since non-infected cultures did not show amplification. In addition, the great sensitivity of the method is demonstrated by detecting the presence of the parasite in cultures even when the level of infection is minimal.

## 1.4. WEB RESOURCES

Through the following YouTube link you can access different videos of interest for Genetics subjects.

https://www.youtube.com/playlist?list=PLBa9sJUx0zXWnO2Wu4H6qmJrEOIFNCzaI

Of special use for this practice are the video-tutorials for the preparation of an agarose gel, and the one that bears the name of this practice "Application of PCR to genetic diagnosis: detection of parasites that infect molluscs", which show each and every one of the steps carried out in the laboratory during the practical session.

## 1.5. QUESTIONS

**1.** What criteria should be followed when designing primers for this type of analysis?
**2.** What is a negative control and a positive control in the PCR technique?
**3.** What would you do if you saw amplification in samples that we clearly know are not infected?
**4.** What makes PCR an ideal diagnostic technique?
**5.** Could we completely rule out an infection if we observed absence of amplification on a sample after PCR?
**6.** Could PCR detect the presence of several parasites at the same time in an experiment similar to the one carried out in this practice?

# 2.- CLONING A PCR PRODUCT

## 2.1. AIM

The aim of this practice is to learn the procedure to clone a DNA fragment. To do so, we will construct a recombinant DNA molecule which, in our case, will consist of a cloning vector and a DNA fragment from the parasite *Perkinsus olseni*. We will use as a cloning vector a plasmid that has a number of particularly favorable characteristics 1) it facilitates the insertion of a DNA fragment, 2) it replicates autonomously in prokaryotic cells (*E. coli*) and 3) it allows distinguishing between colonies that have incorporated recombinant plasmids and those that incorporated plasmids without insertion. In this way, the aim is to present a very useful technique that is routinely used in Molecular Genetics laboratories.

## 2.2. THEORETICAL BASIS

In Molecular Biology, the term cloning refers to a technique by which a DNA fragment of interest is introduced into a vector, and this "genetic construct" is then introduced into bacterial cells, so that it can be maintained and multiply (replicate) within them.
Therefore, the main components of a cloning experiment are: a) the DNA fragment to be cloned, which is called the insert once it is integrated into the vector,
b) the cloning vector, where the insert is introduced allowing its incorporation into the cell, and c) the bacteria where the construct formed by insert plus vector (recombinant plasmid) is introduced, allowing many copies of the insert to be obtained.
These types of experiments are included in what is known today as recombinant DNA technology, given that DNA molecules made up of fragments of different origins are constructed.
The insert can be any DNA fragment, whatever its origin. However, the maximum insert size is limited by the capacity of the vector used.
In the case of the most common plasmids, the size of the insert usually does not exceed 10 kilobases (Kb), and a larger insert usually generates a recombinant construct whose size hinders its efficient penetration into bacterial cells. When it is necessary to clone larger DNA fragments, we can resort to other vectors, such as the lambda phage, in which it is possible to clone fragments of about 15 Kb, cosmids, which can accept up to 40 Kb, BACs (Bacterial Artificial Chromosome), which can accept up to 200 Kb, YACs (Yeast Artificial Chromosome), which can accept up to 2 Mb and MACs, which allow cloning fragments of several Mb.
To obtain the insert of interest, a source of DNA including the insert must be obtained, e.g. from the genome of an animal or plant. Then a technique must be selected that allows us to isolate the fragment of interest, e.g. by digestion of the genomic DNA with restriction enzymes (see protocol below), or by PCR amplification of the insert (see script and practice for this technique). In both cases we will obtain a DNA fragment of known size, so we will perform an agarose gel electrophoresis, identify the appropriate fragment and the DNA will be recovered by a DNA purification technique from agarose gels.
In this practice we will use a plasmid as a cloning vector. A plasmid is a circular bacterial DNA molecule that, not being essential for the survival and multiplication of the bacterium, can coexist and replicate in the cell protoplasm as an extra-chromosomal molecule and be transmitted to daughter cells. Therefore, for a plasmid to be used as a cloning vector, it has to be able to maintain and replicate within the cell. This is possible because the plasmid contains a sequence called the origin of replication, specific to each bacterial species, where DNA polymerase binds and replication of the plasmid begins. In terms of the number of copies within a bacterium, plasmids can be classified into two categories: 1) relaxed, if there are

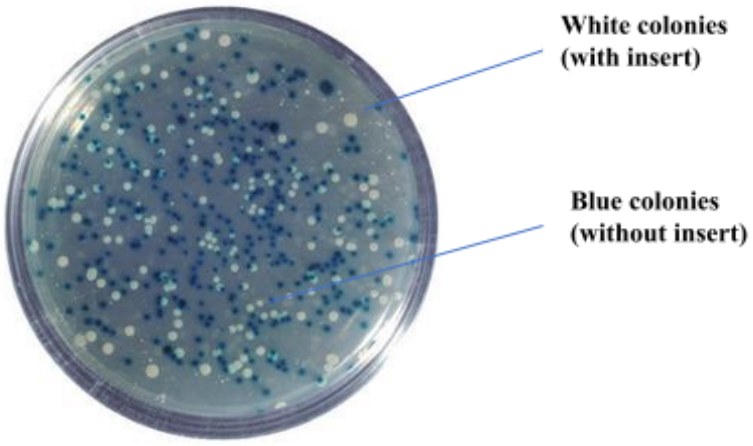
multiple copies, and 2) stringent, if there are only one or very few copies per cell. Relaxed plasmids tend to be more advantageous as they allow efficient multiplication of the plasmid and thus the insert.

To facilitate the integration of the insert, plasmids have a region containing several target sequences specific to various restriction enzymes (polylinker site). DNA fragments obtained by restriction enzyme digestion can be inserted into this region, the targets of which are located in this multiple cloning site. To clone PCR-amplified inserts, open (non-circular) vectors whose 3' ends terminate with a protruding thymine nucleotide (which has no complementary nucleotide on the other strand) are commonly used. These plasmids take advantage of the fact that Taq polymerase (the enzyme that allows DNA amplification by PCR) adds an adenine nucleotide to each 3' end of the amplified DNA. The adenines at the ends of the amplified fragment can be paired with the thymines at the ends of the plasmid, which facilitates the insertion of the amplified (the PCR product) into the plasmid.

During the introduction of plasmids into bacteria (a process known as transformation) only a proportion of the bacteria will incorporate the plasmid (transformation efficiency is never 100%). The plasmids commonly used for this purpose also contain one or more antibiotic resistance genes, allowing selection of the transformed cells (which have incorporated the plasmid). For this purpose, after the transformation process, all bacteria are cultured in a medium containing the antibiotic to which the plasmid confers resistance and, as a consequence, only those that have incorporated the plasmid will survive. In addition, the plasmid may contain a system that allows it to discriminate between cells carrying the vector with the insert and those carrying the recirculated (non-inserted) vector.

One widely used system uses the sequence of the β-galactosidase gene (lacZ gene, from the lac operon of E.coli) interrupted by the multiple cloning region (polylinker). Expression of the β-galactosidase gene requires the presence of IPTG, a molecule that acts as a continuous inducer of the gene.

The β-galactosidase protein, in the presence of one of its substrates, X-gal (5-Bromo-4-Chloro-3-Indol-β-D-galactoside), produces a blue precipitate, as X-gal is hydrolyzed by the enzyme, giving rise to galactose and 5-bromo-4-chloro-3-hydroxindole, which is oxidized to give rise to 5,5'-dibromo-4,4'-dichloro-indole, an insoluble blue compound. Thus, if we grow bacteria transformed with a plasmid containing the β-galactosidase system on a solid medium in the presence of IPTG and X-gal, bacteria that have incorporated recombinant plasmids (with multiple cloning site inserts) will have the β-galactosidase gene inactivated, and the blue precipitate will not form (white colonies), while those transformed with plasmids without an insert will be able to produce the enzyme, as they have an intact gene, and will produce blue colonies (Figure 1).




There are other similar strategies that can be used for the same purpose. For example, using plasmids containing the sequence of a lethal gene interrupted by the multiple cloning site. In

this case, inclusion of the insert in the multiple cloning site will disrupt the lethal gene, with the bacteria transformed with insert plasmids being the only ones to survive.
Nowadays, for conventional uses, researchers do not need to build their own vectors, as there are a wide variety designed and produced by biotech companies for all kinds of cloning uses (e.g. pGEM-4Z vector (uses the β-galactosidase gene as a selection marker) and TOPO vector (uses a lethal gene).The bacterial species and strain used during the cloning process also requires a number of special characteristics. It must be non-pathogenic (obviously to avoid risks to research staff and the general population) and it must be easy to culture (non-pathogenic strains of Escherichia coli are used).
It is preferable for it to have efficient replication (multiplication) and be modified in a way that prevents recombination between the plasmid (vector) and its own chromosome (otherwise, the insert risks being lost). Naturally, a bacterium can acquire a physiological state that enables ("permeabilizes") it to undergo a transformation process. In this situation, the bacterium is said to be "competent". However, this natural competition occurs at a very low frequency and is not useful for cloning purposes. Therefore, in the laboratory, this state is artificially induced by various methods, a process known as "artificial competition". Such permeabilization can be induced by chemical methods.
For this to occur, cells are cooled in the presence of divalent cations such as Ca2+ (in the form of CaCl2), which prepares the cell membranes to be permeable to plasmid DNA. The cells are then incubated on ice with the DNA and then briefly subjected to a heat shock (e.g. 42°C for 30-120 seconds), which facilitates DNA entry into the cell. Permeabilization can also be achieved using physical elements, such as electric current. In this case, bacterial cells are subjected to an electric current of high voltage (around 2000V for bacteria) and short duration (several μs). As in the case of vectors, a wide variety of "competent" bacterial strains are available, provided by biotech companies for all kinds of cloning uses (most notably the E. coli DH5α strain).
Among the many uses of cloning, we can cite the multiplication of copies of a fragment, since, as the recombinant plasmid replicates inside the cell and the cell multiplies, many DNA molecules are obtained. Cloning also allows discrimination between different sequences or variants of amplified DNA. Typically, each transformed bacterium acquires a single recombinant plasmid. When grown on a solid medium, each bacterium will give rise to a colony of bacteria identical to the original and with the same insert. Sequencing the inserts from different colonies will give us an idea of the variability of the original DNA sequences. Another utility of cloning is the generation of a genomic library, which consists of a set of bacterial clones, each of which carries a DNA fragment of the genome of the species under study. Each fragment is included in a clone, and between all the clones, they make up the entire genome. Gen libraries may also contain cDNA fragments (complementary or copy DNA), obtained by mRNA reverse transcription. In this case the number of clones is smaller as only genes that were expressed in the tissue used to extract the mRNA will be represented. Likewise, the inserts will generally be smaller in size as the cloned genes will not contain introns.
Cloning can also allow a gene to be expressed inside a bacterial cell. This requires cloning the in-phase fragment with the open reading pattern of the gene (usually the cDNA obtained from the messenger RNA) into an expression vector. The expression vector has a special promoter that allows controlled induction of transcription of the inserted sequence. As a result, bacteria can synthesize the protein encoded by the insert, allowing the production of enzymes and other proteins of scientific, pharmacological or commercial interest.
In this practice, we will clone DNA fragments that we have previously amplified by PCR. These fragments contain a region of the NTS spacer DNA of the ribosomal DNA (rDNA) of the parasite *Perkinsus olseni*. As a cloning vector we will use the plasmid pGEM-42, which has the characteristics described above.

## 2.3. METHODOLOGY

To carry out the cloning, we will follow the following steps:

***Obtaining the fragment to be cloned and the cloning vector***

The DNA to be cloned will be the product obtained in the PCR practice (see the script and practice corresponding to this technique). The cloning vector corresponds to the commercial plasmid pGEM-42 (Promega). In a PCR reaction, the Taq-polymerase has a transferase-terminal activity, not dependent on the DNA template, which adds an adenine nucleotide at the 3' ends of the amplified products. The pGEM vector is in linear form and has a thymine nucleotide at its 3' ends. This allows much more efficient binding between the amplified fragment and the vector (Figure 2).

***Ligation***

Ligation of the DNA fragments obtained by PCR with the pGEM vector. This process involves an enzyme called ligase, which establishes a phosphodiester bond between the last base of the PCR-amplified product (A) and the first base at the ends of the vector (T) without incorporating a new nucleotide. This results in binding between the DNA strands corresponding to the vector and the insert (Figure 2).
Steps for reaction:
Into an Eppendorf microtube add:

- 7 µl PCR product (100-200 ng)
- 1 µl buffer 10X
- 1 µl the pGEM- T vector
- 1µl ligase enzyme

2. Incubate for 30 minutes at room temperature.

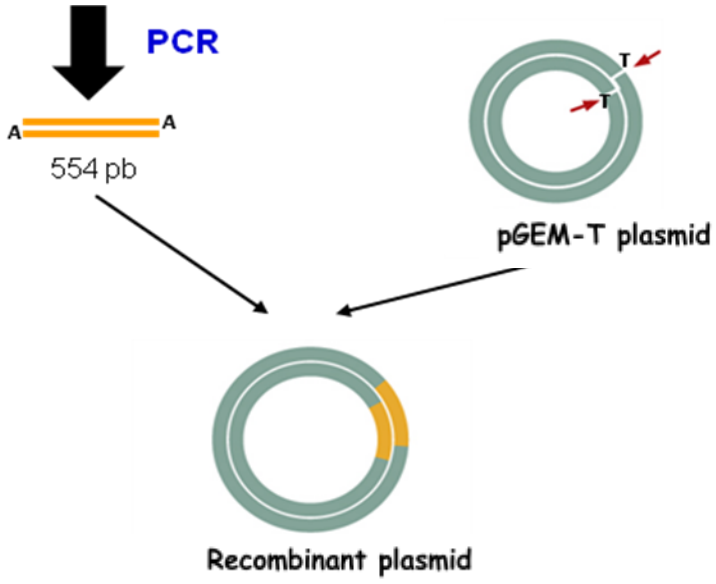



Figure 2: Ligation of a PCR product with the pGEM-T cloning vector

***Transformation***

As stated in the introduction, transformation is the process by which plasmids are introduced into bacterial cells. For this purpose, we will use competent bacteria of E. coli strain JM109. We will proceed as follows:
Place an Eppendorf tube containing 50µl of competent bacteria on ice for a few minutes.

- Add 10 µl of the ligation solution.
- Leave for 20 minutes on ice.
- Heat shock, place the Eppendorf tube with the mixture in a bath at 42°C
    - for 45 seconds.
- Immediately place the microtube on ice for 2 minutes.
- Add 1000µl of liquid LB culture medium.
- Incubate for 30-40 minutes at 37°C with agitation.
- Plate 60µl of the liquid culture on solid LB medium with Ampicillin, X-gal and IPTG.
- Incubate the plates in an inverted position overnight in an oven at 37°C.

***Observation of results***

After an incubation period of 16-24 hours, the resulting plate will show white colonies (with the recombinant plasmid).

## 2.4. WEB RESOURCES

Access a video tutorial through the following YouTube link:

https://www.youtube.com/playlist?list=PLBa9sJUx0zXWnO2Wu4H6qmJrEOIFNCzaI

## 2.5. QUESTIONS

**1.** What are competent cells? What are their characteristics?
**2.** In the cloning process, in which step is the recombinant plasmid introduced into the bacterium?
**3.** Could the pGEM-42 vector be used to clone a DNA fragment cut by restriction enzymes? And one amplified by a high-fidelity DNA polymerase?
**4.** What are the possible causes for the absence of white colonies on the plate?
**5.** What are the possible causes for the absence of colonies on the plate after proper incubation at 37°C?

# 3.- DNA AND PROTEIN SEQUENCE DATABASES

## 3.1. AIM

The aim of this practical course is to introduce the student to the knowledge and handling of DNA and protein sequence databases.

## 3.2. THEORETICAL BASIS

### 3.2.1. Biological information

Like all sciences, biology is constantly generating ever-increasing amounts of information. On a daily basis, biologists are constantly making discoveries and producing data (information) on aspects related to living beings. This information ranges from basic characteristics (e.g. the molecular structure and three-dimensional configuration of a protein) to more complex aspects (e.g. the taxonomy, phylogenetic relationship and ecology of organisms).
At the same time, in order to know the status of a research topic, biologists need continuous access to information and data previously obtained by other researchers. In addition to the literature, geneticists, for example, need information on commonly used methodologies, techniques and reagents. But they also need other information about the species under investigation, such as information about sequences of genes (or DNA in general) or proteins, their variants, their function, the interactions of these genes with other genes, their relationship with sequences in other organisms, what is known about their expression pattern, silencing (or mutation) effect, etc.
The information already available on a particular topic is the basis on which new ideas are developed, and the knowledge of this information prevents repeated research on well-known facts. It could be said that the advance in scientific knowledge has a first, very important and necessary, step which is the review of the research work that has been carried out so far on the subject in question.

### 3.2.1. Storage of biological information

Much of the information that is acquired is easily forgotten unless it is stored in one or more ways. In the past, and even in some civilizations today, information storage is carried out through the so-called collective memory, which is transmitted orally from parents to children. The disadvantages of this traditional way of transmitting information are the limitation of the amount of information that can be ``stored'' and the (almost inevitable) risk of deformation of the information. Storing information in written form offers virtually unlimited information storage capacity and absolute reliability. Before the existence of computers, in science, as in other disciplines, the only way to publicize, store or obtain information from experiments already carried out was by publishing them in scientific journals. The most relevant information published ended up in scientific books and textbooks. Under these conditions, obtaining bibliographies, methods or previous information could become an obstacle as one had to have physical access to the journal(s) containing the information sought. This meant that, in addition to having to buy as many books as possible, it was necessary to subscribe to scientific journals and keep all the copies in a way that allowed one to know where the information was and to be able to retrieve it when needed.
The development of personal computers, limited at first by their limited storage capacity, was a significant breakthrough as it made it possible to store information in digital form. But it was still necessary to rely on physical material (floppy disks) to obtain the digitized information or

to transfer it between computers. However, in the same way as science was done before the discovery of electricity and even the typewriter, it was not until the 1980s that researchers could rely on the powerful tool that is the internet. Since its appearance, the internet has meant a quantitative and qualitative leap in the publication, storage, search and retrieval of data. With access to the internet, from anywhere in the world, a researcher can obtain everything from bibliographies to data on the gene or protein of interest, including sequences, variants, homologous sequences, expression data, effect data, mutation data, function data, etc. In addition, virtually all scientific journals are now available online (many require a subscription but others are freely accessible). Even many articles published before the computer came into existence are now digitized. The internet, together with the increasingly powerful storage capacity of computer hard disks, offered the possibility of centralized ways of storing and organizing information in the form of databases.

Databases are now, in regards to genetics, vital research tools. For today's genetics, it is essential to have access to sequences of DNA (including genomes), RNA (including transcriptomes) and proteins (including proteomes) that have already been identified. Information on gene pathways and networks that provide information on gene interactions is often also required. This, together with information on the expression, function and evolution of the gene of interest, is available in increasingly comprehensive databases. It is no exaggeration to say that today's geneticist cannot do research without access to databases.

As far as the analysis of DNA sequences or proteins is concerned, researchers now have at their disposal databases where these sequences are stored along with their variants, homologous sequences, and a large amount of information on their chromosomal location, properties, expression, function, phylogenetic relationships, etc. Obviously, the provenance of these sequences is science itself, since every time a research group identifies a sequence, or genome, it uploads them to the database, and uploading sequences to the databases is a requirement for publication in scientific journals of findings related to that sequence.

Sometimes, the enormous logistics required to build and maintain a database depend on individual scientific projects (in the case of organism-specific databases) or on a governmental or even intergovernmental effort (as in the case of more widely used general databases). Examples of the first case include databases on model organisms (Figure 1):

Examples of organism-specific databases (Figure 1):
Fruit fly, *Drosophila melanogaster* (http://flybase.org/)
The nematode *Caenorhabditis elegans* (http://www.wormbase.org/).
The plant *Arabidopsis thaliana* (http://www.arabidopsis.org/).

Most relevant general databases (Figure 2):

- *The DNA DataBank of Japan (DDBJ, http://www.ddbj.nig.ac.jp/index-e.html)*
  - *The European Molecular Biology Laboratory (EMBL,) and its "sister" The European Nucleotide Archive (http://www.ebi.ac.uk/ena/)*
  - *GenBank, a database of the US-based The National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/genbank/).*
  - *From them stems the international project The International Nucleotide Sequence Database collaboration (http://www.insdc.org/).*


Figure 1: organism-specific databases


Figure 2. Most relevant general databases

In addition to research staff, databases also rely on algorithms (software) that allow the automation of the process of obtaining, storing, organizing and efficiently presenting and accessing their content. Other algorithms, integrated in the databases, allow the analysis of sequences of interest and their comparison with other sequences (examples of these are the well-known sequence alignment and comparison programs *Clustal* and *Blast*).

### 3.2.3. Identification and formatting of nucleotide and amino acid sequences in databases

Four types of searches can be used to retrieve a sequence from a database. The sequence can be found by using the name of the gene and corresponding organism (or by using the name of the gene and then selecting the organism of interest); Figure 3 shows the result of a search in the gene directory of the NCBI database for the sequence of the collagen genes in humans. However, the sequences in the databases are catalogued and labeled with a unique accession number and identifier, and informative labels about their origin and other characteristics (see sequence formats). This offers the possibility of directly finding the sequence by searching for its accession number or identifier. In the case of the human collagen type 3 alpha 1 gene, the sequence can be obtained by searching the GenBank gene

directory (the most comprehensive database) by the accession number X15332, or by the identifier COL3A1.
Two more indirect ways of obtaining the sequences are by Blast search:
http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastHome
with homologous sequences (sequences of the gene belonging to phylogenetically close organisms), or by browsing the corresponding chromosome using genome browsers such as Genome Browser (http://genome.ucsc.edu/cgi-bin/hgGateway) or Ensembl (http://www.ensembl.org) if the genome of the organism and the location of the sequence of interest are known. In the case of the collagen gene, it would be necessary to navigate around nucleotides 189833342 and 189883227 in band 32 of the long arm of chromosome 2 (chromosome: 2; Location: 2q32.2) (Figure 3)

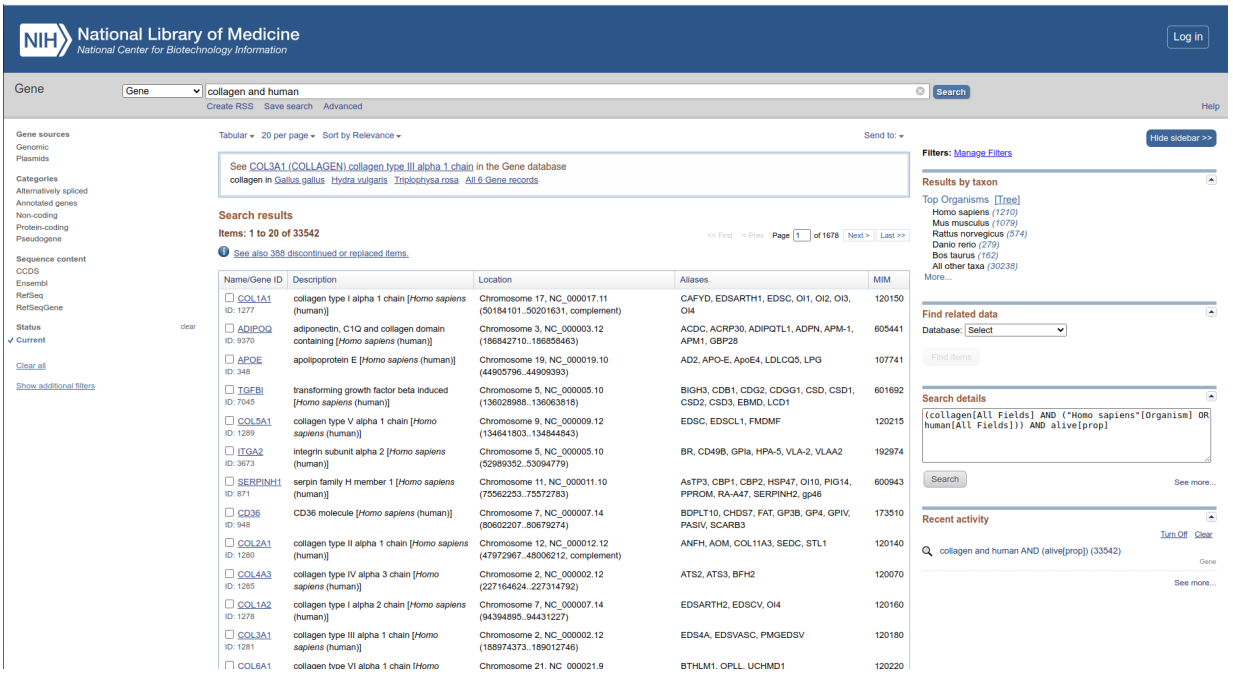



Figure 3: Search for human collagen gene sequences at NCBI

Once obtained, the sequence can be presented in one format or another depending on the database. Here we will introduce the three most commonly used formats, namely, the “European” EMBL format, the “American” GenBank format (both of which include information and several labels identifying the sequence and its origin) and the “simple and universal” fasta format, which may include no more than a header with the name of the sequence.
As mentioned above, the fasta format is the simplest as it includes only a comment part, or title, the beginning of which is marked by the symbol “>”, and which is usually the name of the sequence, its provenance and database accession number, followed by a line break and the nucleotide or amino acid sequence, usually presented in lines of 80 or 120 residues although, apart from the first line break between the title and the sequence, the format ignores spaces and accepts sequences in the form of continuous residues without space or line break. The end of the sequence is simply the last character (residue) of the sequence (see example below). Because it is so simple, the fasta format is the base format required by the vast majority of sequence analysis programs and algorithms, and therefore the one most commonly used by researchers to handle sequences (sequence alignment, phylogenetic trees, blast searches, etc.). The fasta file can be a plain text file or have one of the extensions “.fas” or “.fasta”. A fasta sequence file can have one or several sequences, each with its own identifying line (starting with “>”).

**Example of fasta format** (Figure 4). the dots inside the sequence indicate that we have removed residues to save space, as the whole sequence is about 5kb.

```
>embl|X15332|X15332 Human COL3A1 mRNA for pro alpha-1 (III) collagen

cagaactattctccccagtatgattcatatgatgtcaagtcgggcggagtagcagtaggaggactcgcaggct
atcctggaccagctggcccccccaggcccccccggcccccctggtacatctggtcatcctggttcccctggatc
tccaggataccaaggacccccctggtgaacctgggcaagctggtccttcaggccctccaggacctcctggtgct
ataggtccatctggtcctgctggaaaagatggagaatcaggtagacccggacgacctggagaccgaggattgc
ctggacctccaggtatcaaaggtccagctgggatacctggattccctggtatgaaaggacacagaggcttcga
tggacgaaatggagaaaaagggtgaaacaggtgctcctg...ccctggtccttgctgtggtggtgttggagccc
ctgccattgctgggattggagctgaaaaagctggcggttttgccccttattatggagatgaaccaatg
```

Figure 4: Example of fasta format.

On the other hand, both the EMBL and GeneBank formats are more elaborate and include more identifiers and sequence information. Both share the characteristic of having, in their initial part, annotations indicating the accession number of the sequence and, like fasta, can have one or several sequences, each one marked by its identifier. The left column of the EMBL file contains two letters (abbreviation of the English term) indicating the nature of the corresponding field annotation (e.g. ID is the identifier, KW is the keyword, etc.). The EMBL format starts with a sequence identifier (ID) followed by annotations such as accession number (AC), creation and update dates (DT), description (DE), keywords (KW), organism or species of origin (OS), species classification (OC), bibliographic reference data (pages (RP), authors (RA), title of work (RT), journal, volume, year and pages of publication (RL), and comments (CC). The letters FT mark other features of the sequence such as translation, protein identifier, etc. The beginning of the sequence is marked with the letters SQ and its end with the symbol “//”. Each line of the sequence contains sixty residues, each of which is separated by a ten by ten space. The line ends with a tabulation and the position of the last residue of the corresponding line. The GenBank format has a similar structure to the EMBL format with the following differences: the first line of the file starts with the word “LOCUS” and contains information about the sequence (accession number, name, etc.), and instead of using abbreviations in the first column, as in EMBL, the GenBank format uses a complete word descriptive of the field annotation (in this way the GenBank format is more intuitive than EMBL). The beginning of the sequence is marked by the word “ORIGIN” and, as in EMBL, the end is marked by the symbol “//”. As in the EMBL format, each line of the GenBank sequence contains sixty residues separated by a ten by ten space. In the case of GenBank, however, the line begins with a number marking the position of the first residue of the corresponding line (in EMBL it is the last residue that is marked).

**Example of EMBL format** (Figure 5). Bold dots indicate the same as above.

```
ID   X15332; SV 1; linear; mRNA; STD; HUM; 3234 BP.
XX
AC    X15332;
XX
DT    06-JUL-1989 (Rel. 20, Created)
DT    05-AUG-1995 (Rel. 44, Last updated, Version 2)
XX
DE    Human COL3A1 mRNA for pro alpha-1 (III) collagen
XX
KW    COL3A1 gene; collagen.
XX
OS    Homo sapiens (human)
OC    Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
      Mammalia;
OC    Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini;
      Hominidae;
OC    Homo.
XX
RN    [1]
RP    1-3234
RA    Janeczko R., Ramirez F.;
RT    ;
RL    Submitted (19-MAY-1989) to the EMBL/GenBank/DDBJ databases.
RL    Janeczko R., Ramirez F., Suny Health Science Centre, 450 Clarkson
      Avenue-RL Box 44, Brooklyn NY 11203, U S A.
XX
RN    [2]
RX    DOI; 10.1093/nar/17.16.6742
RX    PUBMED; 2780304.
RA    Janeczko R.A., Ramirez F.;
RT    "Nucleotide and amino acid sequences of the entire human alpha 1 (II
RT    collagen";
RL    Nucleic Acids Res. 17(16):6742-6742(1989).
XX
DR    GDB; 174873.
DR    H-InvDB; HIT000321499.
XX
CC    The sequence overlaps with that reported by Chu et. al. in
CC    J. Biol. Chem. 260:4357-4363(1985), by Toman et. al. in
CC    Nucl. Acids Res. 16:7201-7201(1988) and by Mankoo et. al. in
CC    Nucl. Acids Res. 16:2337-2337(1988).
XX
FH    Key Location/Qualifiers
FH
FT    source 1..3234
FT    /organism="Homo sapiens"
FT    /map="2q31"
FT    /mol_type="mRNA"
FT    /db_xref="taxon:9606"
FT    CDS <1..>3234
FT    /codon_start=1
FT    /product="alpha-1 (III) collagen"
FT    /protein_id="CAA33387.1"
FT    /translation="QNYSPQYDSYDVKSGGVAVGGLAGYPGPAGPPGPPGPPGTSGHPG
FT    SPGSPGYQGPPGEPGQAGPSGPPGPPGAIGPSGPAGKDGESGRPGRPGDRGLPGPPGIK
FT    GPAGIPGFPGMKGHRGFDGRNGEKGETGAPGLKGENGLPGENGAPGPMGPRGAPGERGR
FT    PGLPGAAGARGNDGARGSDGQPGPPGPPGTAGFPGSPGAKGEVGPAGSPGSNGAPGQRG
FT    EPGPQGHAGAQGPPGPPGINGSPGGKGEMGPAGIPGAPGLMGARGPPGPAGANGAPGLR
FT    GGAGEPGKNGAKGEPGPRGERGEAGIPGVPGAKGEDGKDGSPGDPGANGLPGAAGERGA
FT    ..........CCGGVGAPAIAGIGAEKAGGFAPYYGDEPM"
XX
SQ    Sequence 3234 BP; 664 A; 861 C; 1106 G; 603 T; 0 other;
cagaactatt ctccccagta tgattcatat gatgtcaagt cgggcggagt agcagtagga   60
ggactcgcag gctatcctgg accagctggc cccccaggcc cccccggccc ccctggtaca  120
tctggtcatc ctggttcccc tggatctcca ggataccaag gacccctgg tgaacctggg  180
caagctggtc cttcaggccc tccaggacct cctggtgcta taggtccatc tggtcctgct  240
ggaaaagatg gagaatcagg tagacccgga cgacctggag accgaggatt gcctggacct  300
ccaggtatca aaggtccagc tgggatacct ggattccctg gtatgaaagg acacagaggc  360
ttcgatggac gaaatggaga aaagggtgaa acaggtgctc ctggattaaa gggtgaaaat  420
..... attg gagctgaaaa agctggcggt tttgcccctt attatggaga tgaaccaatg 3234
//
```

Figure 5. Example of EMBL format

**Example of GenBank format** (Figure 6). Bold dots indicate the same as above.

```
LOCUS       NM_000000         5400 bp     mRNA     linear     PRI 20-JAN-2011
DEFINITION  Homo sapiens collagen, type III, alpha 1 (COL3A1), mRNA.
ACCESSION   NM_000000
VERSION     NM_000000.3 GI:110224482
KEYWORDS    .
SOURCE      Homo sapiens (human)
ORGANISM    Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires;
            Primates; Haplorrhini; Catarrhini; Hominidae; Homo.
REFERENCE   1 (bases 1 to 5400)
AUTHORS     Kronenberg D, Bruns BC, Moali C, Vadon-Le Goff S, Sterchi EE, Traupe H, Bohm M, Hulmes DJ, Stocker W, Becker-Pauly C
TITLE       Processing of procollagen III by meprins: new players in extracellular matrix assembly?
JOURNAL     J. Invest. Dermatol. 130 (12), 2727-2735 (2010)
PUBMED      20631730
REMARK      GeneRIF: meprins could be important players in several remodeling processes involving collagen fiber deposition
COMMENT     REVIEWED REFSEQ: This record has been curated by NCBI staff. The reference sequence was derived from
            BP374009.1, BC028178.1, X14420.1 and AC006694.7.
FEATURES    Location/Qualifiers
source      1..5400
            /organism="Homo sapiens"
            /mol_type="mRNA"
            /db_xref="taxon:9606"
            /chromosome="2"
            /map="2q31"
gene        1..5400
            /gene="COL3A1"
            /gene_synonym="EDS4A; FLJ34534"
            /note="collagen, type III, alpha 1"
            /db_xref="GeneID:1281"
            /db_xref="HGNC:2201"
            /db_xref="HPRD:00365"
            /db_xref="MIM:120180"
exon        1..196
            /gene="COL3A1"
            /gene_synonym="EDS4A; FLJ34534"
            /inference="alignment:Splign"
            /number=1
CDS         118..4518
            /gene="COL3A1"
            /gene_synonym="EDS4A; FLJ34534"
            /note="Ehlers-Danlos syndrome type IV, autosomal dominant;
            collagen, fetal; collagen alpha-1(III) chain; alpha1 (III)
            collagen"
            /codon_start=1
            /product="collagen alpha-1(III) chain preproprotein"
            /protein_id="NP_000081.1"
            /db_xref="GI:4502951"
            /db_xref="CCDS:CCDS2297.1"
            /db_xref="GeneID:1281"
            /db_xref="HGNC:2201"
            /db_xref="HPRD:00365"
            /db_xref="MIM:120180"
            /translation="MMSFVQKGSWLLLALLHPTIILAQQEAVEGGCSHLGQSYADRDVWKPEP
            CQICVCDSGSVLCDDIICDDQELDCPNPEIPFGECCAVCPQPPTAPTRPPNGQKGDPGPPGIP
            GRNGDPGIPGQPGSPGSPGPPGICESCPTGPQNYSPQYDSYDVKSGVAVGGLAGYPGPAGPPG
            PPGPPGTSGHPGSPGSPGYQGPPGEPGQAGPSGPPGPPGAIGPSGPAGKDGESGRPGRPGERG
            LPGPPGIKGPAGIPGFPGMKGHRGFDGRNGEKGETGAPGLKGENGLPGENGAPGPMGPRGAPG
            ERGRPGLPGAAGARGNDGARGSDGQPGP....VRLPIVDIAPYDIGGPDQEFGVDVGPVCFL"
sig_peptide 118..186
            /gene="COL3A1"
            /gene_synonym="EDS4A; FLJ34534"
proprotein  187..4515
            /gene="COL3A1"
            /gene_synonym="EDS4A; FLJ34534"
            /product="collagen alpha-1(III) chain proprotein"
mat_peptide 577..3780
            /gene="COL3A1"
            /gene_synonym="EDS4A; FLJ34534"
            /product="collagen alpha-1(III) chain"
STS         2303..2528
            /gene="COL3A1"
            /gene_synonym="EDS4A; FLJ34534"
            /standard_name="GDB:178411"
            /db_xref="UniSTS:155007"
exon        2347..2400
            /gene="COL3A1"
            /gene_synonym="EDS4A; FLJ34534"
            /inference="alignment:Splign"
            /number=32
            /gene_synonym="EDS4A; FLJ34534"
STS         5334..5400
            /gene="COL3A1"
            /gene_synonym="EDS4A; FLJ34534"
            /standard_name="WI-16343"
            /db_xref="UniSTS:68589"
STS          5359..5419
            /gene="COL3A1"
            /gene_synonym="EDS4A; FLJ34534"
            /standard_name="COL3A1"
            /db_xref="UniSTS:480020"
polyA_signal  5468..5473
            /gene="COL3A1"
            /gene_synonym="EDS4A; FLJ34534"
polyA_signal  5481..5486
            /gene="COL3A1"
            /gene_synonym="EDS4A; FLJ34534"
polyA_site  5490
            /gene="COL3A1"
            /gene_synonym="EDS4A; FLJ34534"
ORIGIN
          1 ggctgagttt tatgacgggc ccggtgctga agggcaggga acaacttgat ggtgctactt
         61 tgaactgctt ttcttttctc cttttgcac aaagagtctc atgtctgata tttagacatg
        121 atgagctttg tgcaaaaggg gagctggcta cttctcgctc tgcttcatcc cactattatt
        181 ttggcacaac aggaagctgt tgaaggagga tgttcccatc ttggtcagtc ctatgcggat
        241 agagatgtct ggaagccaga accatgccaa atatgtgtct gtgactcagg atccgttctc
       5461 ..........cacccataat aaaatatcat attaaaattc
//
```

Figure 6. Example of GenBank format

## *Genome Browser*

Among the algorithms and utilities that a sequence database can offer is a viewer that allows us to see information about the sequence we are interested in, taking into account its chromosomal location. It can therefore only be used in the case of sequences from organisms whose genomes are partially or completely sequenced, assembled and annotated. This is the tool called Genome Browser. As the name suggests, it allows the researcher to navigate the genome (each chromosome separately). This navigation is not only possible in a horizontal direction (i.e. to see which sequences border our sequence or locus of interest) but also vertically (zoom) allowing the movement between various levels of focus from cytogenetic (for example to see the chromosomal banding information in the region) to the sequence itself and its characteristics (promoter, transcription factor binding site…). In addition, the Genome Browser allows you to include all kinds of information and annotations on each chromosome sequence. Thus, if we go to the Genome Browser for the human genome and search for the collagen gene that we used as an example earlier (COL3A1) we will see that it is indeed (Figure 4) located in the locus between nucleotides 189833342 and 189883227 of the assembly of human chromosome 2, an area that corresponds to the cytological (chromosomal) band 32 of the long arm of this chromosome. We will see that in addition to the information on which chromosome, arm, and region our sequence is located, Genome Browser also offers us information on its nature (gene, promoter, intron, etc.), expression data, variants (including SNPs), data on function, gene interactions, structural data of the protein, the orthologues of the sequence, its phylogenetic relationships, bibliography, etc., in short, all kinds of annotation (information) available on that sequence. All this makes Genome Browser the most informative tool in the case of sequences of organisms with sequenced and (albeit partially) assembled genomes.

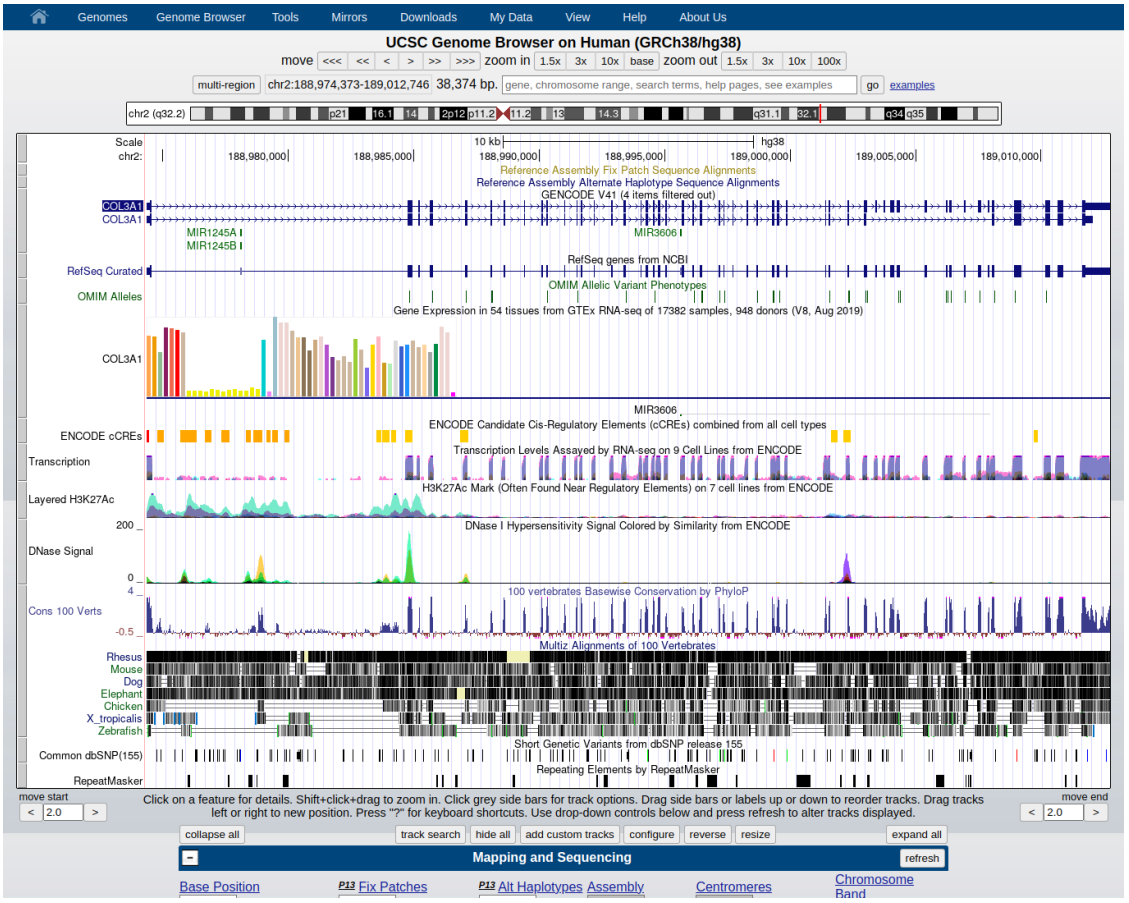


Figure 7. Screenshot showing a part of the result of the search for the human *Col3A1* gene sequence in Genome Browser.

# 3.3. METHODOLOGY

## 3.3.1 Introduction to Bioinformatics

Genetic analysis techniques have evolved extremely rapidly in recent years, having gone from being manual, slow, expensive and producing relatively little information, to being automatic, increasingly faster and cheaper and producing enormous amounts of information. With mass sequencing technologies, for example, whole genome sequences can be obtained in a short time. The storage, processing and analysis of all this information requires the use of fast and powerful computational tools. Bioinformatics is the discipline charged with developing the necessary tools for this (development or programming profile), as well as the use of these tools to carry out the analyses that, in the end, result in biological knowledge.

Bioinformatics tools can be classified as tools for storing and retrieving information (databases) and programs for manipulating and analyzing this information. From the point of view of Genetics, the former are fundamentally databases of DNA and protein sequences, mutations, expression, regulation, methylation, etc., and will be the subject of work in this practical session, while some of the latter (computational gene prediction, multiple alignment and phylogenetic reconstructions, computational analysis of differential gene expression) will be worked on in the following practical sessions.

## 3.3.2 DNA and protein sequence databases

Databases are structured information storage systems that allow the location and retrieval of data of interest in a fast, simple and efficient way, among huge amounts of data, by means of a program called a database engine. In this practice we will look at some of the most widely used DNA and protein sequence databases.

### 3.3.2.1 GenBank

GenBank (http://www.ncbi.nlm.nih.gov/genbank/) is the genetic sequence database of the US National Institutes of Health (NIH), an annotated collection of all publicly available DNA sequences (Nucleic Acids Research, 2008 Jan; 36 (Database issue): D25-30) (Figure 8). The database is hosted on the servers of The National Center for Biotechnology Information in the United States.

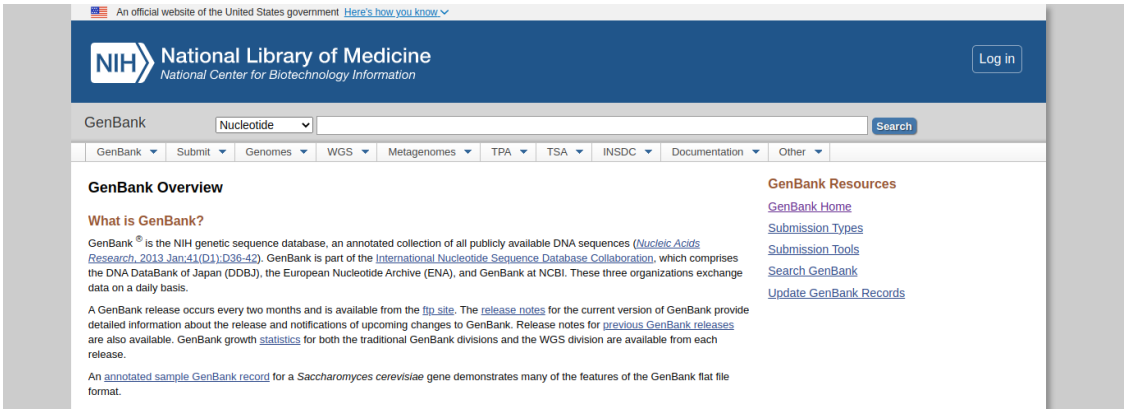



Figure 8. GenBank access webpage

The drop-down menu allows you to choose the database to use (Figure 9), along with a text box followed by a Search button.

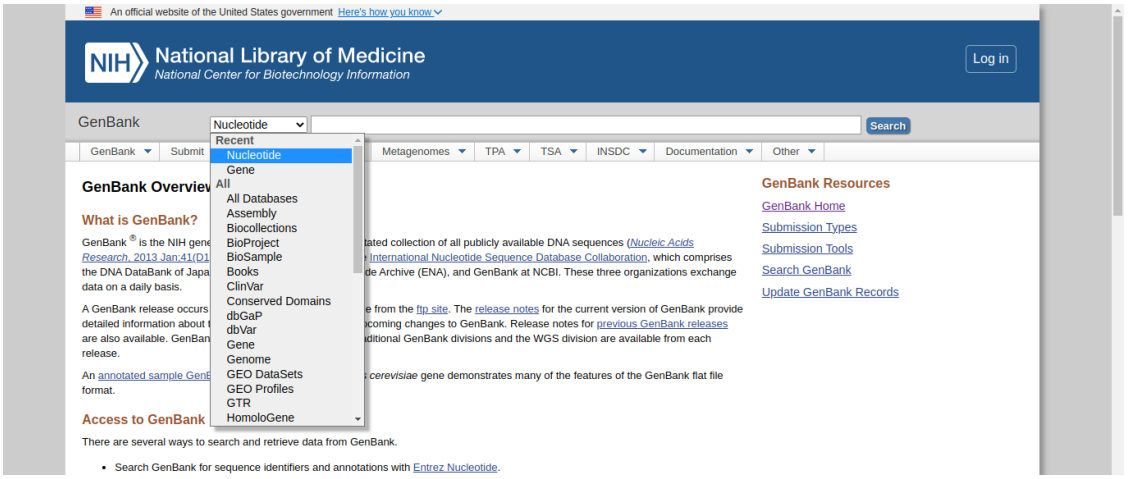



Figure 9. Performing a search in GenBank

To perform a search, we first select the database to use (Nucleotide for DNA, Protein for proteins, PubMed for bibliography, etc.) and then enter a search string in the text box; finally, we click on Search. As an example, if we wanted to search for the sequence of the gene coding for human coagulation factor VIII, we would choose the nucleotide database and type Homo sapiens coagulation factor VIII gene in the text box. The result of that search is shown in Figure 10.

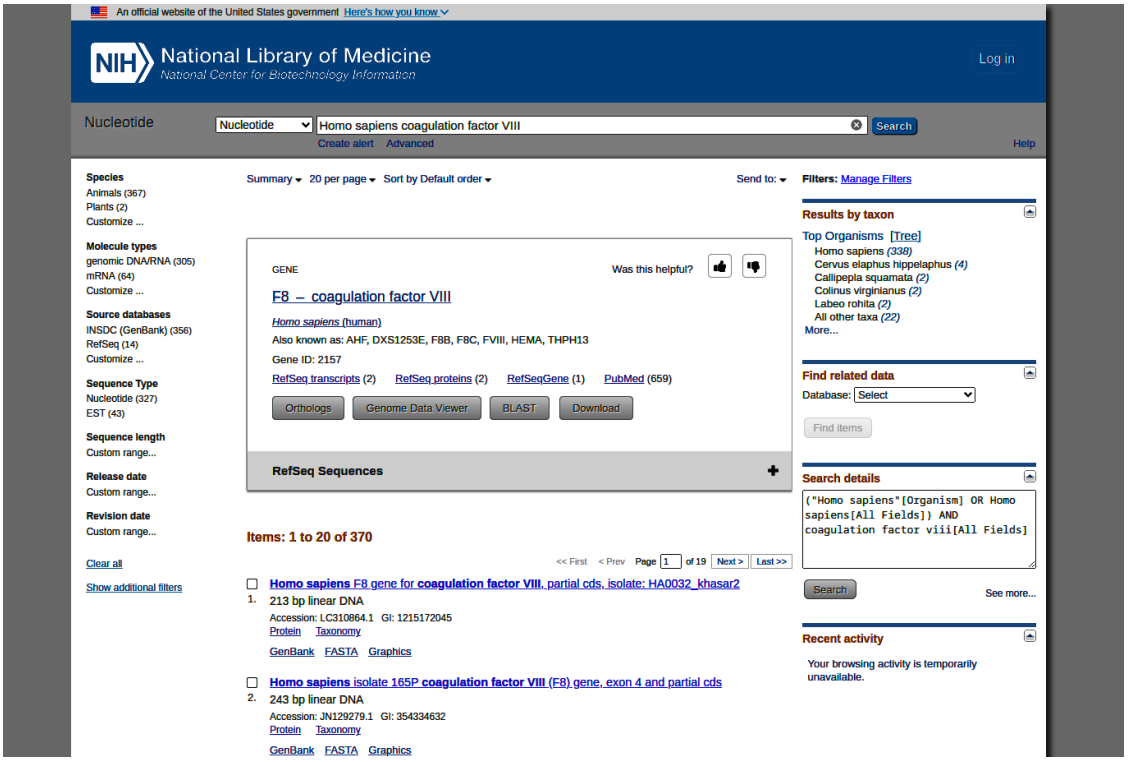



Figure 10. Result of a search in GenBank

As can be seen, at the time it was performed, the search string matched 191 records in the database. For each of the results, the sequence name linked (in blue and underlined) to the database record, the type (DNA or RNA) of sequence and its length, the accession number of

the database record (which uniquely identifies it), and links to the sequence in GenBank and FASTA formats, as well as to a graphical sequence browser and a list of related sequences are shown.

By clicking on the link with the name of the sequence, we access the information stored in the corresponding record, which is structured in different information fields (Figure 11).


Figure 11. DNA sequence record stored in GenBank format

Some relevant fields of the GenBank format are the following:

**Locus**: Contains an identifier (not necessarily unique) of the sequence, as well as its length (1319 base pairs in the example), the type of sequence (linear DNA) and the date of its publication in the database.

**Definition**: Contains more detailed information about the sequence stored in that record.

**Accession**: The accession number of the sequence in the database, which identifies it uniquely.

**Source**: Fields containing information about the origin of the stored sequence, the species to which it belongs and its taxonomic classification.

**Reference**: Fields containing bibliographic references about the sequence, its publication in scientific journals or databases, etc.

**Features:** Contain the annotation of the sequence, which describes what is specifically contained in the different positions of the sequence. In the case of the example in figure 8, the gene coding for coagulation factor VIII in humans.

**Origin**: This is the last field of the record, which stores the nucleotide sequence. The end of the record is marked by the characters "//" on a new line.

The sequence can be accessed in FASTA format (by clicking on the corresponding link at the top left of the page) (Figure 12).

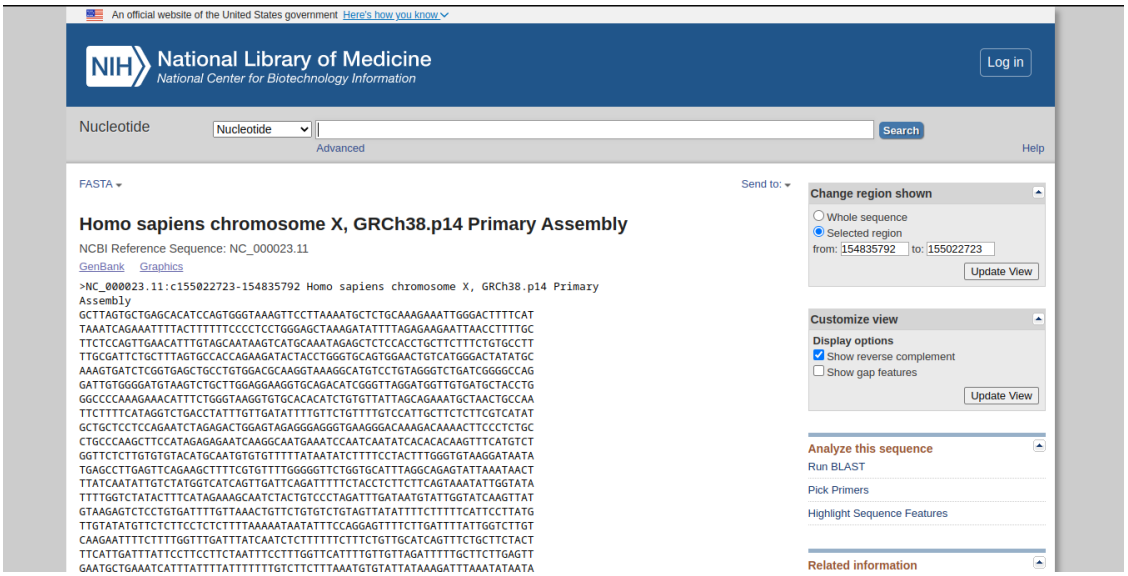



Figure 12. Sequence of coagulation factor VIII in FASTA format

Searching for an amino acid sequence is done in GenBank in an analogous way, by choosing the protein database from the drop-down menu and typing the search string in the text box. An example can be seen in Figure 13, which shows the record corresponding to the protein encoded by the gene in the example above, i.e. coagulation factor VIII in humans.

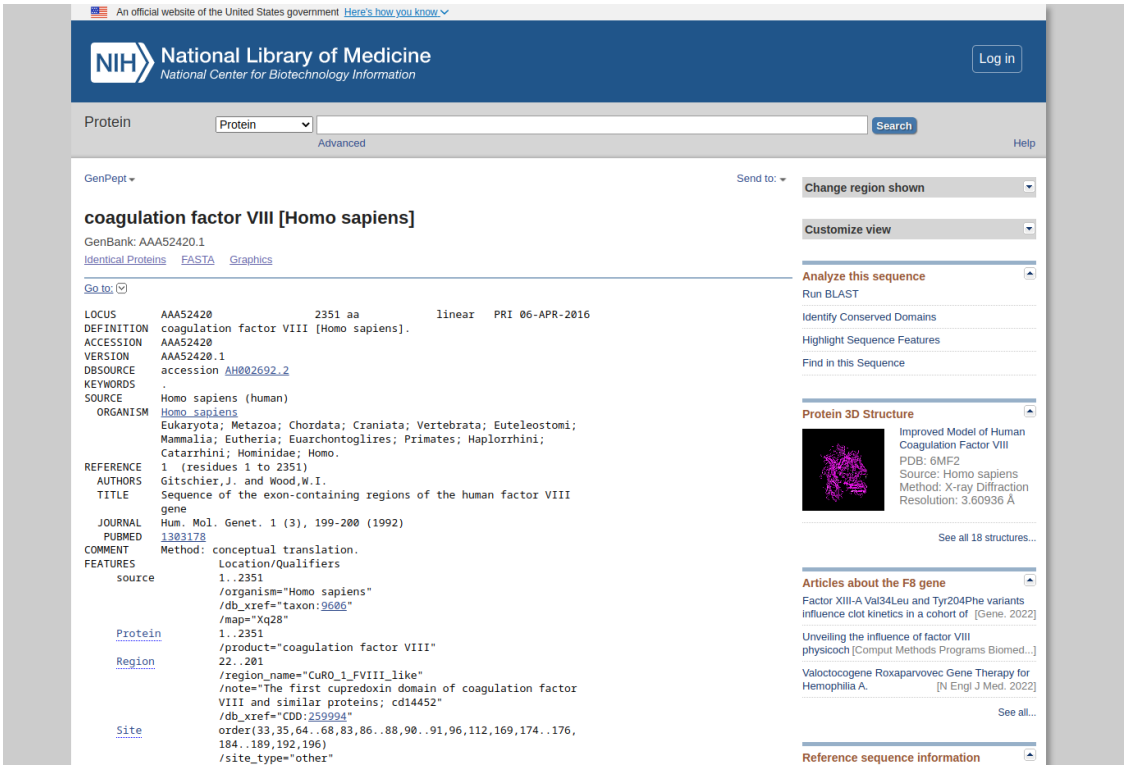



Figure 13. Record of a protein sequence stored in GenBank format

### *3.3.2.2 EMBL*

The EMBL sequence database belongs to the European Molecular Biology Laboratory (EMBL), and is hosted on the servers of the European Bioinformatics Institute (EBI, http://www.ebi.ac.uk/). Like GenBank, EMBL contains databases of DNA and protein sequences, structure, expression, complete genomes, scientific literature, etc. (Figure 14).

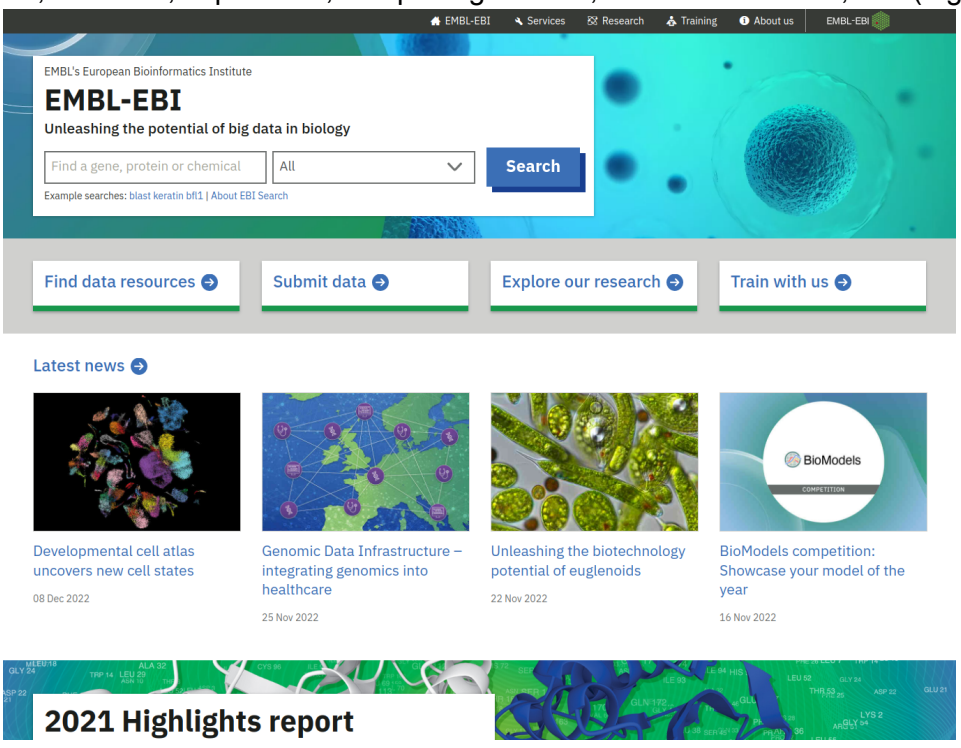



Figure 14. EMBL access page

The use of EMBL databases and the storage format of their sequences (Figure 15) are very similar to those of GenBank.

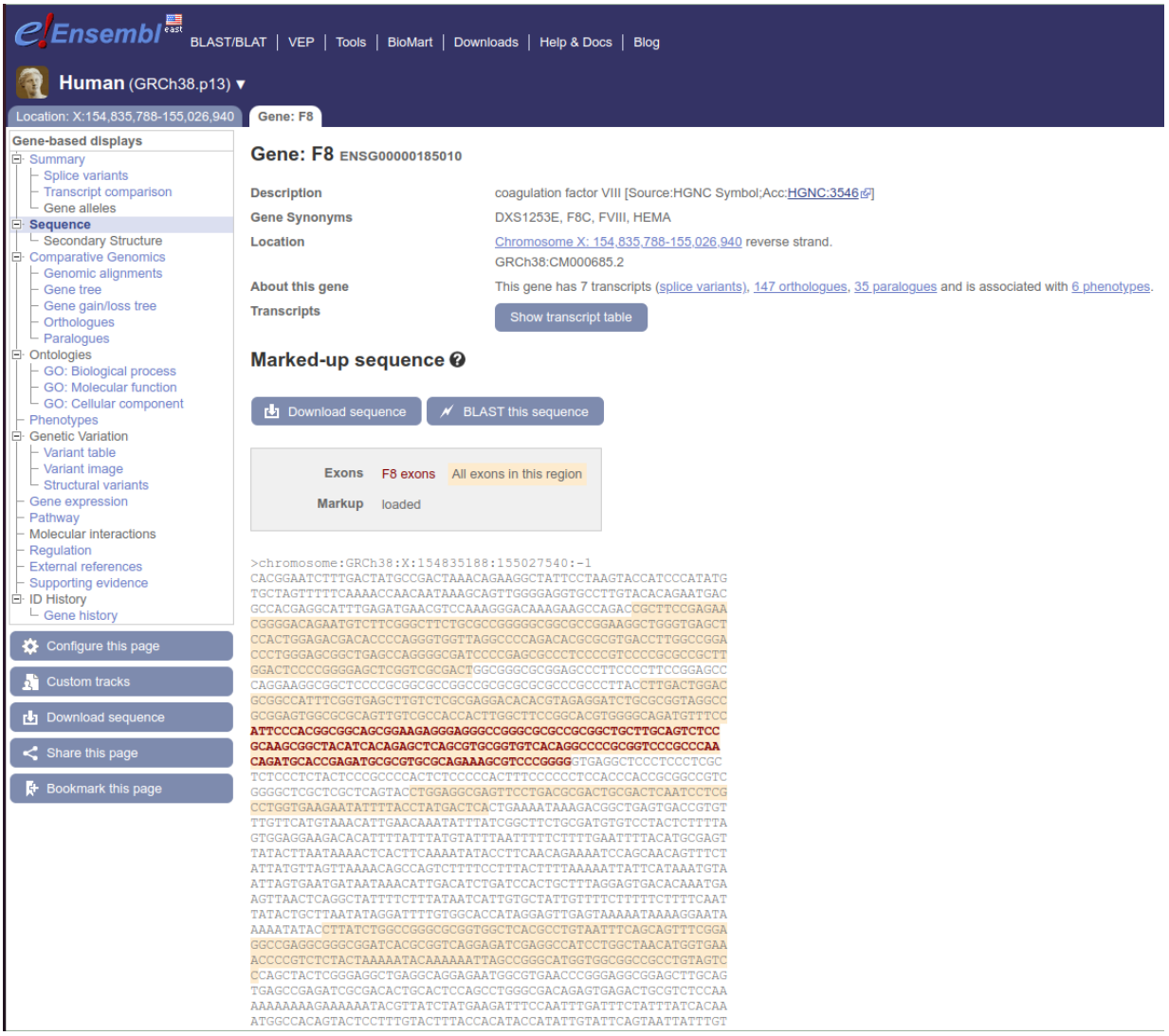



Figure 15. DNA sequence in EMBL format

### *3.3.2.3 UniProt*

UniProt (http://www.uniprot.org/uniprot/) is one of the most widely used protein databases (Figure 16). Querying the database is similar to the previous databases.

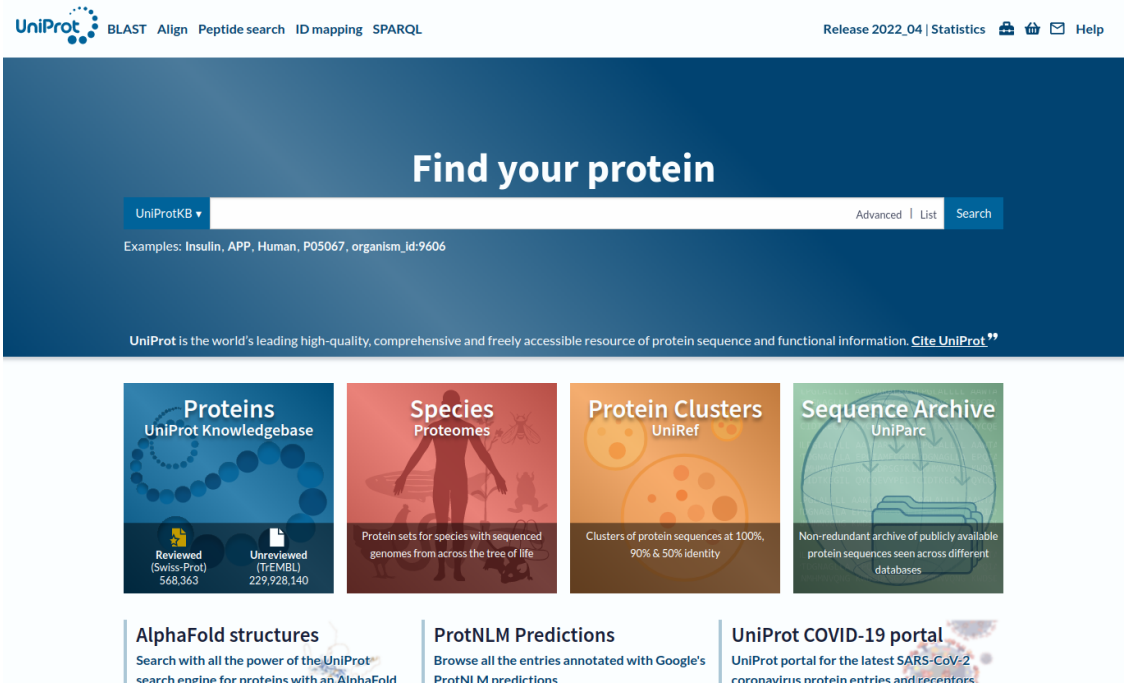



Figure 16. UniProt login page

### *3.3.2.4 Whole genome databases*

Databases that store whole genomes already exist, such as the EBI's Genomes Pages (http://www.ebi.ac.uk/genomes/) (Figure 17).

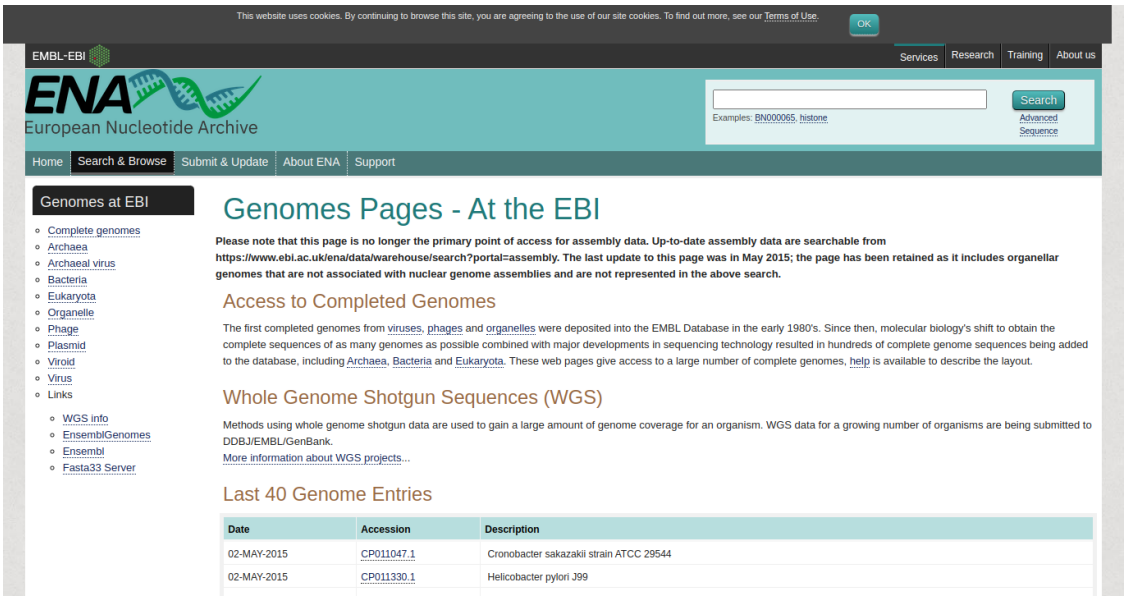



Figure 17. Genomes Pages access page

### *3.3.2.5 Scientific literature databases*

There are also databases of scientific literature, hosted on the servers of the research centers mentioned above, such as Entrez, EMBL, PubMed, NCBI Bookshelf, etc.

### 3.3.3 Searching databases

In addition to searching for DNA or protein sequences by name, species, etc., we may be interested in searching for sequences that show similarity (homology) to a given problem sequence (example in Figure 18), i.e. what is known as database crawling.

MAVMAPRTLV LLLSGALALT QTWAGSHSMR YFSTSVSRPG RGEPRFIAVG YVDDTQFVRF
DSDAASQRME PRAPWIEQEG PEYWDRNTRN VKAHSQTDRV DLGTLRGYYN
QSEDGSHTIQ
RMYGCDVGSD GRFLRGYQQD AYDGKDYIAL NEDLRSWTAA DMAAEITKRK
WEAAHFAEQL
RAYLEGTCVE WLRRHLENGK ETLQRTDAPK THMTHHAVSD HEAILRCWAL SFYPAEITLT
WQRDGEDQTQ DTELVETRPA GDGTFQKWAA VVVPSGQEQR YTCHVQHEGL
PEPLTLRWEP
SSQPTIPIVG IIAGLVLFGA VIAGAVVAAV RWRRKSSDRK GGSYSQAASS DSAQGSDVSL
TACKV

Figure 18. Sequence of an anonymized protein

One of the best-known database tracing algorithms is BLAST, implemented by the BLASTn (DNA) and BLASTp (protein) programs (https://blast.ncbi.nlm.nih.gov/Blast.cgi).

We are going to trace the protein databases with the example sequence in Figure 18 using the BLASTp program (Figure 19 a and b). In the graphical interface of the program, we find a text box where we can paste the problem sequence, as well as a Select File button that allows us to choose a file containing the problem sequence on our computer. Further down, we find the Blast button for the execution of the trace.

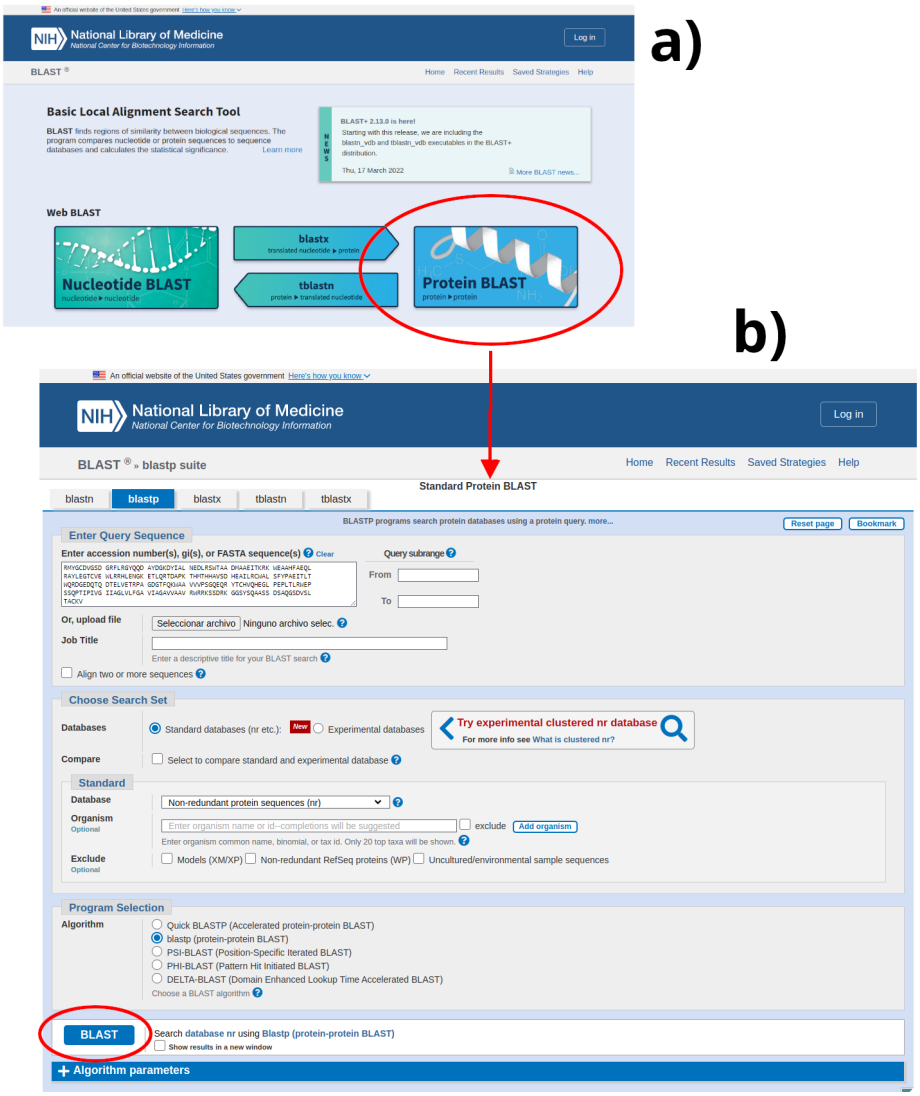



Figure 19. BLASTp access page

The result of a BLASTp scan has three parts, an interactive graphical summary (Figure 20), a detailed result in the form of a table (Figure 21 and a list of the sequence alignments found (Figure 22). As can be seen in all of them, the problem protein was the human class I histocompatibility antigen.

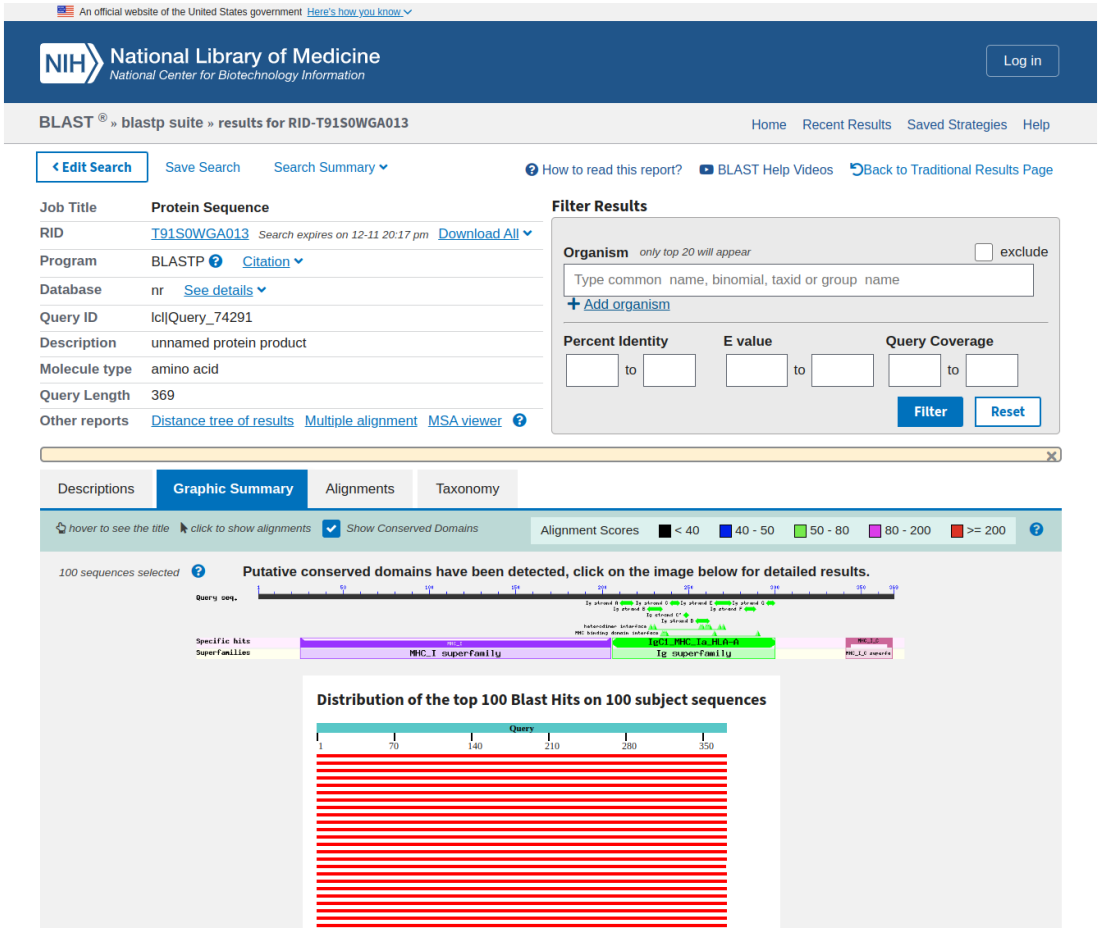



Figure 20. Graphical summary of the result of a BLASTp screen

The table showing the results (Figure 21) presents in the first column the accession number of each of the sequences in the database that show similarity (found by a sequence alignment algorithm) to the problem sequence. The accession number is also a link to the record storing the sequence in each case. The second column contains the sequence description. The following columns present the alignment score, the percentage overlap of the sequences and finally the probability E-value, which represents the probability that the similarity between the anonymous problem sequence and the one found in the database is random. Small values indicate that the similarity is not due to chance and, therefore, the sequences are related or, as in the case of the first sequence obtained (E = 0), they are the same sequence.

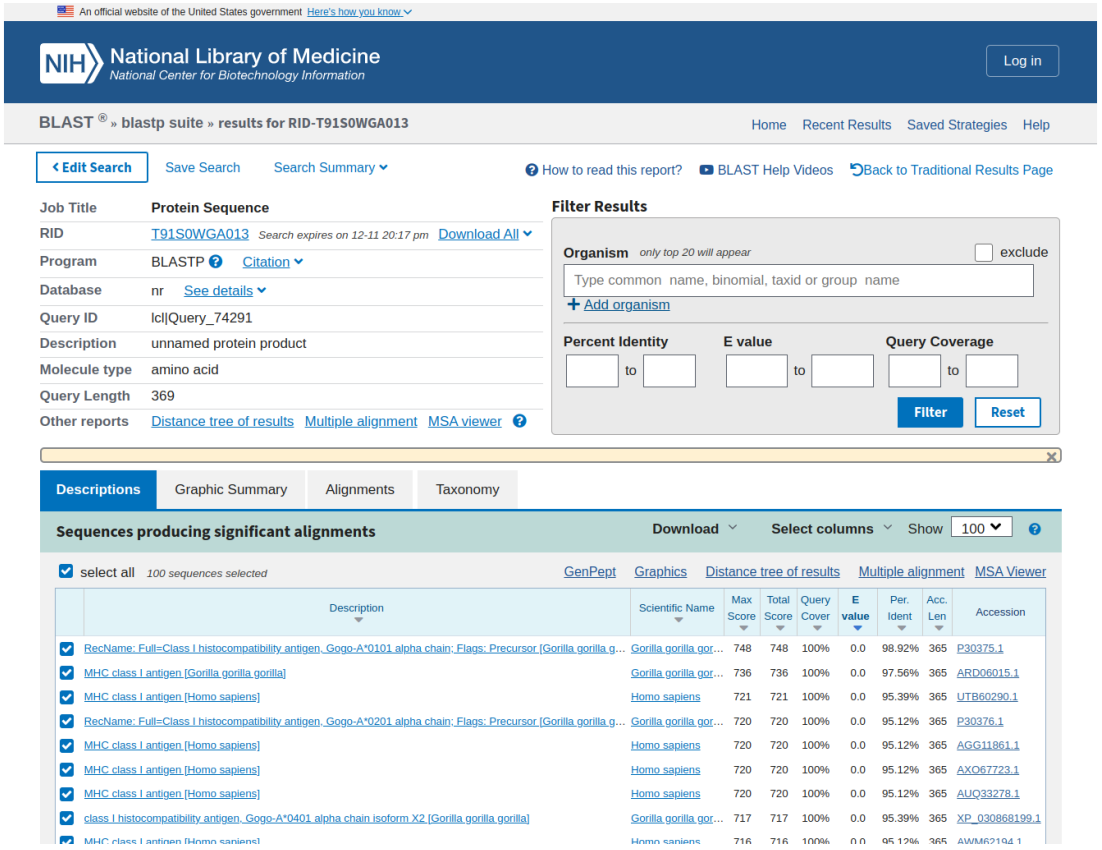



Figure 21. Detailed result of a BLASTp scan

Finally, the alignments of the problem sequence with each of the sequences obtained from the database are shown (Figure 22), in which the complete sequences can be observed and, between them, the consensus sequence. It is easy to observe the coincidences and differences between the aligned sequences.

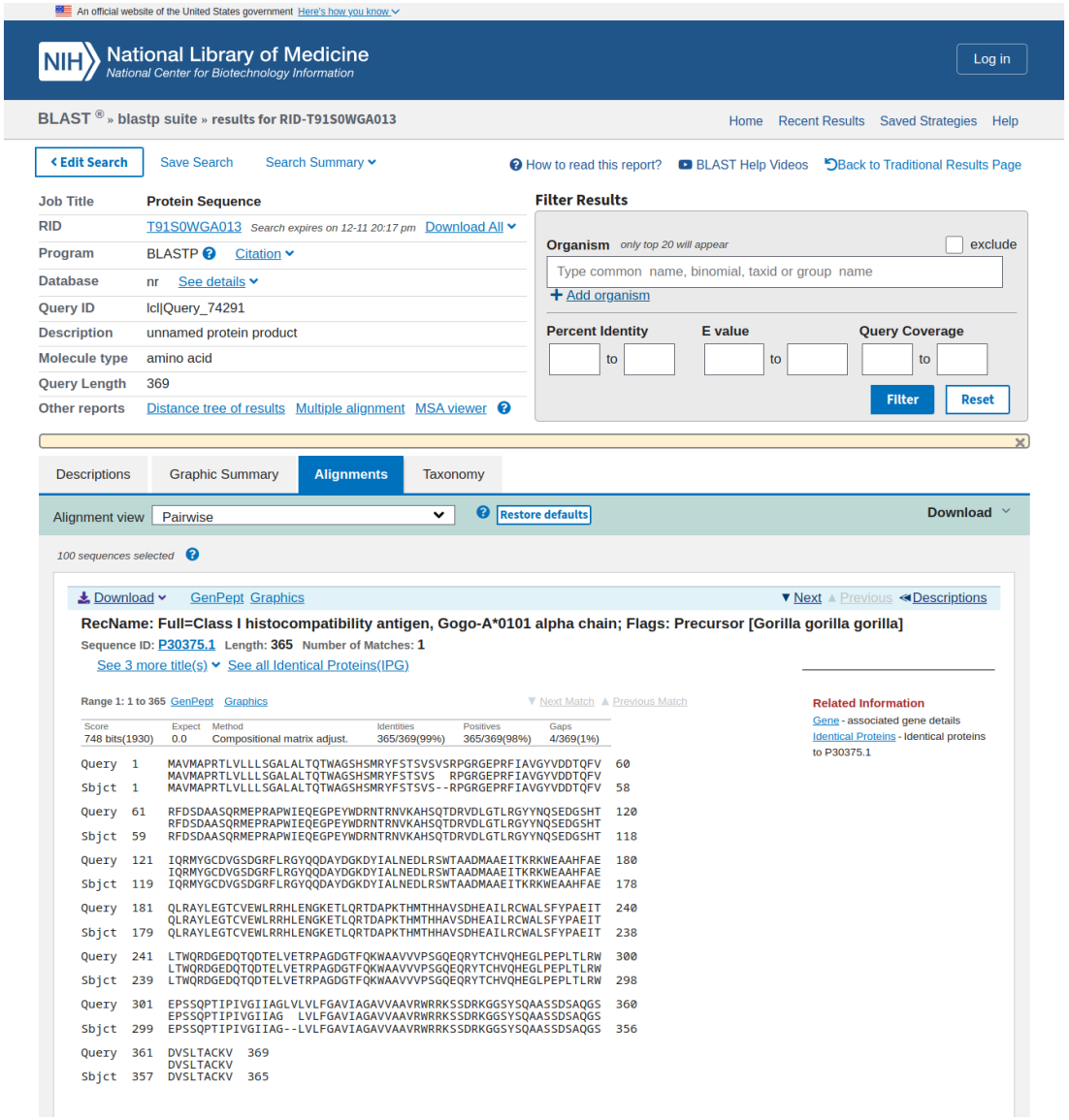



Figure 22. Alignment in a BLASTp scan

## BLAT

This program works in a similar way to BLAST (in fact it stands for Blast Like Alignment Tool) and is used to find an input sequence within a genome. BLAT reports all positions (chromosome, start and end) of the genome where the alignment of the input sequence gives a score and similarity above minimum thresholds. The BLAT output and its interpretation is explained in the figure 23:

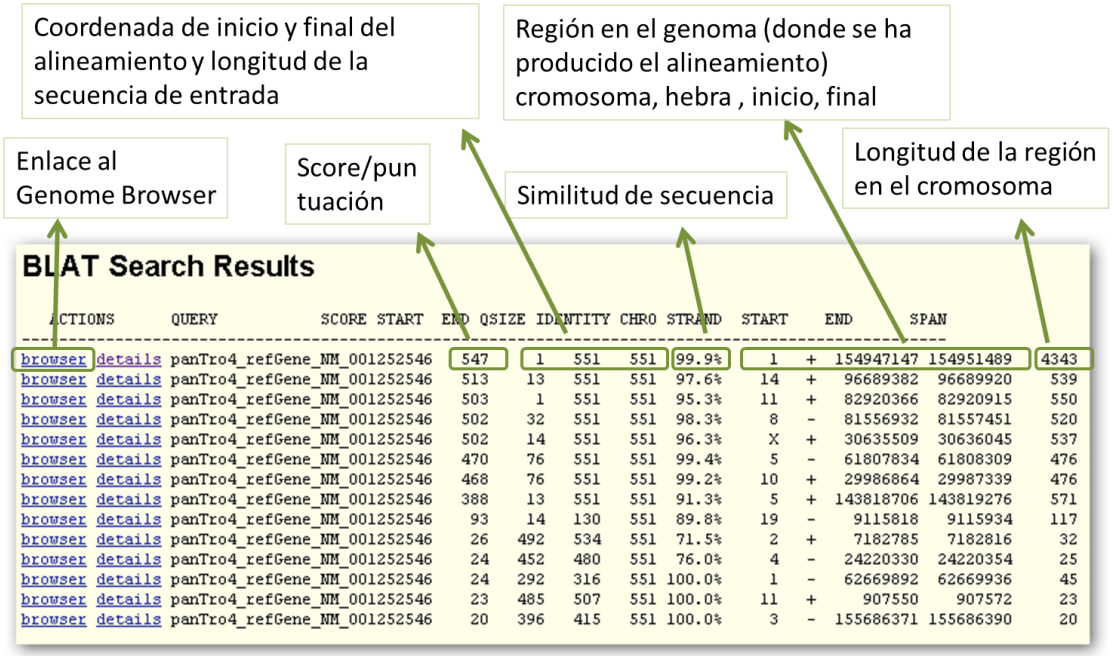



Figure 23. Tracing a sequence in a genome with BLAT

### 3.3.4 Genome browsers

A genome browser is a graphical representation of a genome, as can be deduced from the above. There are different genome browsers, but all of them allow visualization of annotations and other genomic features. In general, genome browsers are computer applications that can be stand-alone or operate through the internet, and that allow access to a large amount of information on genomes, such as identifying DNA sequences corresponding to specific genes within a complete genome (accessed through a specific database), identifying functional elements, carrying out comparisons between species, etc.

Some of the most widely used genome browsers are:

Apollo Genome Annotation Curation Tool (http://apollo.berkeleybop.org/current/index.html) This genome browser offers many possibilities, including the ability to annotate. It is Java-based, so it can be used on Windows, Mac OS X, or any Unix-based operating system.

Generic Genome Browser (GBrowse) (http://www.gmod.org/wiki/GBrowse) Developed by GMOD (http://www.gmod.org/wiki/Main_Page), it allows users to quickly configure a browser to their needs.

UCSC Genome Browser (http://genome.ucsc.edu/cgi-bin/hgGateway) Developed by the Genome Bioinformatics Group of UC Santa Cruz (University of California), it provides different genomes to analyze.

Ensembl (http://www.ensembl.org/index.html) Ensembl is a joint project between EMBL-EBI and the Wellcome Trust Sanger Institute, and provides access to different eukaryotic genomes for analysis.

## 3.4 EXERCISES AND QUESTIONS

The following are examples and exercises based on the BLAST and BLAT algorithms:

- Given a problem amino acid sequence, search for different proteins that have similarity (homology) to it.
- Given an anonymous DNA sequence, determine the type of sequence involved by searching a database.
- Locate a specific sequence in the human genome and summarize the information extracted from the analysis.

**BLAST:**

Example: Dicer protein sequence(isoform 2):

https://bioinfo5.ugr.es/GII/bioinfoI/Dicer_iso2.fa

Go to the BLAST page (https://blast.ncbi.nlm.nih.gov/Blast.cgi).

Before starting the analysis:

1. What type of sequence is involved?
2. What 'type' of BLAST should we use?
3. What are the differences between BLASTn and BLASTp?
4. Which database will we use?

Choose the BLAST program to use (BLASTn or BLASTp), paste the Dicer sequence (isoform2) in the text box and track the databases by clicking on BLAST at the bottom of the page on the left (see screenshot in figure 19). The following images are screenshots of the results (Figures 24 and 25):

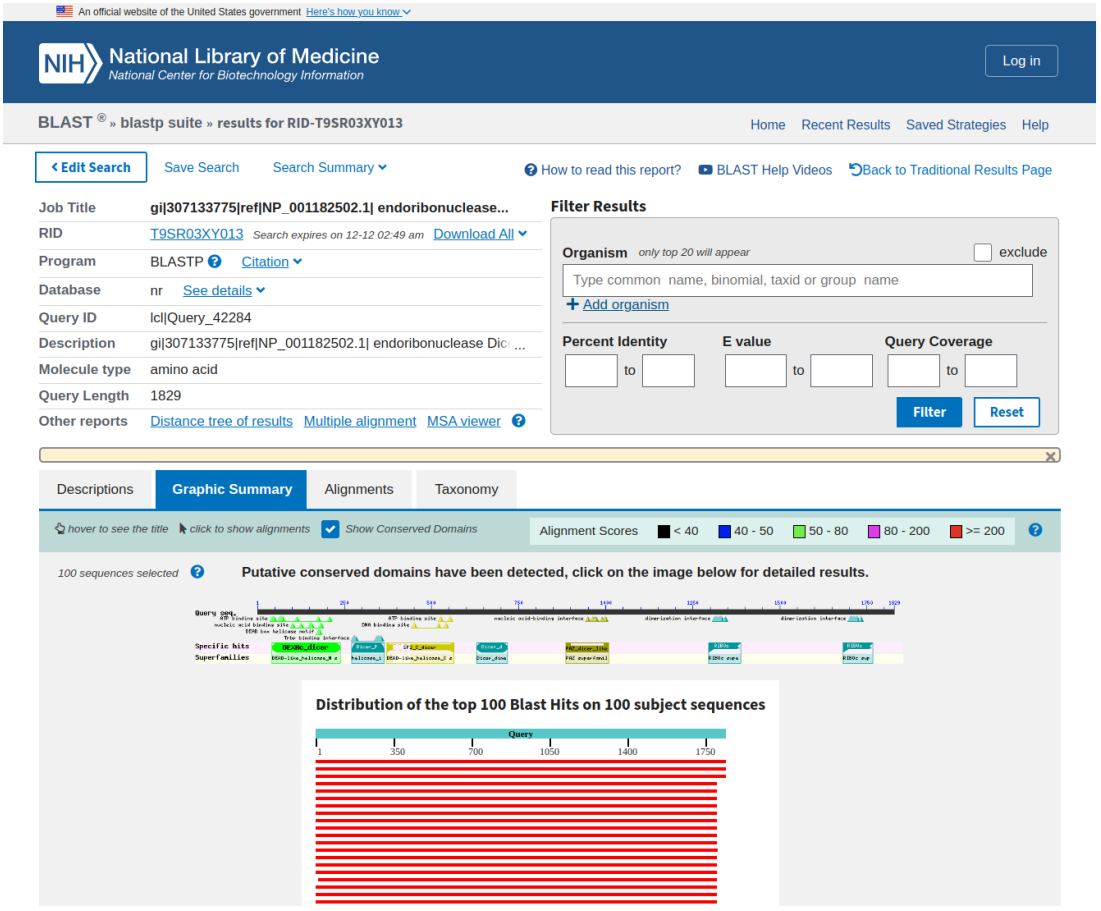



*Figure 24. Graphical summary*

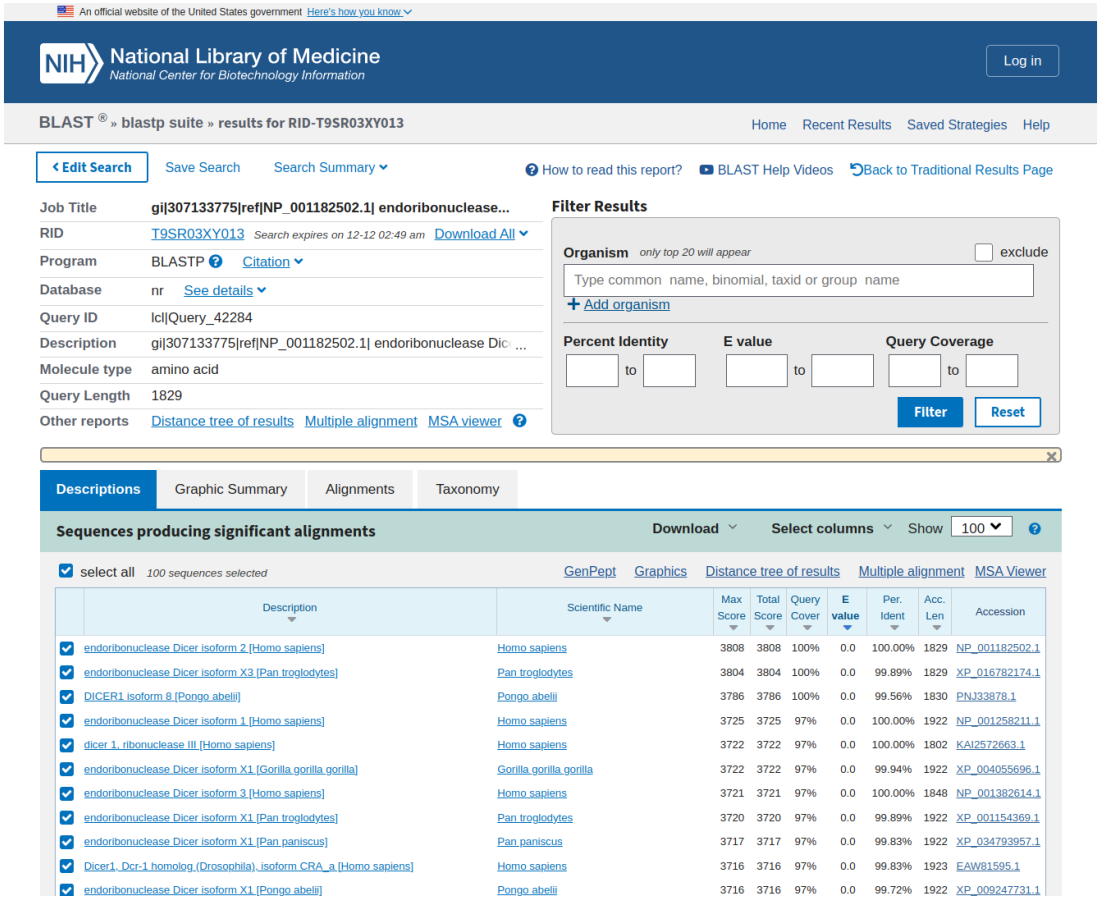



Figure 25. Descriptions

The analysis of the data yielded by BLAST can be summarized in the following points:

- Each line represents the alignment against a sequence in the database (exceeding certain thresholds).
- The first line represents the alignment of the input sequence against the same sequence in the database (Query Cover=100% and Ident=100%) and the length of NP_001182502 is identical to the length of the alignment.
- We observed alignments with a high Ident value against sequences from other species (they correspond to homologous sequences present in other species).
- We observe in the fourth line an alignment against a sequence in Homo sapiens with Ident = 100% and Query Cover = 97%: this alignment corresponds to another isoform, and therefore not all the input sequence has been aligned (only 97%).
- If we click on the accession number of the sequences (last column in Descriptions), we are taken to the NCBI page where we can consult the information about the sequences with which the analyzed sequence has identity. See the NCBI gene page for information on this gene: (http://www.ncbi.nlm.nih.gov/gene/?term=NP_001182502)

**Exercise: Sequence identification by comparison**

A molecular test (PCR) was performed on a patient with flu symptoms. The size of the amplified product was not as expected, so sequencing of the amplified DNA fragment has been carried out.

>anonSec2
CCCGAGGAGATCTGGAGCTGCTGGTGCGGCAGTAAAGGGAGTCGGAACGATGGTGATG
GAACTAATTCGG
ATGATAAAGCGAGGGATTAACGATCGGAATTTCTGGAGAGGTGAAAATGGGCGAAGAACA
AGAATTGCAT
ATGAGAGAATGTGCAACATCCTCAAAGGGAAATTCCAAACAGCAGCACAAAGAGCAATGA
TGGATCAGGT
ACGGGAAAGCAGAAATCCTGGGAATGCTGAGATTGAAGATCTCATATTTCTGGCACGGTC
TGCACTCATC
CTGAGAGGATCAGTGGCCCACAAGTCCTGCTTGCCTGCTTGTGTGTACGGGCTTGCCGT
GGCCAGTGGAT
ATGACTTTGAGAGAGAAGGGTACTCTCTGGTCGGGATTGATCCTTTCCGTCTGCTGCAAA
ACAGCCAGGT
CTTTAGTCTAATTAGACCAAATGAGAATCCAGCACATAAAAGTCAATTGGTGTGGATGGCAT
GCCATTCT

This sequence can be downloaded from:

https://bioinfo5.ugr.es/GII/bioinfoI/anonSec2.fa

a) From which organism does this sequence come from?
b) Is the sequence obtained from the patient contained in the database?
c) Is alignment a property that two sequences have?
d) How do we define the similarity between two sequences?
e) Is similarity between two sequences the same as homology between two sequences?
f) What is the alignment score?
g) What is the significance of the e value reported by BLAST?
h) Why do we often observe more than one result in a BLAST search?

BLAT:

**Example: Localization of an active Alu sequence in the human genome**

Alu retrotransposons represent the most frequent transposable element in the human (and most primate) genome. There are more than 1.1 million copies in the human genome, which represents 11% of the entire genome. Although they had the highest amplification rate more than 30 million years ago, there are still active Alus in the human genome that produce insertional polymorphisms (related in some cases to certain diseases).

Go to the BLAT page ( https://genome.ucsc.edu/cgi-bin/hgBlat) (Figure 26).

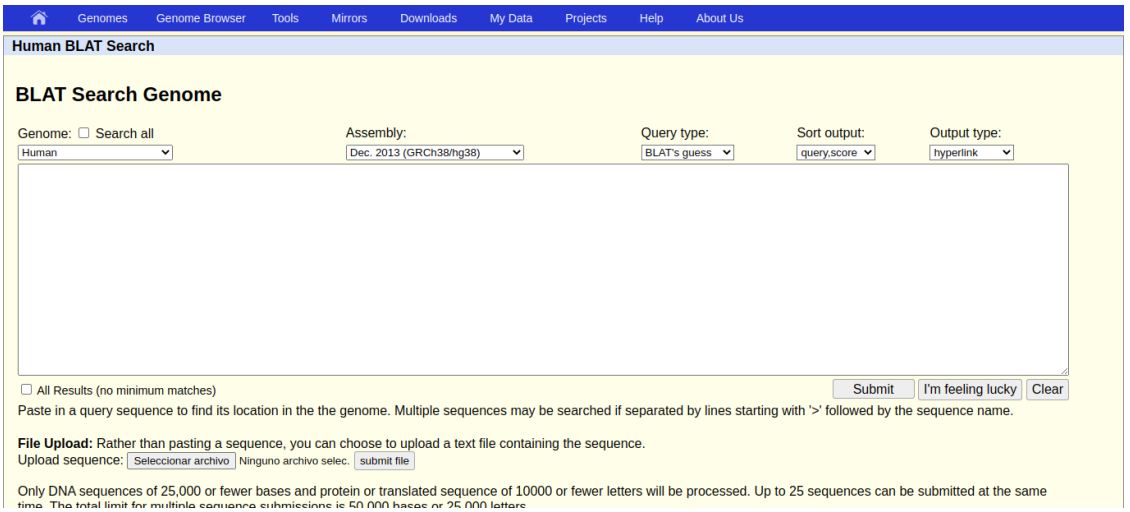



Figure 26. BLAT main page

We will use the BLAT program to locate the insertion positions of this active Alu (https://bioinfo5.ugr.es/GII/bioinfoI/AluYa5_active.fa) in the hg19 assembly (*Homo sapiens*). To do this, paste the sequence of the Alu in the text box and trace the genome by clicking on Submit under the box with the sequence. The following image is a screenshot of the results (Figure 27):

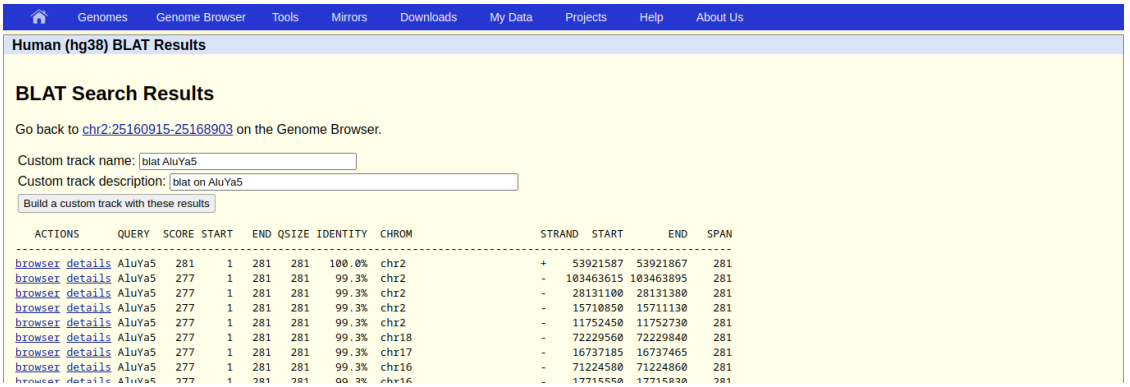



Figure 27. BLAT search results

The analysis of the BLAST data can be summarized as follows:

- Each line corresponds to a position in the genome where the Alu sequence has aligned with higher sequence similarity than a given threshold value.
- Despite aligning at different loci, the active Alu sequence is found at a single locus (only the first line represents an alignment where the entire Alu sequence, 281bp, shows 100% similarity to the sequence present at that position or locus).
- The remaining positions at which the Alu sequence has aligned correspond to Alu sequences that have accumulated mutations over generations and thus differ at the sequence level.

**Exercise: Insertion polymorphism by Alu sequence**

The locus between coordinates 27144072-27144384 on chromosome 12 (chr12) corresponds to an AluYb9. The insertion of this Alu element is polymorphic in the human population

(insertion polymorphism). To detect whether a given individual has this Alu insertion, the region can be amplified by PCR. We have the sequences of two candidate primers:

5' primer: GCAGACAGTACCCACTTATTTTTGT

3' primer: GAAGAAACAAATGCTTTATAGAACCA

a) Are these two primers suitable to amplify a region containing this AluYb9?

b) If not, which pair of primers could we use?

c) What will be the length of the PCR product amplified by the appropriate primer pair?

d) What will be the size of the products amplified from the DNA of an individual heterozygous for the Alu insertion?

e) Draw a schematic representing the PCR products obtained for the possible genotypes for this insertion polymorphism.

f) Which biological questions can be answered by BLAST and BLAT?

**Notes**

- Primer sequences should be pasted into the BLAT text box in FASTA format:

  *>5′primer*

  GCAGACAGTACCCACTTATTTTTGT
  *>3′primer*:
  GAAGAAACAAATGCTTTATAGAACCA

- Auxiliary programs: EMBOSS revseq:

http://emboss.bioinformatics.nl/cgi-bin/emboss/revseq

# 4.- COMPUTATIONAL GENE PREDICTION

## 4.1. AIM

Genetic information is currently undergoing an enormous increase , mainly due to the fact that many whole genome sequencing projects have been completed. Once a genome sequence is available, the recognition of protein coding regions is of paramount importance. For this purpose, several computer programs have been developed, which predict the number and location of genes, including the exact location of exons and introns (in eukaryotes) from an uncharacterized sequence. The aim of this computational practical is to provide an introduction to the knowledge and use of this type of program.

## 4.2. THEORETICAL BASIS

### 4.2.1. Web resources

Bioinformatics programs are mostly run on command line using a UNIX environment. However, a large variety of interfaces have been developed to facilitate their use. Some of these interfaces consist of web pages that collect data provided by the user and return the results by the program via web or e-mail.

A software list with links to the original pages of each application is available via the following link:

https://www.sanger.ac.uk/science/tools/#

https://bioinformaticshome.com/db/

Other software directly related to computational gene prediction can be used and downloaded from the following link:

http://opal.biology.gatech.edu/GeneMark/

In this link you can find different versions of the GeneMark application for gene prediction in prokaryotes, eukaryotes or a "self-training" version. Additionally, it contains links for free download programs for non-commercial use with a two-year renewable license.

## 4.3. METHODOLOGY

### 4.3.1. Sequence search and analysis (A)

Firstly, we will obtain a DNA sequence on which we can use computational gene prediction programs. The chosen sequence is a fragment of the human Y chromosome, between nucleotides 2,784,990 and 2,789,726 (assembly GRCh38/hg38). We will obtain this sequence from the Ensembl database.

Ensembl is a database that contains the genomes of a multitude of organisms, which are annotated using different programs that localize a variety of characteristics in the sequences, including gene, exon and intron location. The database is accessible online or via programs that access the database remotely using libraries written in perl language.

To get our sequence, we will first go to the home page of Ensembl: http://www.ensembl.org

In the drop-down menu at the top of the page we choose "Human", and in the search text field we will enter the chromosome of interest and the start and the end nucleotides as follows:

Y:2784990-2789726

As shown in the figure below:


Figure 1: Ensembl database sequence search.

After clicking on the "Go" button, a representation of the chosen genome region will appear. We want the nucleotide sequence of that region, so we will click on "Export data" button, located on the left.


Figure 2: Ensembl database sequence export.

A new window will appear in which you can choose different options regarding data exportation. The default option is to export the sequence in FASTA format, which is precisely what we need, so we only have to click on the "Next" button. We can then download the sequence in different formats. We will choose "Text" to obtain the sequence. Afterwards, we can archive the sequence, copy and paste it in a text editor or in the web interface of the other program.

### 4.3.2. ORF prediction

Open Reading Frames (ORF), consist of a sequence fragment starting in a start codon and ending in a stop codon. If the distance between the two codons is large enough, these ORFs could be indicative of the presence of a coding region. We will search for ORFs with the ORFfinder program at:

https://www.ncbi.nlm.nih.gov/orffinder/

By pasting the sequence into the text box:

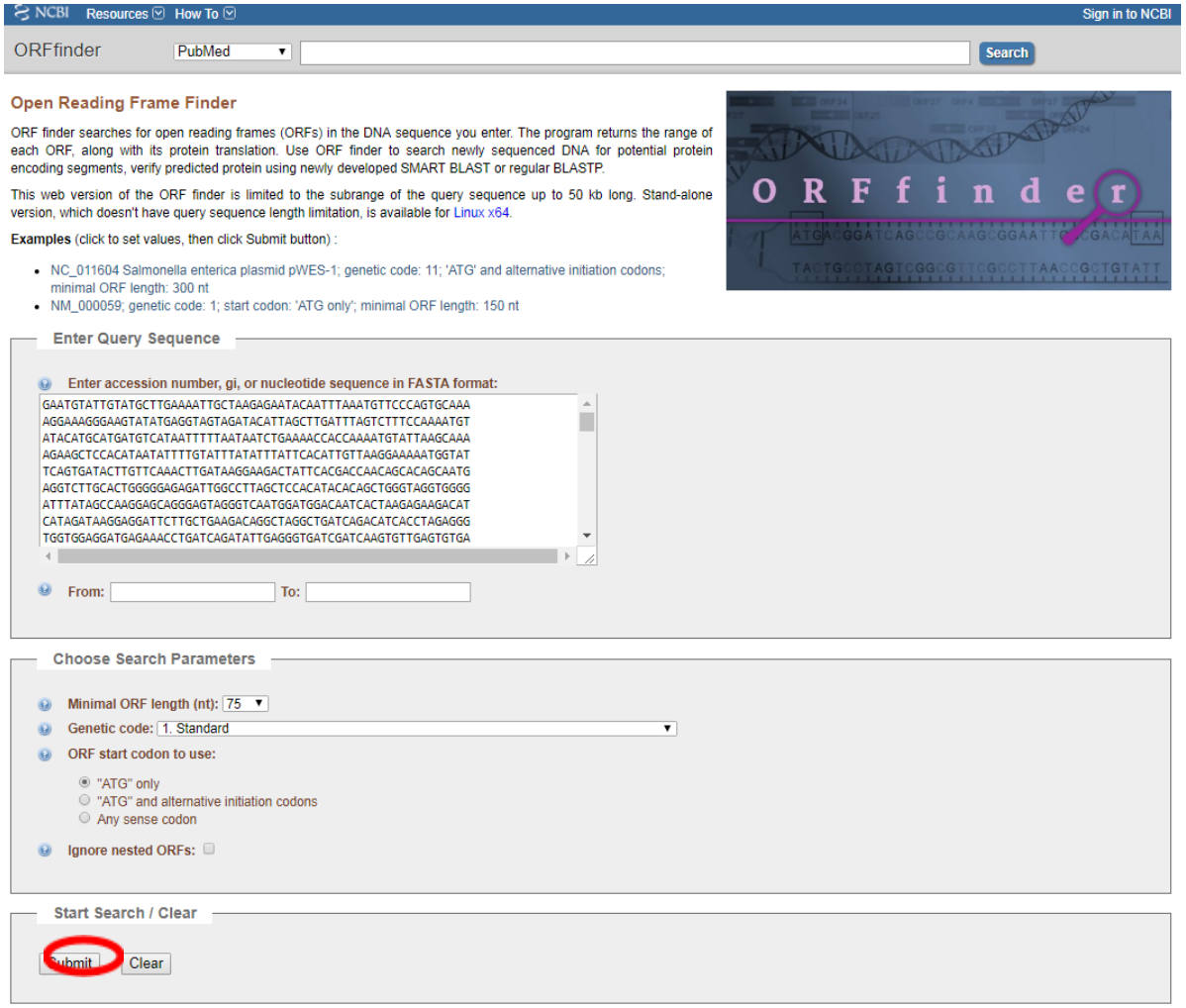



Figure 3: ORFfinder interface.

After clicking on the "Submit" button, the program will search the ORFs and display the result as follows:

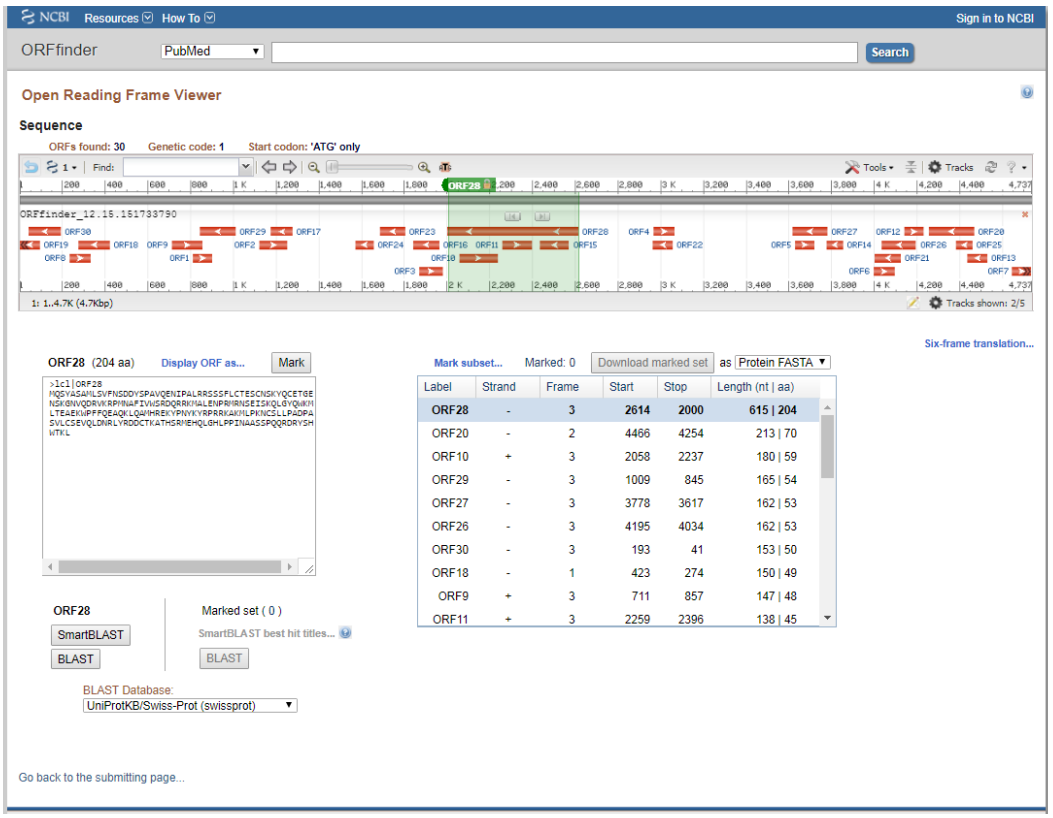



Figure 4: ORFfinder results.

In the previous image, we can see a graphical representation of all the possible ORFs marked in red and named in blue. Below the graph, in the box on the right, for each possible ORF we can find the strand (+ or -), the reading phase (1, 2 or 3), the nucleotides where they begin and end, and their total length in nucleotides | amino acids. The ORFs are ordered by length. The ORF with the longest length, which could be indicative of a coding region, is marked in both the graph and the box. Below the graph, the box on the left shows the amino acid sequence of the potential coding region (the possible translation of that nucleotide fragment or potential ORF).

Then, we should see if this reading frame belongs to any known protein. Using this same website it is possible to perform a BLASTp search through the "nr" (non-redundant) database, which contains all known sequences without redundant data. In this step we have to select the option in the drop-down menu below the box with the sequence and click on the "BLAST" button (shown in the figure below). When we click on this button, the data is sent to the NCBI (National Center for Biotechnology Information). We can see now the sent data and the chosen options for the search.

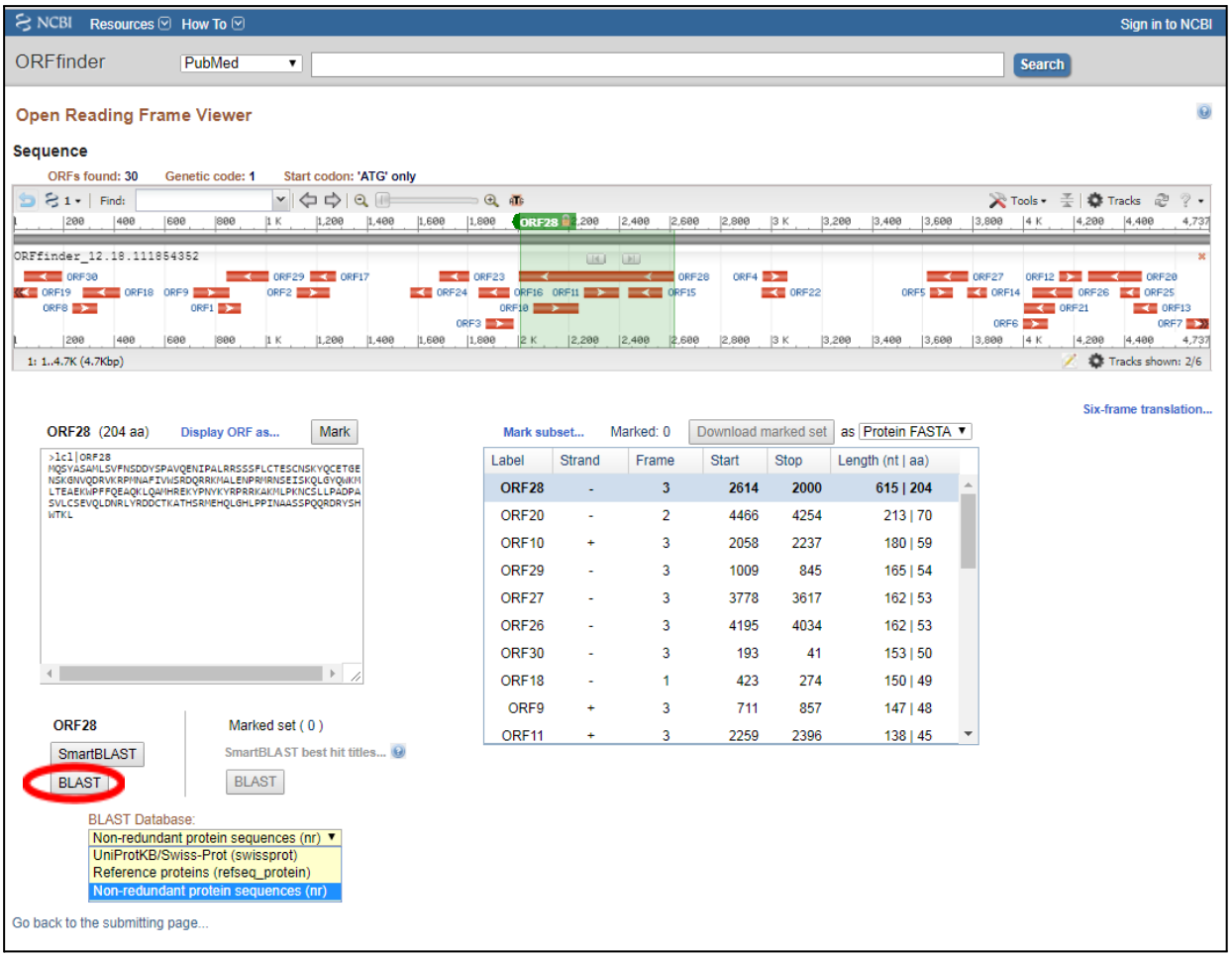



Figure 5: ORFfinder BLAST link.

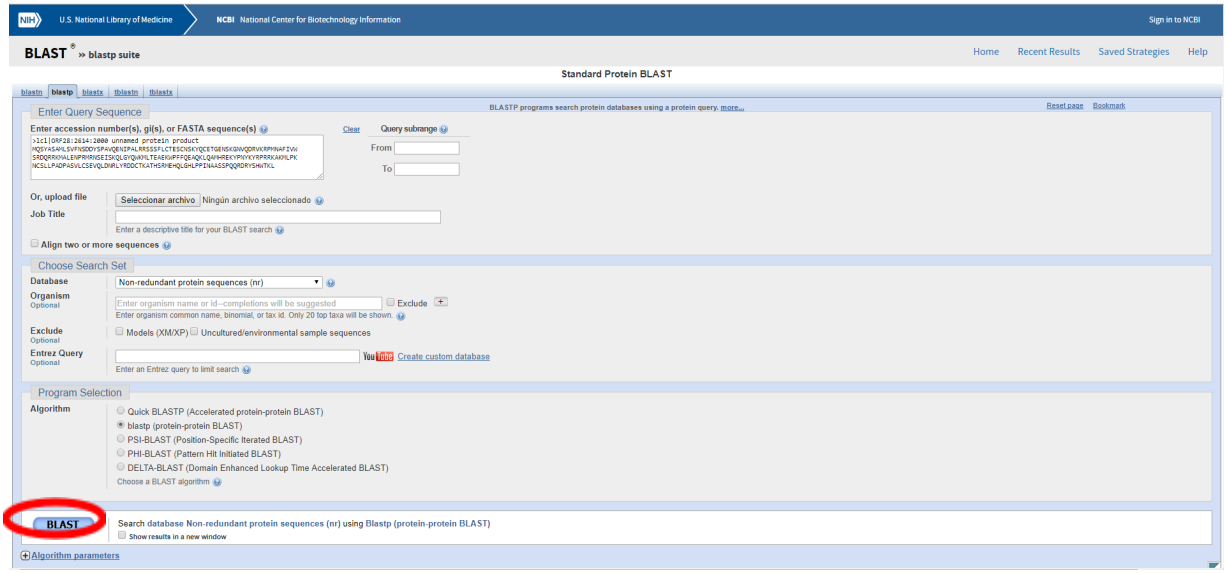



Figure 6: BLASTp interface.

Once on the NCBI page, click again on the "BLAST" button. A part of the BLAST results are shown below:

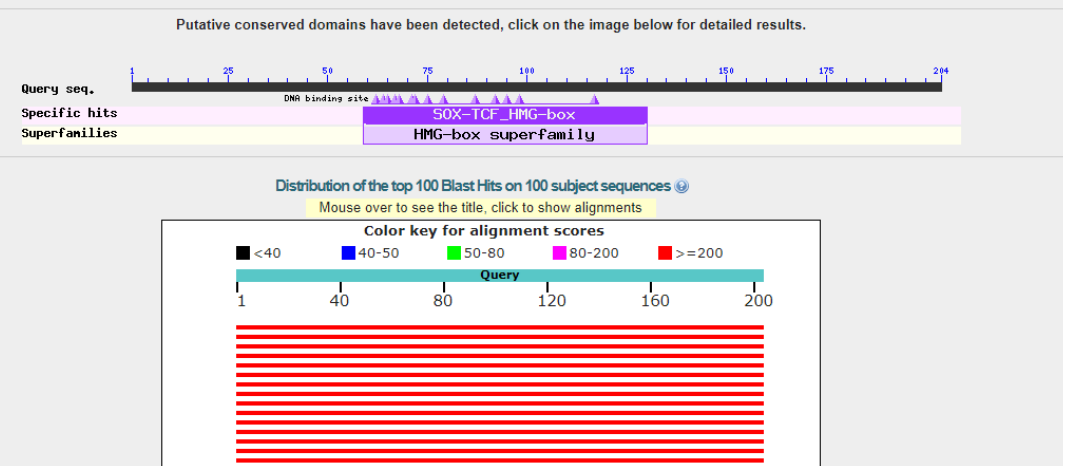



Figure 7: BLASTp graphical results.

In the upper part we can see that a conserved HMG-type domain has been located. The lower lines represent the found sequences with homology with the query sequence, as explained in the previous computational practical. By placing the cursor on the first red line, the text box above the lines will display information about the sequence that this line represents. In this case it is:

Sex-Determining Region Y protein [Homo sapiens] Score=428, E=4e$^{-152}$

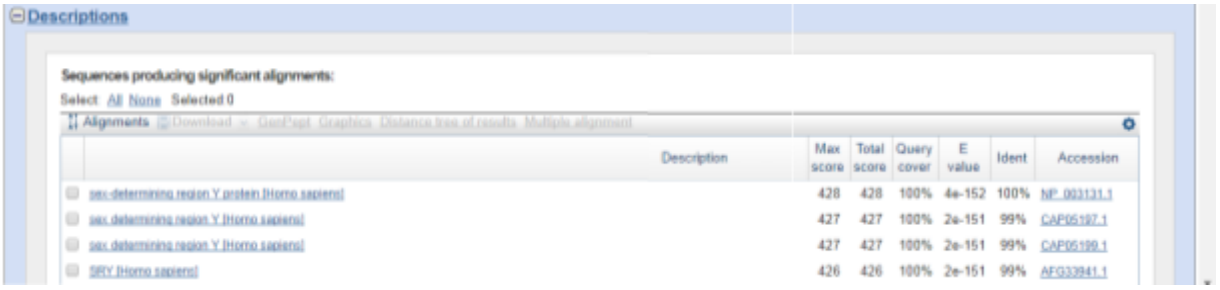



Figure 8: BLASTp result table result.

Here, as also explained in the previous computational practical, the column **'Score'** shows the score obtained in the alignment and '**E value'** is the "Expect" value, i.e. the expected number of random matches in the database.

If you click on the HMG box at the top, you will access a page with additional information about this type of domain:

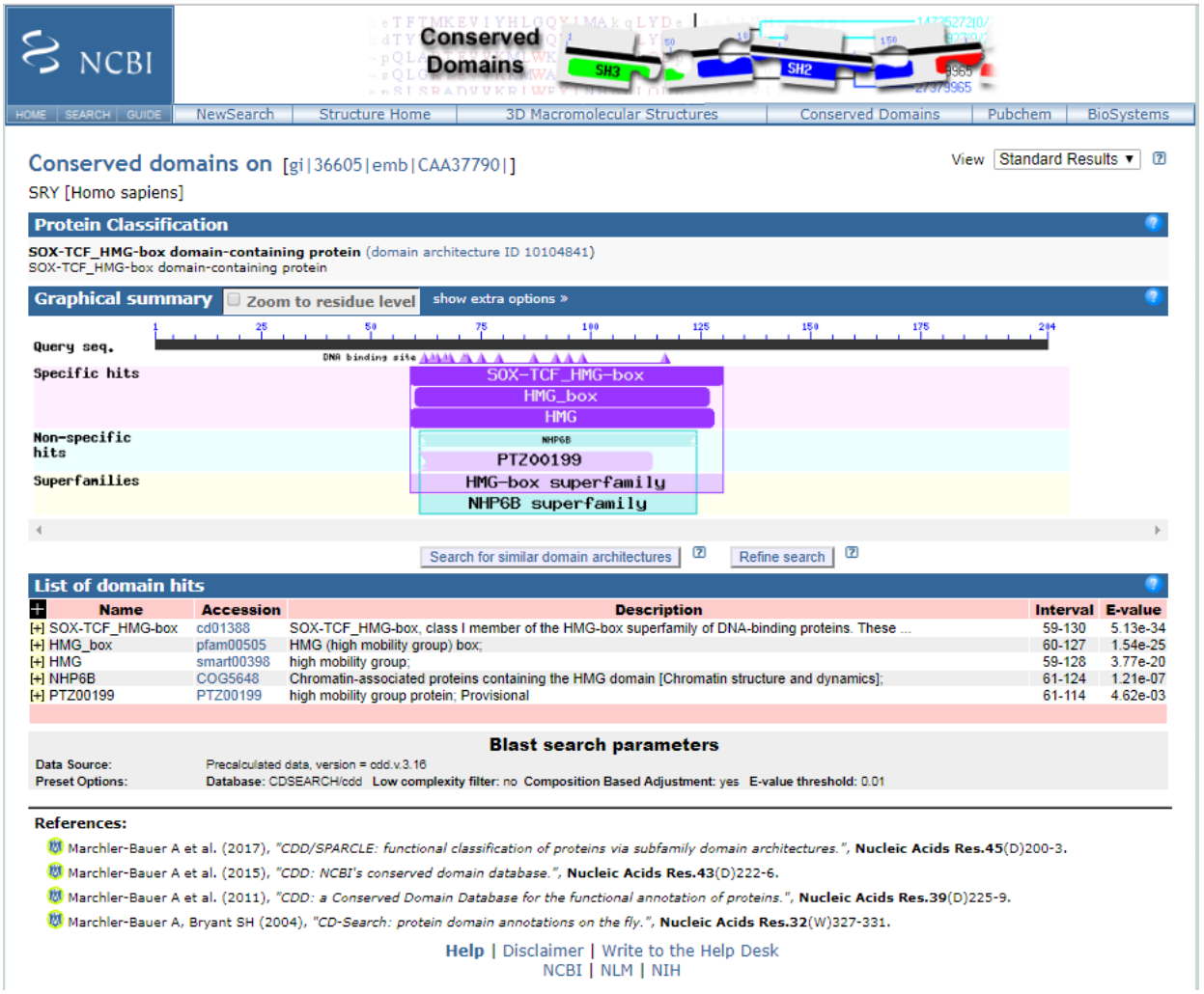



Figure 9: BLASTp graphical result detail.

There is another option available in ORFfinder for BLAST, called SmartBLAST. This option allows a BLAST analysis to be performed in a slightly different way. SmartBLAST processes the sequence search by presenting a summary of the five characterized proteins in different species (whenever possible) included in the reference database that show homology to the problem sequence. If SmartBLAST cannot find five matches in the reference database it will use matches from the non-redundant (nr) database. SmartBLAST obtains these results using a combination of an optimized BLASTp search, a new implementation of BLAST that aims to find closely related matches and to generate a multiple alignment. In addition, SmartBLAST presents the matches found for the “problem sequence” in the conserved domain database. Additional matches to the nr database are presented after the first five matches.

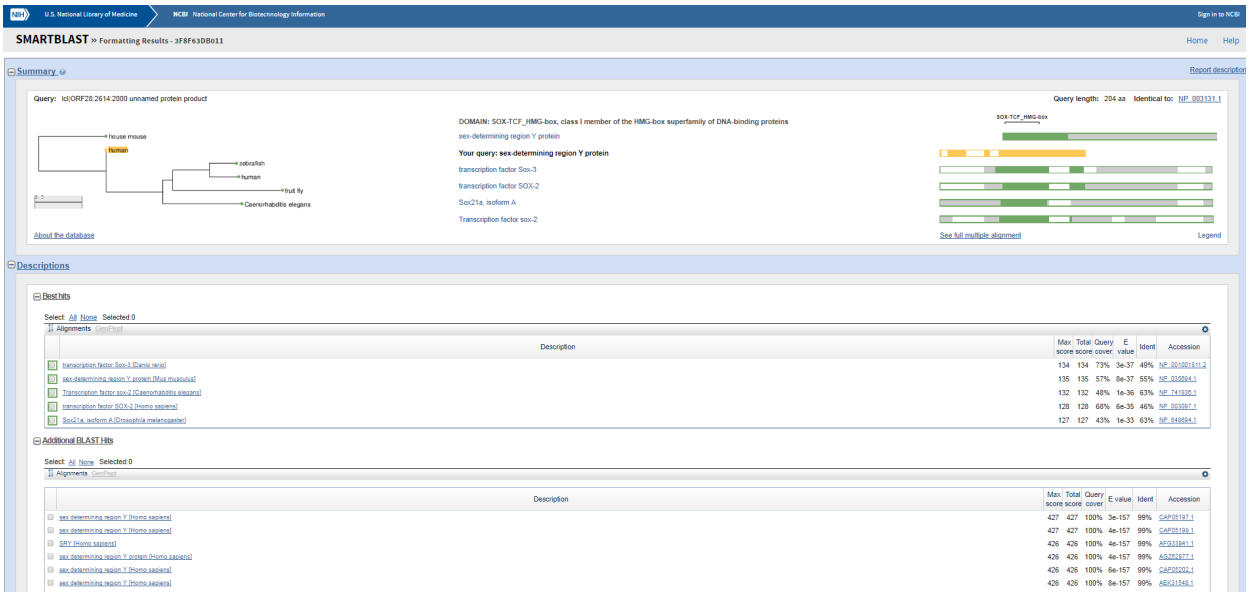



Figure 10: SmartBLAST results.

### 4.3.3. Sequence search and analysis (B)

We previously retrieved a sequence of the human Y chromosome. Similarly, we will now retrieve the sequence corresponding to nucleotides 72,120,505 to 72,127,125 of chromosome 17 (GRCh38/hg38 assembly). To do this, follow the steps described above for the SRY gene sequence (At http://www.ensembl.org, Human, search for 17:72,120,505-72,127,125, Export data, Next, Text).

We will analyze the sequence in ORFfinder. The result obtained is slightly unclear:

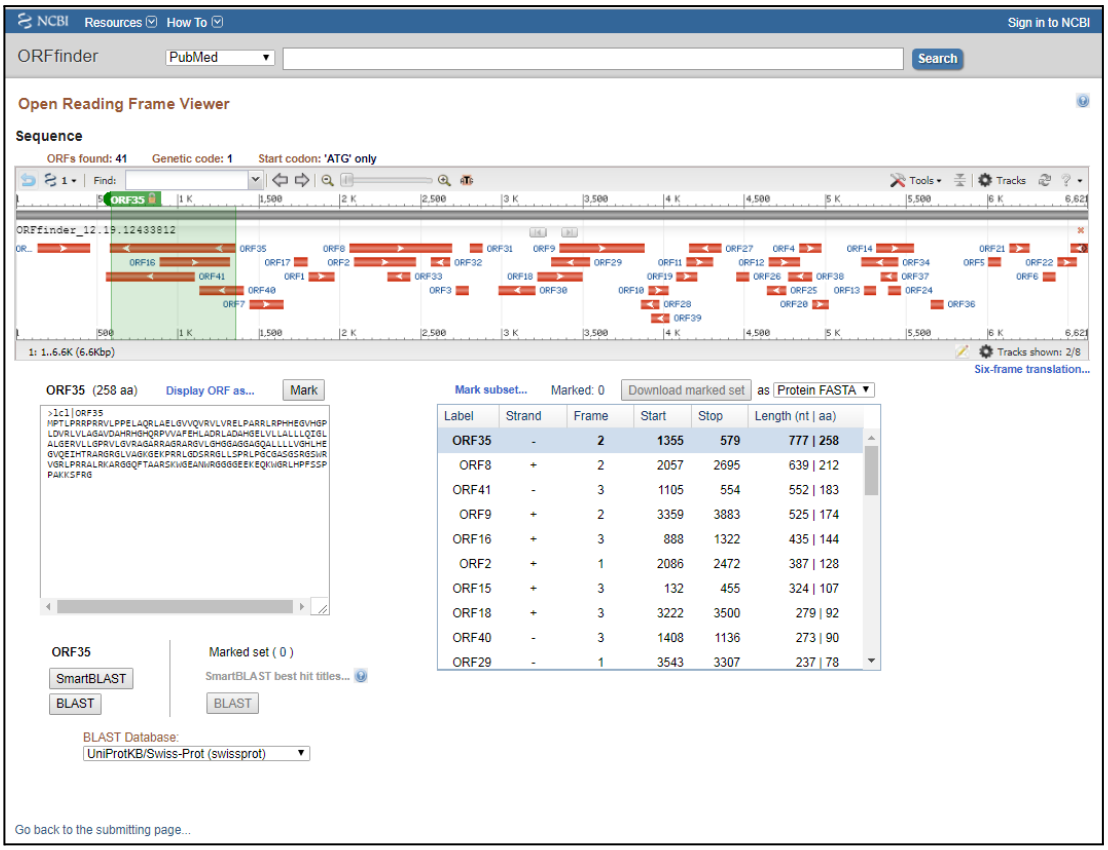



Figure 11: ORFfinder results.

Many open reading frames are observed. However, there are none that clearly differ from all the others in size. We already know that this region of chromosome 17 contains a gene, so the efficiency of this method for locating genes may be questioned. The problem lies in the fact that the gene contained in this portion of DNA has several introns that interrupt the open reading frame. Considering that most eukaryotic genes are interrupted by introns, this really implies a problem for the correct estimation of gene location when based only on the presence of open reading frames.

Therefore, it is necessary to estimate the position of the possible intron beginnings and ends present in the sequence, called "donor" and "acceptor" sites, respectively. For this purpose, we will analyze the sequence with the NetGene2 program:

https://services.healthtech.dtu.dk/services/NetGene2-2.42/ç

Once in the NetGene2 page, we paste our sequence in the text box below and click on the “Send file” button:

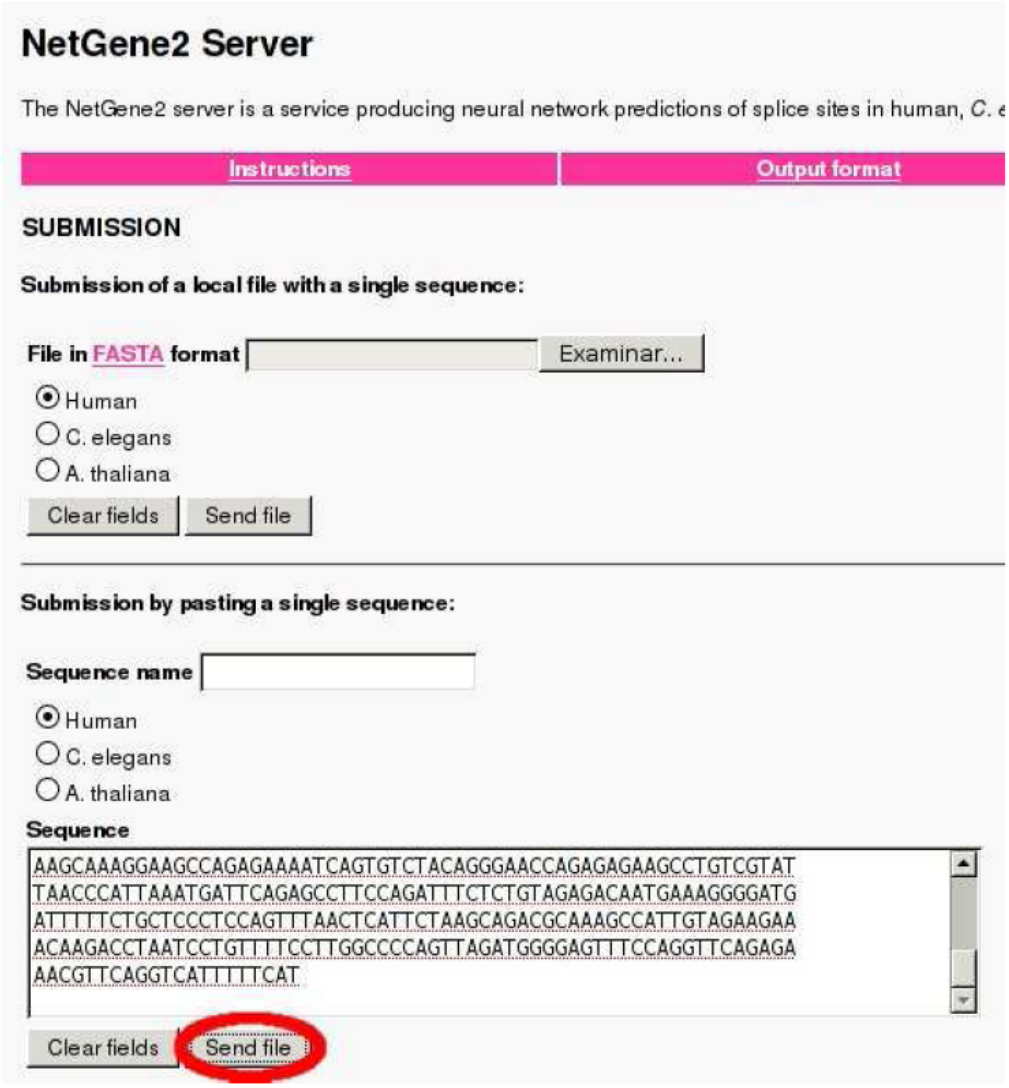



Figure 12: NetGene2 Server interface.

We will now observe a page that automatically refreshes at regular intervals until the task is completed, which is when the results will appear. Some of the results obtained with the problem sequence are shown below:

```
********************* NetGene2 v. 2.4 ************************


The sequence: sequence1 has the following composition:

Length: 6621 nucleotides.
23.8% A, 27.3% C, 25.2% G, 23.7% T, 0.0% X, 52.6% G+C

Donor splice sites, direct strand
---------------------------------
              pos 5'->3'  phase strand  confidence  5'      exon intron    3'
                  474       0      +        0.79       TGGCTCTAAG^GTGAGGCGGA
                 1164       0      +        0.34       GCCCATGCCG^GTGCGCGTCA
                 1319       2      +        0.95       AGCTCTGGAG^GTAGGACCCG H
                 1532       0      +        0.63       GAGGGGGGTG^GTAAGTGGAA
                 1545       1      +        0.50       AGTGGAAGAG^GTGAGGGAGG
                 1800       2      +        0.32       CTGGAAATAG^GTGGGAGTGT
                 1894       1      +        0.49       GTTGGGGGCG^GTAAGTCGAG
                 2011       1      +        0.35       GACCGCTCAG^GTCAGACTGC
                 2469       1      +        1.00       GAGCACTCGG^GTGAGTCGCC H
                 4682       0      +        0.37       GAAGCATTTG^GTAAGCTTTA
                 4820       0      +        0.62       TAAGAAAGAG^GTAAAAGGCA
                 5114       2      +        0.35       TCCTCAAAGG^GTATGGTCAT
```

Figure 13: NetGene2 result table.

In this output we can observe the possible donor sites (exon/intron cut-off). The columns, from left to right, show: the position of the exon/intron cut-off point, the phase in which it is located, the strand, confidence level and sequence. Confidence levels close to 1 may indicate functional sites. Similarly, possible acceptor sites are shown on the page. Finally, a graphical representation of the results is shown:

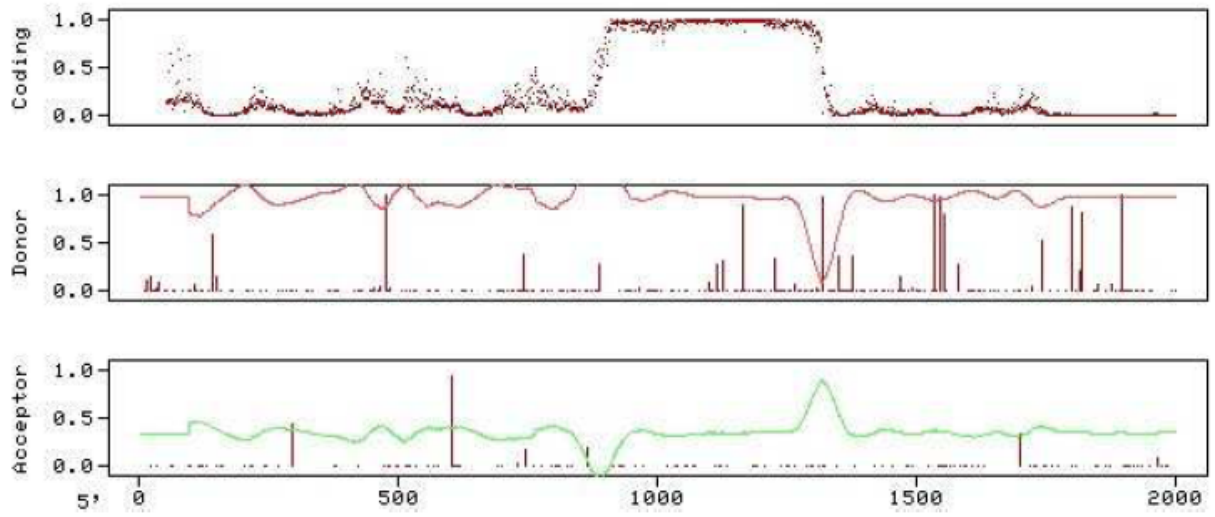



Figure 14: NetGene2 graphical results.

The three graphs correspond, from top to bottom, to the coding potential, the location of donor sites and the acceptor sites. The vertical lines correspond to the possible donor and acceptor sites. The length of the lines corresponds to the confidence levels.

The curves observed in the second and third graphs are derived from the changes in slope of the coding potential curve. To identify donor or acceptor sites with real biological potential, their positions should coincide with the boundary of the potential coding regions. Thus, the boundary between exons/introns and between introns/exons should coincide with vertical lines of a length close to 1. In addition, we should observe significant dips in the respective curves that, in turn, coincide with slope changes in the coding potential curve.

The first potential donor site (exon/intron) corresponds to position 1319, and the next position that could act as an acceptor (intron/exon) is 2214. These two points would correspond to a first intron, so the coding sequence of the potential gene should start before position 1319 and end in the vicinity thereof.

Looking at the ORFfinder results, we see that the second open reading frame in phase 3 ends at nucleotide 1322. If we carry out a BLAST from ORFfinder with the amino acid sequence encoded by this frame, we obtain the following:

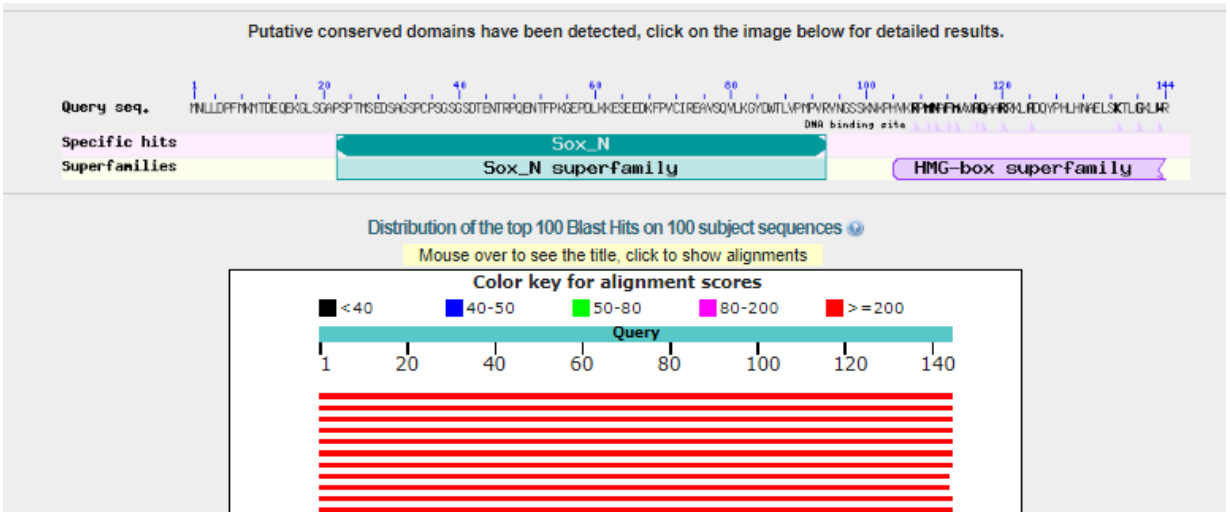



Figure 15: BLASTp graphical results.

Indeed, it corresponds to conserved domains. It is a fragment of the SOX9 gene, we can conclude this because the first matches correspond to this gene in different species. In fact, one of them corresponds to the human SOX9 gene.

The next exon should start after position 2214, where the first donor site is located. Read phase 1 shows an ORF between nucleotides 2086-2472, and phase 2 between 2057-2695. A BLAST with the first did not show significant results, but the second shows the end of a conserved HMG box.

Therefore, we conclude that the ORF of phase 3 and the following one of phase 2 correspond to two exons of the same gene, interrupted by an intron located in the

middle of the region, which encode a conserved HMG-type domain. Following this strategy, we are able to locate the rest of the sequences that correspond to SOX9 exons.

### 4.3.4. Intron location through “dot plot”

We will simulate a laboratory experiment in which the mRNA of a gene of interest would be isolated; however, a part or all of the gene sequence should previously be known. Subsequently, the sequence of this mRNA would be obtained and compared to the genomic sequence of the same gene, highlighting the regions corresponding to exons and introns. Instead of obtaining the messenger sequence in the laboratory, as we actually know the gene we will obtain it from a database. To perform this simulation, we will go to the page:

http://www.ncbi.nlm.nih.gov/

In the upper part, we will select the nucleotide database. In the text line, we will write the keywords "sox9 mrna homo sapiens" and we will click on the button “Search”:

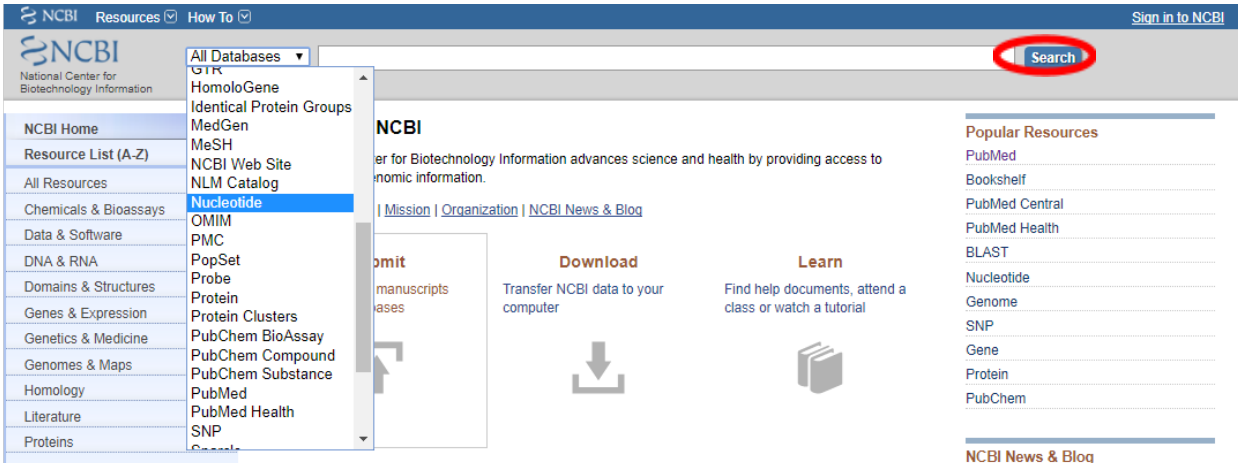



Figure 16: NCBI nucleotide search interface.

In the obtained result we will observe:

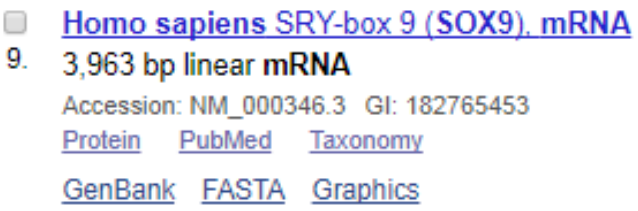



Figure 17: NCBI nucleotide search result.

By clicking on the gene description, we can retrieve the messenger sequence. A dot plot analysis can be created at:

http://emboss.toulouse.inra.fr/cgi-bin/emboss/dottup

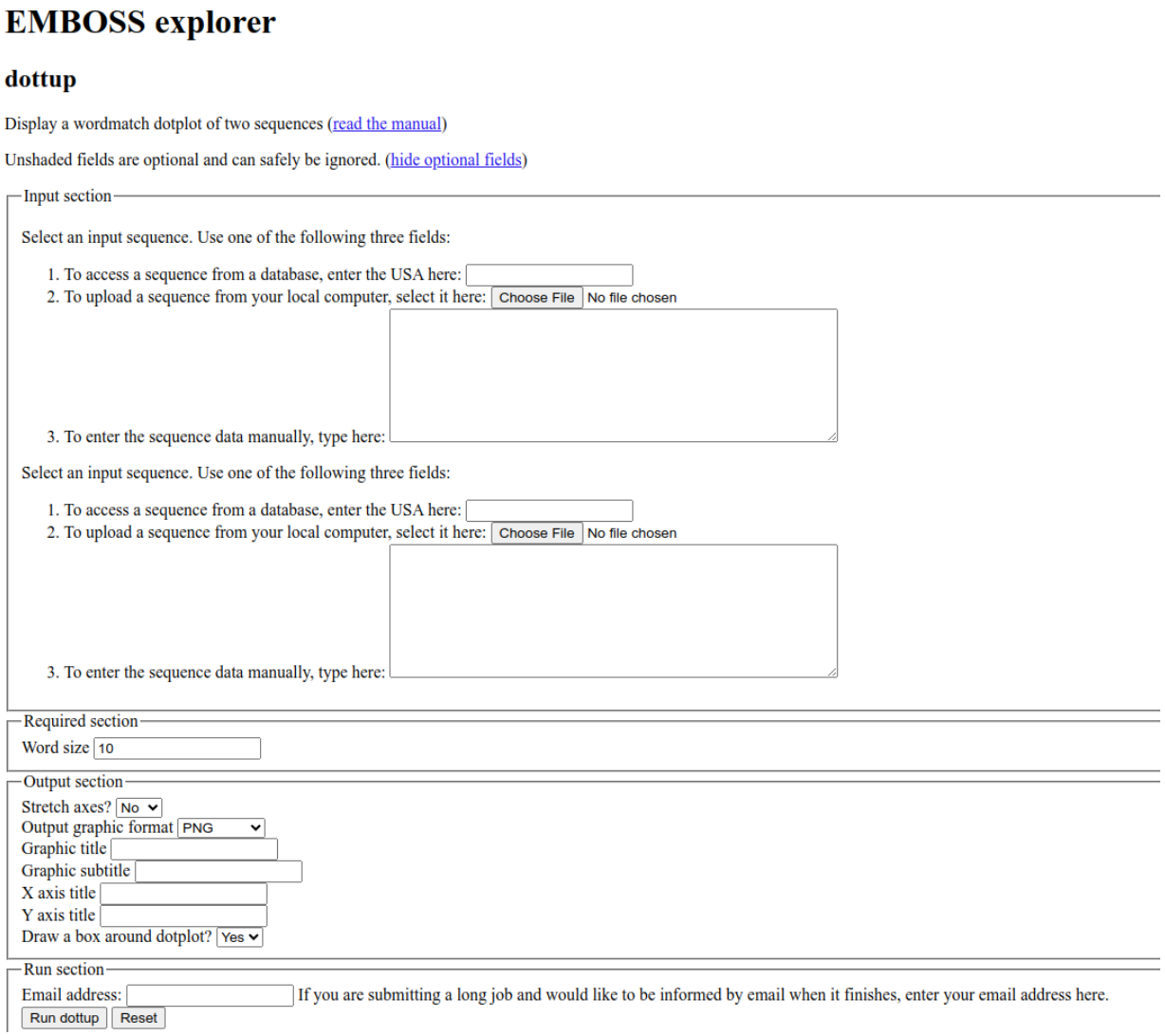



Figure 18: Dottup interface.

The web will show us the introns and exons position when we compare both sequences (we can paste each sequence in a box or select the files in each case):

**OUTPUT FILE** .stdout

```
Created dottup.1.png
```

**IMAGE FILE** dottup.1.png

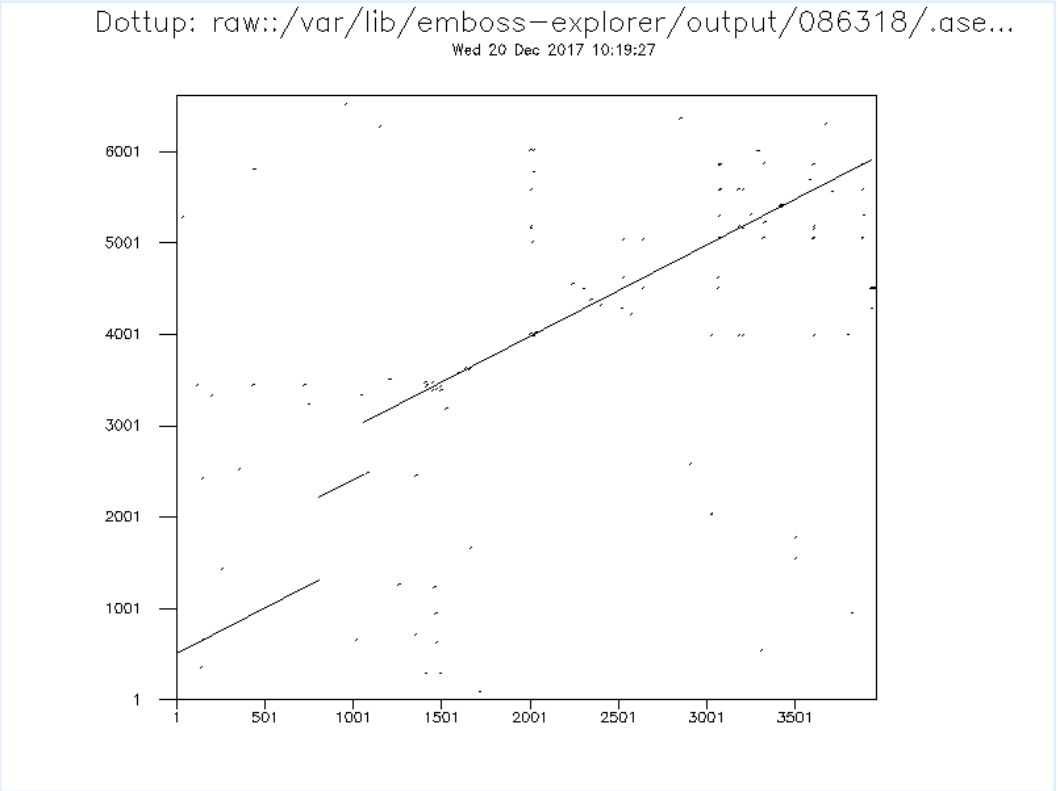



Figure 18: Dottup results.

The translation of the mRNA will reveal where the stop codon is located, and provides an explanation of why the coding potential decays before the end of the third exon, as can be seen in the NetGene2 output.

## 4.4. EXERCISES AND QUESTIONS

In relation to the query made using the ORFfinder program:

**1.** What do the red bars in the graph indicate?
**2.** Within these bars, what do the arrows indicate?
**3.** What do each of the columns that appear in the box at the bottom right of the graph indicate ("Strand", "Frame", "Start", "Stop" and "Length")?
**4.** Of all the possible open reading frames, which one is the most likely to be a coding region?
**5.** Why do all the open reading frames start with the ATG triplet? Which triplet/s do they end with?

**6.** What could happen if an open reading frame is interrupted by an intron?
**7.** What conditions would occur for an intron to be encompassed in an open reading frame?
**8.** How could it be verified that an open reading frame encodes a known protein?
**9.** What happens with the ORFfinder program if the analyzed sequence is a gene composed of several exons and introns?

In relation with the query made using the NetGene2 program:

**1.** What do the numerical tables donor splice sites and acceptor splice sites indicate? What does the last column in these tables indicate?
**2.** In the confidence column, what do the high values indicate?
**3.** What is represented in the upper graph?
**4.** In the two lower graphs, what do the vertical lines and the horizontal line (red or green) represent?
**5.** Based on the graphs, how would you identify places with a high probability of being functional donors/acceptors?
**6.** Why in the case of SOX9 do the beginnings and ends of the ORFs located with ORFfinder not exactly coincide with the donor and acceptor sites predicted by NetGene2?
**7.** Why can two exons of the same gene appear in different read phases?
**8.** Why does the coding potential decay before reaching the end of the last exon?

# 5.- A) MULTIPLE ALIGNMENT OF DNA AND PROTEIN SEQUENCES

## 5.1A. OBJECTIVE

When homologous sequences of nucleotides (DNA) or amino acids (proteins) of different species are compared in order to analyze the existing differences between them and their evolutionary relationships, an essential previous step in this analysis is to establish a multiple alignment of all the sequences. The objective of this exercise is to acquire the necessary skills to carry out multiple alignments of sequences and to become familiar with the use of the software that allow us to do so.
The procedure to be followed has several steps. The first consists in aligning all the sequences two by two. Therefore, we will initially describe how to proceed when making an alignment between two homologous sequences.

## 5.2A. THEORETICAL BASIS

***Alignment of two homologous nucleotide or amino acid sequences***
By comparing two homologous DNA or protein sequences, an alignment can be established by base-to-base pairing of the bases of each of the two sequences. For example, in the case of DNA:

**5'-AATGTCATGCGCTGAATCCCCCC-3'**
**5'-AAGGTCTTGCCCT-AATGCCCCC-3'**

If the two sequences being compared have different lengths, it is because one or both of them have incorporated or lost some residue (nucleotide or amino acid, depending on the sequences being compared). Thus, the first thing to identify is the location of insertions and deletions that may have occurred in each species since they diverged from a common ancestral species.
In base-to-base alignment matching, we may encounter one of three possible nucleotide/amino acidic sites or positions:
- Coincidences (*matches*): the same base/amino acid in the two sequences.
- Non-coincidences (*mismatches*): a different base/amino acid in each sequence.
- Insertions/deletions (*gaps*): gaps are represented by dashes (-) and mean that in one of the two sequences an insertion or a deletion occurred at that position.

When we compare a partial sequence of a gene/protein obtained from a species with the complete sequence of that gene/protein, the alignment will be performed by proposing a huge terminal gap that would represent the unknown information (missing data). These alignment positions are often represented by the question mark (?) in the incomplete sequence.
Obtaining the correct alignment is essential so that all subsequent evolutionary and phylogenetic analyses are not affected. Such alignment can be done manually if there are not many gaps and if the sequences are short and not very divergent. However, methods have been developed that facilitate the work and the fidelity of the result in any type of comparison:
1. The **dot matrix** method follows this procedure: one of the sequences is arranged on the vertical axis, and the other sequence on the horizontal axis, of a two-dimensional matrix. Each time there is an identical nucleotide/amino acid in both sequences, a dot is placed in the box corresponding to the x-position of the horizontal sequence and the y-position of the vertical sequence. The alignment is obtained by a diagonal line connecting the dots across the array starting in the upper left box and trying to end in the lower right box. The plot can reveal different situations as we are able to see in the following dot matrices for two hypothetical nucleotide sequences:
A. The two sequences are identical:

| | A | G | C | T | T | G | C | A | G | C |
|---|---|---|---|---|---|---|---|---|---|---|
| A | ● | | | | | | | ● | | |
| G | | ● | | | | ● | | | ● | |
| C | | | ● | | | | ● | | | ● |
| T | | | | ● | ● | | | | | |
| T | | | | ● | ● | | | | | |
| G | | ● | | | | ● | | | ● | |
| C | | | ● | | | | ● | | | ● |
| A | ● | | | | | | | ● | | |
| G | | ● | | | | ● | | | ● | |
| C | | | ● | | | | ● | | | ● |

B. The two sequences are equal in length but differ in sequence:

| | A | G | C | T | T | G | C | A | G | C |
|---|---|---|---|---|---|---|---|---|---|---|
| A | ● | | | | | | | ● | | |
| G | | ● | | | | ● | | | ● | |
| C | | | ● | | | | ● | | | ● |
| T | | | | ● | ● | | | | | |
| T | | | | ● | ● | | | | | |
| G | | ● | | | | ● | | | ● | |
| T | | | | ● | ● | | | | | |
| A | ● | | | | | | | ● | | |
| G | | ● | | | | ● | | | ● | |
| C | | | ● | | | | ● | | | ● |

C. The two sequences differ in length (only insertions and/or deletions would explain the differences between them):

| | A | G | C | T | T | G | C | A | G | C |
|---|---|---|---|---|---|---|---|---|---|---|
| A | ● | | | | | | | ● | | |
| G | | ● | | | | ● | | | ● | |
| C | | | ● | | | | ● | | | ● |
| T | | | | ● | ● | | | | | |
| T | | | | ● | ● | | | | | |
| C | | | ● | | | | ● | | | ● |
| A | ● | | | | | | | ● | | |
| G | | ● | | | | ● | | | ● | |
| C | | | ● | | | | ● | | | ● |

D. The two sequences differ in length (insertions and/or deletions would account for some of the differences between them) and in sequence (changes by substitution of one residue for another):

| | A | G | C | T | T | G | C | A | G | C |
|---|---|---|---|---|---|---|---|---|---|---|
| A | • | | | | | | | • | | |
| G | | • | | | | • | | | • | |
| C | | | • | | | | • | | | • |
| T | | | | • | • | | | | | |
| T | | | | • | • | | | | | |
| C | | | • | | | | • | | | • |
| A | • | | | | | | | • | | |
| C | | | • | | | | • | | | • |
| C | | | • | | | | • | | | • |

In a longer sequence and with more changes than those reflected here, it is much more difficult to establish the alignment and there may be more alternative pathways that would explain the differences between two sequences.

In fact, there are usually a very large number of points in the matrix which, together with the absence of a perfect diagonal, makes it difficult to trace the alignment. A method has been devised to improve the definition of the alignment. It consists of comparing the two sequences using "sliding windows" that make comparisons of *n* by *n* residues, instead of nucleotide by nucleotide. A match in this case is determined from a given threshold. Thus, two parameters are fundamental in this type of comparison: window size and stringency. Once a window size has been established it remains constant throughout the analysis. It consists of determining every how many residues a comparison is made. Thus, if the window size is five residues, it means that we compare the two sequences progressively by 5 residues. The astringency determines the threshold: the number of residues that must be matched within that window. This eliminates many of the false identity points in the matrix.

2. A second method consists of defining an alignment as one in which the number of mismatches and gaps are minimized according to certain criteria. The problem is that in order to increase the number of mismatches it is usually necessary to increase the number of gaps. Therefore, according to this criterion, several alignment options are possible, so a procedure has been designed to calculate an index of divergence or dissimilarity between the two sequences being compared. This index will have different values for each of the alternative alignments obtained. The alignment with the lowest divergence index will be chosen as the best of all.

The calculation of the divergence index depends on the gap penalty, which usually has two components: gap-opening penalty and gap-extension penalty. Gap penalties are factors by which the gap values (the number and length of gaps) are multiplied in order to establish equivalence between these values and the value of mismatches (number of substitutions). Thus, the penalty is based on our own experience through the comparison between the calculation of the frequency of insertions and deletions that have occurred in evolution since the separation of the two species whose sequences are being aligned and the frequency with which nucleotide (or amino acid) substitutions have occurred.

In the case of protein sequences, dissimilarities in the different amino acid positions can be valued with different weights depending on whether the change produced is to an amino acid more or less similar in its biochemical properties. Thus, certain groups of amino acids have been established by biochemical affinity whose pairings in an alignment receive higher or lower scores according to different criteria, instead of a score of zero which is what sites

receive where there is dissimilarity and the paired amino acids do not have any biochemical affinity.

***Multiple alignments***

Multiple alignments follow a similar procedure to the one described above, but the complexity of the calculations becomes greater as the number of sequences to be aligned increases. There are different programs that can perform this type of alignment, such as Clustal X, or the online version of Clustal Omega, which implements the Clustal algorithm (Higgins and Sharp, 1988). In this case, alignments are performed in a three-step process. First, all sequences are compared two by two (pairwise alignments). Second, a dendrogram (similar to a phylogenetic tree) is constructed that groups the sequences by similarity. Third, multiple alignment is carried out using the dendrogram as a guide and aligning sequences progressively according to the branching order of the tree; that is, first the two sequences with the highest similarity are aligned and sequences are progressively added to the alignment in order of decreasing similarity.

# 5.- B) PHYLOGENETIC ANALYSIS

## 5.1B. OBJECTIVE

Molecular phylogeny consists of the study of the evolutionary relationships between organisms from molecular data arranged in a multiple alignment of DNA or protein sequences. The objective of this practical is to introduce us to the theory and methodology used in phylogenetic analysis as well as to familiarize us with the use of computer programs for phylogenetic analysis.

## 5.2B. THEORETICAL BASIS

To simplify the wording of this text, from now on we will always refer to DNA sequences, everything that is said also being applicable to the analysis of protein sequences.

In phylogenetic analysis, the objective is the construction of a phylogenetic tree that illustrates the evolutionary history of a group of species. A phylogenetic tree is a graph composed of nodes and branches in which a branch connects two adjacent nodes. The nodes represent the species and the branches define the relationships between those species in terms of descent and ancestry. The branching pattern is called the topology of the tree. A distinction must be made between terminal nodes and internal nodes. The latter represent hypothetical ancestral species while the terminal nodes represent species existing today. Species that are connected by branches to the same internal node share that ancestral node. Branches connecting external nodes to internal nodes are called external or terminal branches while those connecting internal nodes are internal branches. A node can be bifurcated if it has only two descendants or multifurcated if it has more than two. Generally, the most common representation of phylogenies uses bifurcate trees since the speciation process is assumed to be binary: two descendant species from a common ancestral species. A multifurcation or polytomy in a tree can be interpreted in two ways: a) it represents a reality, i.e., one ancestor has given rise to more than two descendant species; b) there is ambiguity in determining the correct bifurcation pattern because the available data are not conclusive.

A natural clade or monophyletic group consists of a group of taxa (species, or group of species such as a genus, family, order or class) that derive from a common ancestry that is not shared with any other taxa outside the group. A taxonomic group (genus, family, order or class) is expected to be monophyletic. However, some currently established taxonomic groups may be non-monophyletic: molecular phylogeny has shown, in some cases, that a taxonomic group has a common ancestor shared with other taxa (paraphyletic group); a polyphyletic group consists of two lineages that have acquired the same character by evolutionary convergence (organisms classified in the same polyphyletic group share phenotypic homoplasies).

A tree can be rooted when there is a node, the root, which is unequivocally the most recent common ancestor of all the species compared. From the root, a single evolutionary pathway gives rise to each of the nodes. A tree without a root is a tree that only specifies the kinship relationships between the compared species without describing the evolutionary steps that have led from a common ancestor to those species.

A scaled tree is one in which its branches are scaled, i.e., the length of each branch is proportional to the number of changes produced between the sequences being compared. In a non-scaled tree, the lengths of the branches are not proportional to the number of changes, therefore the terminal nodes will appear aligned.

For a group of species there are different possible trees and the number of these increases in relation to the number of species compared. However, only one of these trees is the correct tree which, depending on the accuracy of our data and our analysis, may or may not coincide with the tree inferred in our phylogenetic reconstruction.

In any case, we must always keep in mind that in our analysis what we are comparing are homologous DNA sequences obtained from each of the species we are studying. Thus, in principle, what we get is a gene tree. However, each gene may have different evolutionary histories and the rates and modes of these may not consistently reflect the evolutionary

history of the species. Therefore, to obtain as accurate a species tree as possible, it is more correct to analyze the history of different genes and non-gene sequences.
When choosing which type of sequences to use in our phylogenetic analysis, we must take into account the rate of change of the compared sequences. Thus, if the group to be compared is formed by phylogenetically very close species, there is a need for a sequence that evolves faster and that has accumulated enough changes in the diversification process of the compared group. In this case, it is interesting to resort to non-gene sequences that change more rapidly. The use of conserved gene sequences with an important function in the organism would be inadvisable in this instance, since it is very likely that very few changes have occurred in the compared sequences and, therefore, there is little phylogenetic signal with resolving capacity for phylogenetic reconstruction. However, the use of mitochondrial DNA gene sequences, which have a faster rate of evolution than nuclear DNA sequences, is often useful. In contrast, when the comparison is between species of distant taxonomic groups, non-gene sequences can be very disparate and undesirable for phylogenetic analysis. In this case, it is more convenient to use more conserved sequences.

***Phylogenetic reconstruction methods***

Most of the proposed phylogenetic inference methods define an optimization criterion that aims to choose the best tree among all possible trees that could explain the starting data. This criterion gives different values to each possible tree. This value is used to compare the different trees. Algorithms exist to calculate these values and identify the best tree according to the optimization criterion.
The following phylogenetic inference methods are currently available: a) methods based on genetic distance matrices; b) maximum parsimony method; c) maximum likelihood method; d) Bayesian method.

*Methods based on distance matrices*

There are several methods for reconstructing phylogenetic trees based on genetic distance matrices. The first step in all of them is to construct the distance matrix. This is done by estimating the differences between each pair of sequences in the alignment. The simplest way to calculate the genetic distance is to calculate the number of differences (p) between the sequences. However, if p has a high value (the sequences have diverged considerably) it may be that, at each site in the alignment, multiple substitutions and reversals have occurred, so that p is giving an underestimate of the number of nucleotide substitutions that have actually occurred. Therefore, several methods have been developed to calculate corrected distances based on probabilistic models. The calculations of such distances are corrected values of p according to these models. Each model assumes a different evolutionary pattern with respect to nucleotide composition and rates of change for each type of nucleotide substitution, for each nucleotide position, and for each lineage. We will return to these models later when we study maximum likelihood methods.
A typical distance matrix looks like this:

| | Species 1 | Species 2 | Species 3 | Species 4 | Species 5 |
|---|---|---|---|---|---|
| Species 1 | | 0.012 | 0.018 | 0.022 | 0.035 |
| Species 2 | | | 0.013 | 0.020 | 0.032 |
| Species 3 | | | | 0.021 | 0.033 |
| Species 4 | | | | | 0.020 |

The distance values of this matrix are used to reconstruct the tree, the length of the branches being proportional to these values. As mentioned at the beginning, there are different inference methods based on distances, but the most popular is the Neighbor-joining or N-J method. This method is based on an algorithm that tries to find the shortest tree, that is, the one that minimizes the total length of the tree, understood as the sum of the lengths of all its branches. First, the two sequences that are most similar (the shortest genetic distance between them) are identified. That is to say, among all the pairs of sequences compared, the two sequences whose sum of the lengths of their branches is the smallest are identified. This pair of sequences constitutes the first pair of "neighbors", connected through an internal node. The next step is to consider this pair as a single sequence by computing the arithmetic mean distance between them and the rest of the sequences and constructing a new distance matrix. Then the pair of sequences whose sum of the lengths of their branches is the smallest is chosen again, a procedure that is continued until all the internal nodes of the tree are identified.
As an exercise, one could try to manually construct a tree by this method from the distance matrix shown above.

*Maximum parsimony method*

The maximum parsimony method aims to construct a phylogeny with the topology that requires the least number of evolutionary changes to explain the observed differences between the aligned sequences. Sometimes, this criterion is met by two or more trees that will be equally parsimonious. To apply this criterion, each nucleotide site in the sequence is classified as follows:
**-Invariable**: all sequences have the same nucleotide at that position.
**-Informative**: a site is phylogenetically informative from the point of view of maximum parsimony when there are at least two different classes of nucleotides, each represented at least twice in the alignment.
**-Non-informative**: a site that, being variable, does not meet the above requirement.
Once the sites in the alignment have been classified and the informative sites identified, the minimum number of substitutions needed to explain each informative site is calculated for each possible tree. Summing the number of changes for the set of all informative sites for each possible tree, the tree that is explained by the least number of changes is chosen.
If there is more than one tree with that number, a consensus tree can be obtained, from which we can distinguish: a) strict consensus, in which all conflicting branches are resolved by collapsing them to a single multifurcated node; b) majority-rule consensus, in which conflicting branches are resolved by selecting the branching pattern observed in more than 50% of the trees obtained.

*Maximum likelihood method*

The likelihood, L, of a phylogenetic tree is the probability that the data observed in an alignment can be explained from that phylogeny constructed according to a given evolutionary model of nucleotide substitution, i.e. L =P(data|tree+model). The objective of the maximum likelihood method is to find the tree with the highest value of L, among all the possible trees that would explain the observed data.
The question is: What is the probability that a given phylogeny has generated the observed data in an alignment assuming a given evolutionary model of nucleotide substitution?
To answer the question, assuming that each site in the alignment evolves independently, one must calculate L for each site separately (Ln) and as a whole (L = L1 x L2 x L3 x...x Ln). To calculate each Ln, one must consider all possible scenarios through which the current nucleotide in each sequence has been arrived at from an ancestral nucleotide. Some scenarios will be more plausible than others but all will have at least some probability of being the ones that have generated the current situation. Therefore, each Ln has a probability that is equal to the sum of the probabilities of each possible phylogenetic reconstruction that explains the current data from the ancestral situation. These probabilities depend on the evolutionary model we assume and on the length of the branches which, in turn, depends on the rate of

substitution and evolutionary time. For convenience, the likelihood is calculated by logarithmic transformation (lnL) so that lnL = lnL1 + lnL2 + lnL3 +...+ lnLn.
A phylogenetic tree inferred by this method is only valid for the assumed evolutionary model but may not be valid for another evolutionary model. Therefore, a correct choice of the evolutionary model applicable to the analyzed sequences is essential. There are different evolutionary models that attempt to explain the nucleotide substitution pattern followed by the analyzed sequences. From a general model in which it is assumed that each type of nucleotide substitution has a different rate and that each nucleotide appears in the sequence in a different proportion to a simpler model in which we assume that all nucleotides appear with the same frequency (25% for each of the four nucleotides) and there is the same rate of change for all types of nucleotide substitution. Going through different models in which differences in the ratio of nucleotides or lack thereof are taken into account and different types of nucleotide substitutions (different types of transitions and transversions) are considered differentially. In addition, each model may assume that the rates of change differ between different nucleotide sites of the alignment or between different phylogeny lineages. It therefore becomes necessary to test which evolutionary model best fits the sequences analyzed.
*Bayesian inference method*
Bayesian inference of a phylogeny is based on a quantity called posterior probability of distribution trees, which is the probability of a tree conditioned by the observations [P(tree+model|data)]. The conditioning is achieved through Bayes' theorem. It is not possible to calculate the posterior probability of distribution trees analytically. Instead, a simulation technique called Markov chain Monte Carlo (MCMC) is used to approximate this probability.

***Reliability of phylogenetic reconstruction***
To answer the question about the reliability of the tree obtained by any of the existing methods, there are methods that allow us to estimate the statistical support of the topology obtained. One of the most popular is the bootstrap method. Once a phylogenetic tree has been obtained from an alignment of sequences and with a given method, this phylogeny becomes the null hypothesis to be tested by bootstrap. For this purpose, new different alignments (an appropriate number could be between 500 and 1000) are constructed by re-sampling with re-replacement. That is, different random alignments are constructed by replacing a given number of nucleotide positions with other positions of the alignment, each of which has the same probability of replacing the others. Thus in the new alignment, a site may be repeated more than once at the expense of other sites. Thus, if the alignment has this sequence of nucleotide positions:
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
Different re-samplings can result in alignments like these:
1 1 1 4 4 6 6 6 6 10 11 12 12 12 12 12 17 18 19 20 20 20 20 24 24
1 1 3 3 3 6 7 8 8 8 8 8 13 14 15 15 15 15 19 19 19 25 25 25 25
2 2 2 4 5 5 5 8 9 10 10 12 12 12 16 16 17 19 20 21 21 24 24 24 25
From each of the new alignments a new phylogenetic tree is inferred using the same method used with the initial alignment. The percentage of times that each interior branch of the initial tree is confirmed in the set of bootstrapped trees constitutes the bootstrap value of each branch. As a rule of thumb, if the bootstrap value of a given interior branch is greater than 95%, the topology of that branch is accepted to be correct.

## 5.3. METHODOLOGY

The practice will be carried out by running online the Clustal Omega program (https://www.ebi.ac.uk/Tools/msa/clustalo/).

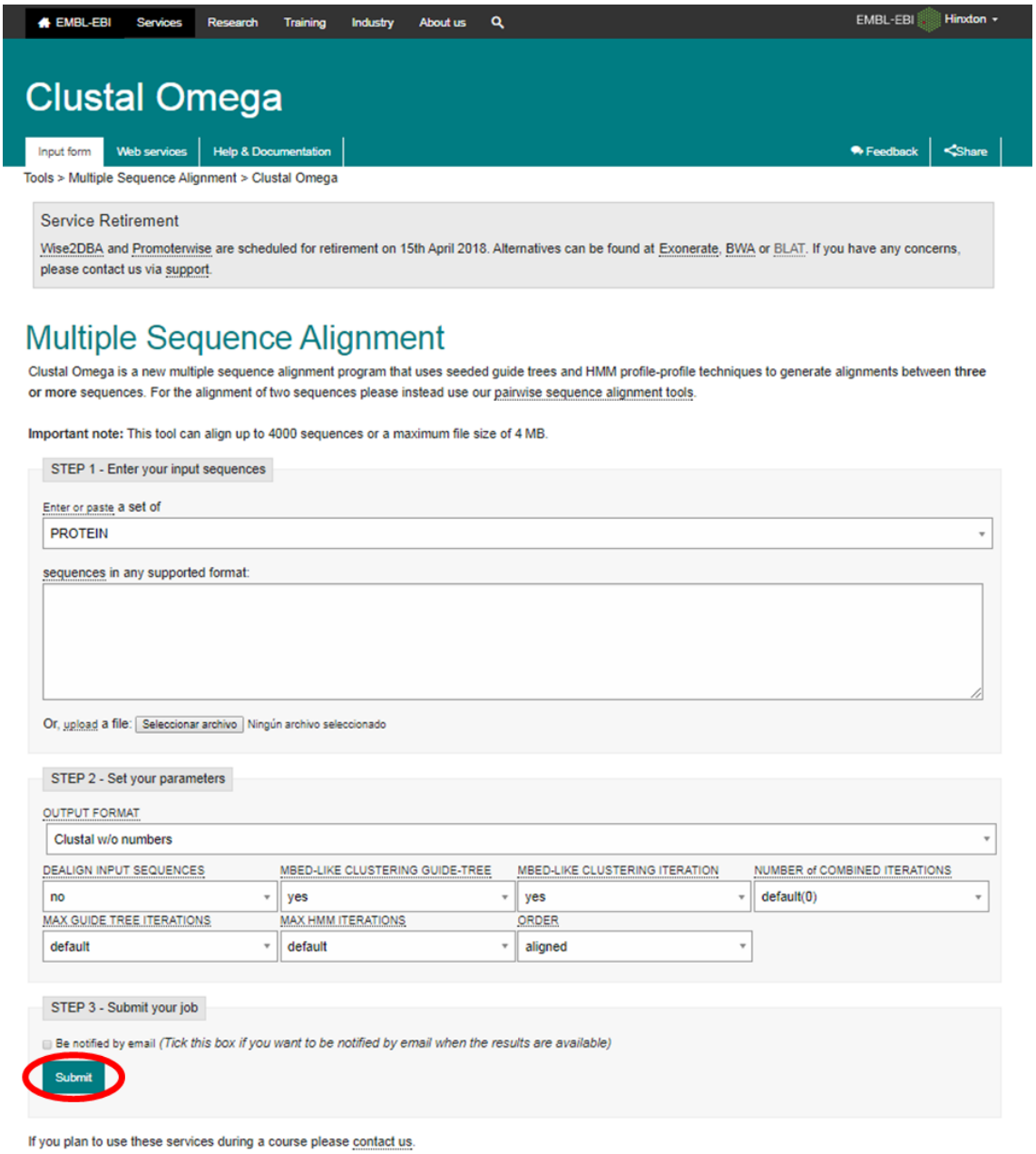



Figure 1: Clustal Omega interface.

First, select in "STEP 1" the type of sequence to be analyzed (protein, DNA or RNA). In the box, load the sequences to be analyzed (paste the sequences or upload the file, in FASTA format in any case), leaving the default analysis options that appear in "STEP2". Click on "Submit" to carry out the multiple sequence alignment.

As indicated above, the alignment is performed in a three-step process:
1.- All sequences are compared two by two (pairwise alignments).
2.- A dendrogram (similar in appearance to a phylogenetic tree) is constructed that groups the sequences by similarity.
The dendrogram is used as a guide, aligning sequences progressively according to the branching order of the tree. The two sequences with the highest similarity are aligned and sequences are progressively added to the alignment in order of decreasing similarity.
The program displays the best alignment combination of the sequences, showing similarities and differences.

The symbol "*" indicates positions in the alignment where the same residue (nucleotide or amino acid) is present in all sequences.
In the case of protein sequence alignments, the symbol ":" indicates that one of the following "strong" amino acid groups with similar chemical properties is highly conserved: STA, NEQK, NHQK, NDEQ, QHRK, MILV, MILF, HY, FYW, while the symbol ". " indicates that one of the following "weak" groups is highly conserved: CSA, ATV, SAG, STNK, STPA, SGND, SNDEQK, NDEQHK, NEQHRK, FVLIM, HFY.
Once the multiple alignment has been obtained, the program offers different options, including downloading the alignment or performing a simple phylogenetic analysis (Send to simple_phylogeny). In this option, the UPGMA analysis will be chosen:

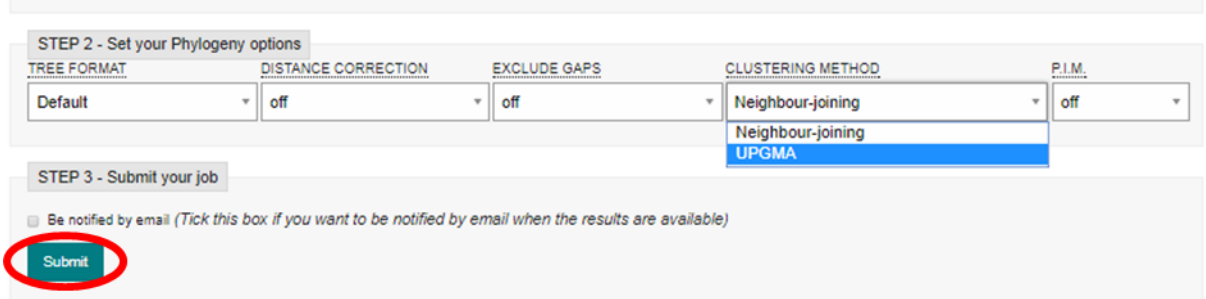
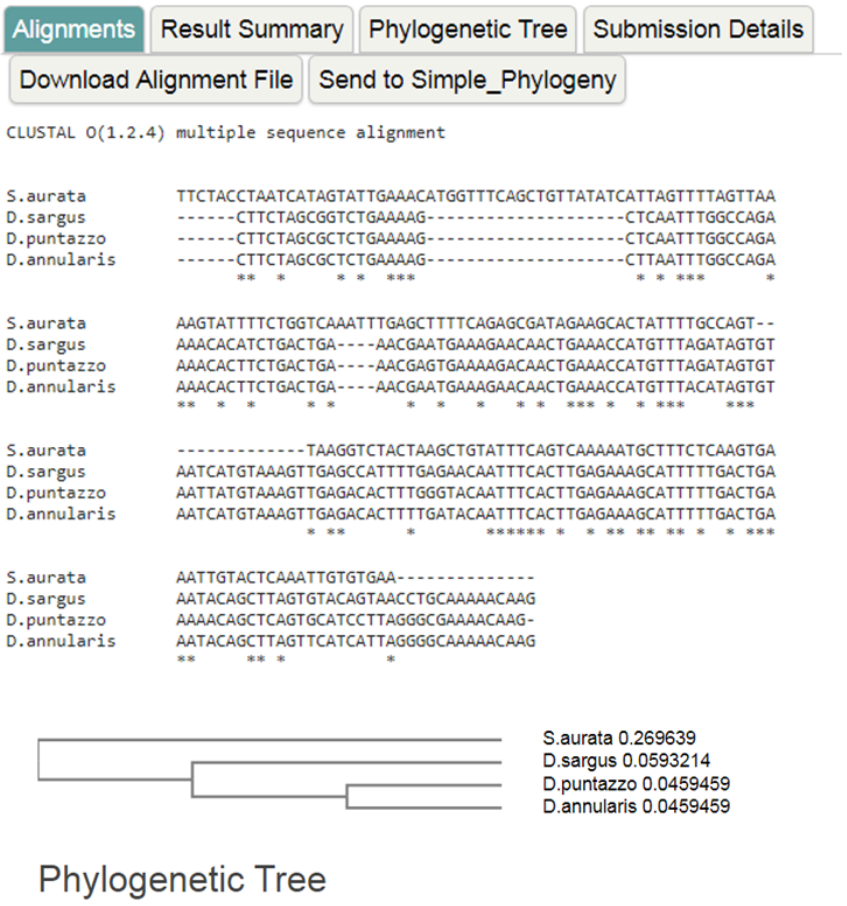



Figure 2: Clustal Omega steps 2 and 3.

Finally, click on "Submit" to obtain the sequence tree.

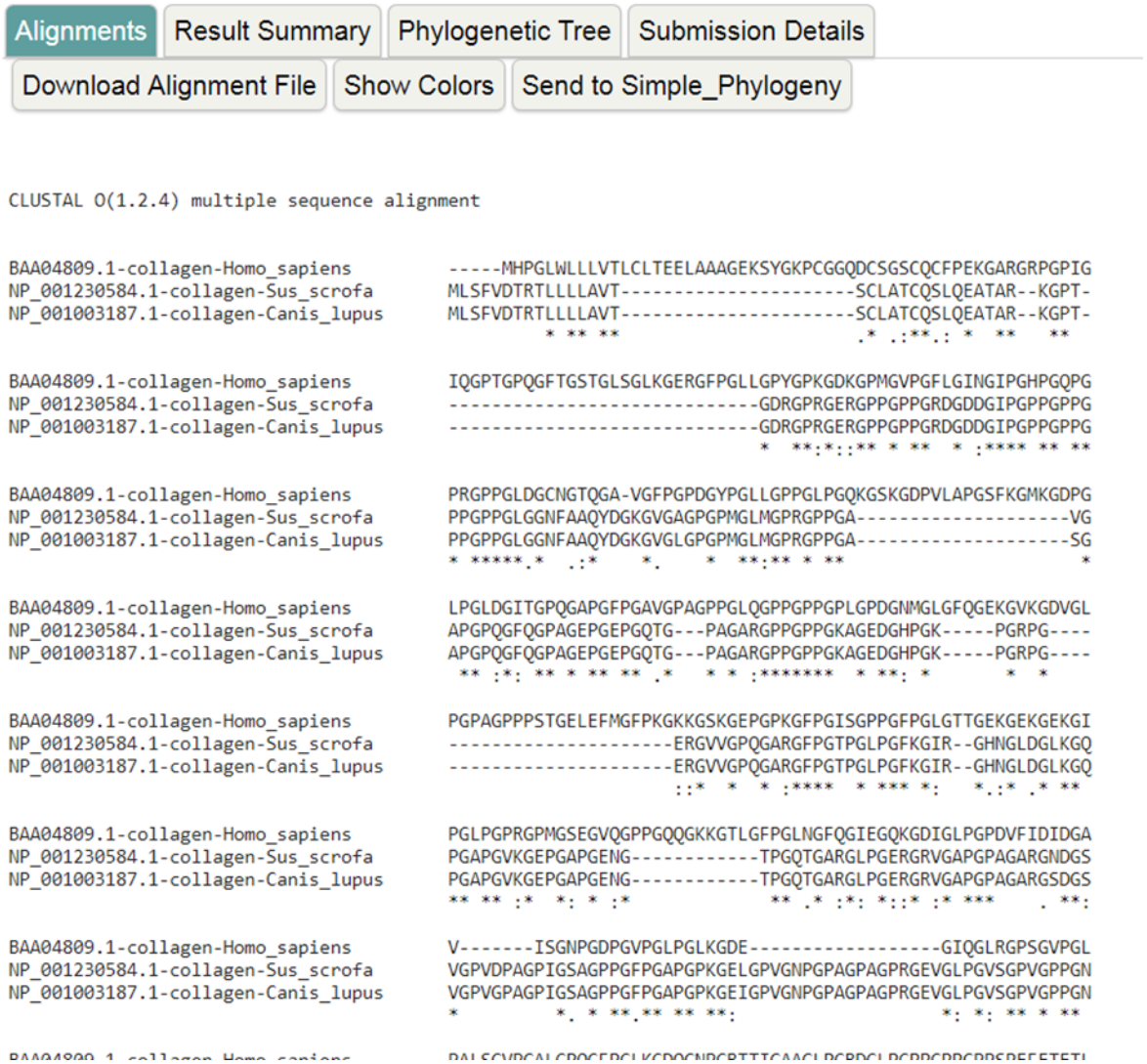
## 5.4. EXERCISES AND QUESTIONS

The following are examples and exercises using nucleotide and protein sequences. A multiple alignment (forward alignment) of homologous sequences will be performed. From the sequence comparison data, a distance matrix will be obtained, and from the distance matrix, a dendrogram will be obtained using an algorithmic method such as UPGMA (Unweigthed Pair Group Method with Arithmetic mean).
Some key concepts in this type of analysis are:
- The phylogenetic reconstruction based on distances that we will carry out is based on the fact that the distance between taxa is a reflection of the phylogenetic relationship between them.
- The evolutionary distance is the average of changes that have occurred in a position between two pairs of sequences throughout their evolution from their common ancestor.
- Molecular phylogeny is an estimation of phylogenetic relationships based on the comparison of DNA or protein sequences belonging to these taxa.
Example of nucleotide sequence alignment: Multiple alignment of four sequences belonging to the EcoRI satellite DNA family in four different species of sparid fish.
Using the sequence file of this repeated DNA family in FASTA format available in the PRADO2 platform of the course, obtain the corresponding multiple alignment and the phylogenetic tree using the Clustal Omega online program.
To do so, follow the methodology described in the previous section for the Clustal Omega online program (https://www.ebi.ac.uk/Tools/msa/clustalo/).
You should obtain an alignment and sequence tree similar to the following figure:




**Example amino acid sequence alignment: Multiple alignment of three sequences belonging to collagen.**

Using now the collagen sequence file, you should obtain an alignment similar to the following partial figure:

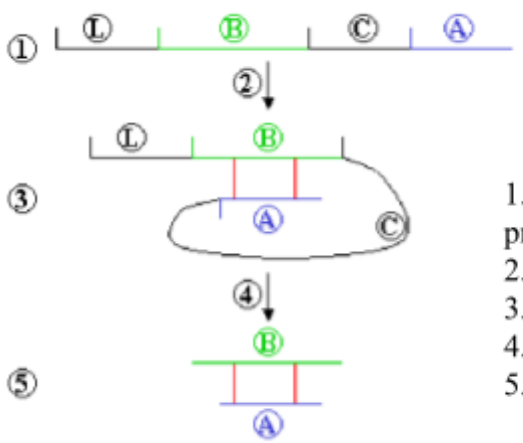



**<u>Exercise 1</u>: Search for homologous sequences in the NCBI database and perform a multiple alignment pertaining to the insulin gene of different species.**

Search the following accession numbers in the database, as explained in the practice corresponding to DNA and protein sequence databases (http://www.ncbi.nlm.nih.gov/genbank), and perform a multiple alignment with these sequences:

AAP36446
NP_001008996
NP_001123565
NP_776351
NP_032413
NP_062003
P01318
P01315

Next, compare the multiple alignment obtained with the following scheme:




Now, include in the above alignment the following sequences:
AAA37041
XP_006008147
XP_006033708
NP_990553

And finally, establish the parentage relationships between the species by making a phylogenetic tree as explained in the previous example (Send to simple_phylogeny; UPGMA; Submit). Explain and discuss the results obtained.

**Exercise 2: Search for homologous sequences related to the first laboratory practice of the course.**
Search the NCBI database for the sequences corresponding to the following accession numbers:
AF509333
AY305326
AF497479
a) To which organism and to which type of sequence does each accession number refer?
b) Perform a multiple alignment with them.
c) Explain and discuss the results of the alignment.

**Exercise 3: Searching for homologous ribosomal DNA sequences**
Perform a multiple alignment and a phylogenetic tree with the following accession numbers corresponding to 18S ribosomal DNA gene sequences in different species:
L11288
AF173605
AF115860
X00686
NR_033238
AF173614
AF173630
AF173611

AF173612

**<u>Exercise 4</u>: Identification of DNA sequence by phylogenetic similarity on the tree of the previous exercise**

Using the file called "Problem Sequence" available on the PRADO2 platform of the course, which contains the sequence of the 18S ribosomal DNA gene of an unidentified species, and based on the multiple alignments and the sequence tree obtained in the previous exercise, try to determine, as far as possible, to which organism it belongs.

# 6.- EXPRESSION OF GENES INVOLVED IN MAMMALIAN TESTIS DEVELOPMENT

## 6.1. AIM

The student will learn a method, based on molecular diagnostics, which is commonly used for sexing mammalian embryos as well as to identify embryonic organs in which the *SOX9* gene is expressed.

## 6.2. THEORETICAL BASIS

***Genetic sex determination in mammals***

In mammals, the presence of a Y chromosome determines male sex, while its absence implies female development. At the beginning of its embryonic development, the gonad is undifferentiated and bipotential, which means that it can follow two alternative and mutually exclusive developmental routes: testis or ovary. In the XY embryonic gonad, the *SRY* gene (located on the Y chromosome; * see note on correct mammalian gene typing at the end of this script) initiates a genetic cascade that induces a subpopulation of somatic cells to differentiate as Sertoli cells, which are subsequently responsible for orchestrating testicular development. These Sertoli cells are organized into sex cords (precursors of the seminiferous tubules of the adult testis), within which the germ cells that cease to proliferate (mitotic arrest) are found. Sertoli cells also control the differentiation of Leydig cells, which are testosterone- and dihydrotestosterone-secreting cells that will masculinize the individual's soma. In the male pathway of mouse gonadal development, the SRY protein binds, together with the steroidogenic factor, SF1, to an enhancer sequence of the *Sox9* gene and activates its expression. Ectopic expression of *Sox9* in the developing XX gonad results is testis development, whereas its mutation in an XY gonad will result in the activation of the ovarian pathway. Therefore, *Sox9*, as well as *Sry*, are necessary and sufficient to activate testicular organogenesis. SOX9 activates the *Fgf9* gene, which in turn stabilizes the expression of *Sox9*, establishing a self-maintaining loop of *Sox9* expression in the male gonad. SOX9 also activates the expression of other genes such as *Amh* (Antimüllerian hormone), *Vnn1* (Vanin-1), and *Ptgds* (prostaglandin synthetase) which are known to be involved in testis differentiation. Taking all these considerations into account, nowadays it is widely accepted that *SOX9* is the main switch gene controlling testis development, and exerts this function not only in mammals, but in all vertebrates.

In the XX gonad, the absence of the Y chromosome, and therefore of the *SRY* gene, implies the silencing of *SOX9* and the upregulation of *RSPO1* and *WNT4*, which initiate a genetic cascade leading to ovarian development. In the absence of SOX9, bipotential supporting cells of the embryonic gonad differentiate as pre-granulosa cells (and not as pre-Sertoli cells), steroidogenic cells differentiate as theca cells (rather than as Leydig cells) and germ cells initiate meiosis, which will cease shortly after prophase I (meiotic arrest). In summary, in the

absence of *Sry*, gonadal organogenesis follows the ovarian genetic pathway and the somatic phenotype of the individual will be female.

The classical view coined by Jost (1953), that the ovarian pathway is the default pathway, changed on the basis of new data in which XX mice were observed to develop partial or complete sex reversal when harboring loss-of-function mutations in genes such as *Wnt4* and *Rspo1*. XX *Wnt4*$^{-/-}$ individuals (homozygous for the mutated allele) showed partially masculinized gonads with expression of the *Sox9* and *Fgf9* genes, Leydig cell differentiation, cell migration from the adjacent mesonephros into the gonad (XY gonad-specific morphological event) and development of a testis-specific vascular pattern. The loss-of-function mutation in the *RSPO1* gene causes complete female-to-male sex reversal, i.e. XX males. This was the first case describing a single mutation in a gene causing complete female-to-male sex reversal, and this mutation places *RSPO1* as an ovarian determinant in mammals. Another gene involved in the ovarian pathway is *FOXL2*, which is necessary for the development and maintenance of ovarian structure. The absence of functional granulosa cells leads to premature initiation of folliculogenesis and premature ovarian failure. However, the absence of female-to-male sex reversal of *Foxl2*$^{-/-}$ mutant mice indicates that it is not an ovary determining gene. In this practical session we will amplify a fragment of the *Sry* gene, and verify that it is present in male cells (XY), whereas female cells (XX) lack the gene.

### *SOX9*: A pleiotropic gene

The *SOX9* gene was initially identified as the one responsible for campomelic dysplasia syndrome (CD), a skeletal malformation associated with XY sex reversal. *SOX9* is a gene that codes for a transcription factor belonging to the SOX (Sry-like HMG box) family of proteins. In humans it is located on chromosomal region 17q24.3-q25.1 and is composed of three exons and two introns. *SOX9* is expressed in a large number of embryonic tissues including chondrocytes, Sertoli cells, otic placode cells, pancreatic cells, intestinal epithelial cells, neural crest cells, lung epithelial cells, notochord cells and several other tissues. This suggests that *SOX9* has multiple functions during mammalian embryonic development, and highlight the role that *SOX9* has in the development of the different organs in which it is expressed. In the mouse, this gene is located on chromosome 11. The first mutant mouse for *Sox9* was described in 2001. These mice harbouring a heterozygous null mutation for *Sox9* reproduced most of the skeletal malformations shown by CD patients, although other abnormalities, such as XY sex reversal, were not evident. These heterozygous *Sox9* mutant mice died around birth, so it was not possible to generate homozygous mutant mice. Because of the latter, conditional mutant mice were generated for the various tissues where *Sox9* is expressed, i.e. animals that only lack the function of the gene in specific tissues or organs. Thus, *Sox9* has been conditionally inactivated in homozygosity in chondrocytes, resulting in the complete absence of cartilage and bone formation. Embryos with *Sox9* inactivated in chondrocytes exhibited generalized chondrodysplasia. *Sox9* has also been conditionally inactivated during mouse testicular development. In this case, XY individuals were observed to develop phenotypically as females having ovaries rather than testes. Despite this, the testis-determining gene, *Sry*, continued to be expressed indicating that *Sox9* acts downstream in the gene cascade that regulates testicular development. Homozygous conditional inactivation of *Sox9* in mice has shown that it is also required for spinal glial cell differentiation, cardiac valve and septum formation, notochord development, pancreatic stem cell maintenance, otic placode invagination, prostate development, neural crest cell survival, and maintenance of spermatogenesis. The second part of this practical will consist of the observation of histological sections on which immunohistochemistry with an anti-SOX9 antibody have been performed.

# 6.3. METHODOLOGY

### 6.3.1. PCR for detection of the Sry gene

For the detection of the *Sry* gene we will use the PCR (Polymerase Chain Reaction) technique. For this purpose, we have designed specific primers, on the one hand for the *Sry* gene, which is located on the Y chromosome and is therefore specific to males, and on the other hand for the *Myogenin* gene, an autosomal gene that will serve as a positive control. We will perform a "duplex PCR", that is, a PCR in which the primers for both genes are present in a single reaction, and we can therefore simultaneously amplify the fragments corresponding to the two genes. The sequences of the primers are as follows:

*-Primers for amplification of the mouse Sry gene:*

Sry-F 5'- GCA AAC AGC TTT GTG GTC AA 3'
Sry-R 5'- GGA AAA GGG GAT GAA ATG GT 3'

-Primers for the amplification of the mouse *Myogenin* gene:

Mio-F 5'- TTA CGT CCA TCG TGG ACA GCA T 3'
Mio-R 5' TGG GCT GGG TGT TAG CCT TAT G 3'

***Amplification Reaction (PCR)***

In a 200µl microtube add, following the order indicated, the following reagents for a final volume of 25µl:

- Sterile water 16,5 µl
- 10% PCR buffer (10x) 2,5 µl
- MgCl2 (25mM) 1,5 µl
- DMSO 1,2 µl
- Primer Sry-F (500 ng/µl) 0,5 µl
- Primer Sry-R (500 ng/µl) 0,5 µl
- Primer Mio-F (500 ng/µl) 0,5 µl
- Primer Mio-R (500 ng/µl) 0,5 µl
- dNTPs (25 mM) 0,2 µl
- Taq polymerase (2U) 0,1 µl
- DNA (100 ng/µl) 1 µl

The microtubes are then placed in the thermal cycler and programmed for 35 cycles according to the following program:

Denaturation: 91ºC, 45 sec.
Alignment: 60ºC, 60 sec.
Extension: 72ºC, 45 sec.

Once the PCR is finished, we will perform an agarose gel electrophoresis using the samples and the individuals will be sexed. After the PCR reaction, a 179 bp fragment will be amplified for the *Sry* gene and a *246* bp fragment for the *Myogenin* gene. Both amplicons can be perfectly separated by agarose gel electrophoresis. In the case of a male, both bands will be distinguishable, whereas in the case of a female, only the 246 bp band will be visible. If no band is observed, this indicates that the PCR did not work (Figure 1).

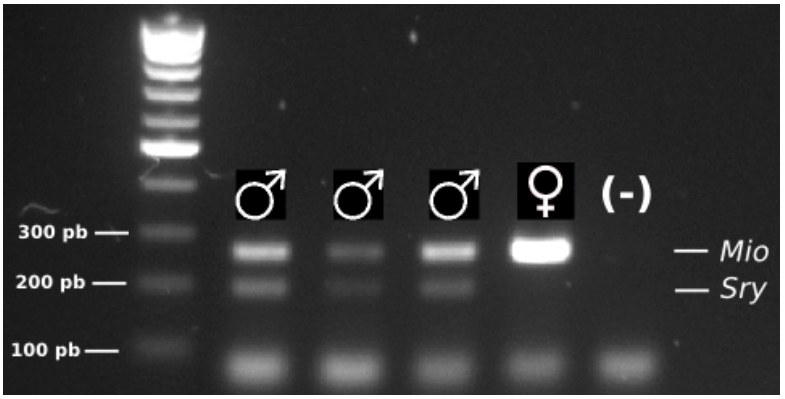



Figure 1: Electrophoresis of the products of a PCR performed for the sexing of mouse embryos. The presence in the gel of a band corresponding to the Sry gene denotes the presence of a male, while its absence indicates that the embryo is female. The Myogenin band serves as the quality control (positive control) of the PCR reaction. (-) is the negative control (no template reaction), indicating the absence of DNA contamination in the PCR reaction mixture.

### 6.3.2. Observation of immunohistochemistry slides for SOX9

Several techniques are currently available to detect gene expression in tissues. One of these techniques is immunohistochemistry, which allows us to identify the cell type where a protein of interest is localized, a situation that in most cases implies that the gene coding for that protein is being expressed in that cell type. In an immunohistochemical technique, the localization of the protein of interest is revealed by an enzymatic reaction, with one of the most widely used at present being that catalyzed by horseradish peroxidase. One of the ways to perform immunohistochemistry using the peroxidase method is to fix the tissue of interest, dehydrate it, embed it in paraffin, and perform histological sections. After deparaffinization and hydration, the histological sections are incubated with a solution containing the primary antibody, specific to our protein of interest. In this situation, in those cells where the protein of interest is present, binding between the protein of interest and the primary antibody will occur. Since the protein of interest is bound inside the cell, the complex will also remain inside the cell. The preparations are then washed intensively to remove the primary antibody that has not bound to the protein of interest, and re-incubated with a solution containing a secondary antibody, which is a specific antibody against immunoglobulin G of the species where the primary antibody was generated. The secondary antibody is conjugated to horseradish peroxidase (anti-Ig-Peroxidase). This causes a complex to form between the protein of

interest, the primary-antibody and the conjugated secondary antibody, which remains inside the cells where the protein of interest is present. The preparations are then washed again to remove the free secondary antibody and incubated with a solution containing $H_2O_2$ and di-amino benzidine (DAB). Peroxidase catalyzes the reaction $2H_20_2$ ---> $2H_20 + O_2$. This causes $O_2$ to be released inside those cells, oxidizing DAB, and finally giving rise to a brown precipitate (Figure 2).

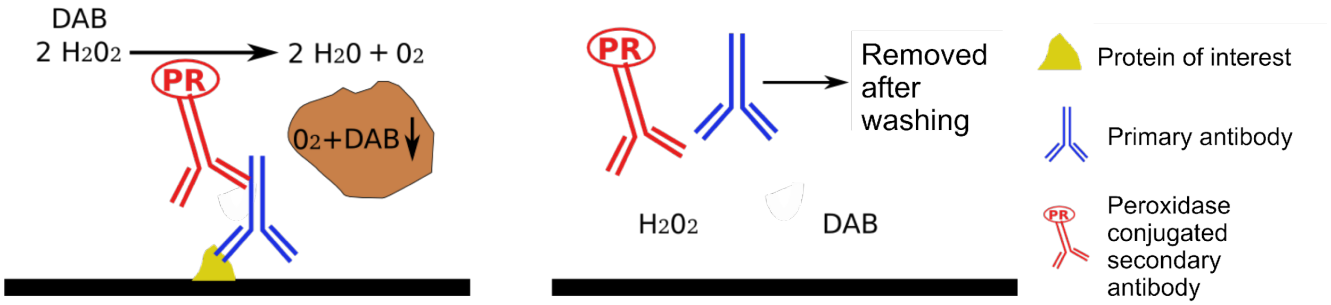



Figure 2: Molecular basis of the immunohistochemistry technique. The presence in the sample of the protein of interest (left schematic) allows anchoring to the preparation of the complex composed of the primary antibody, secondary antibody and peroxidase, allowing the DAB-stained reaction. Its absence (right schematic) allows washing of all components, with no reaction.

Finally, the preparations are counterstained with haematoxylin, dehydrated and mounted with DePeX for observation under the light microscope. After this process, we will observe the cells positive for the protein of interest in brown, while the nuclei of the negative cells are blue (haematoxylin; Figure 3).

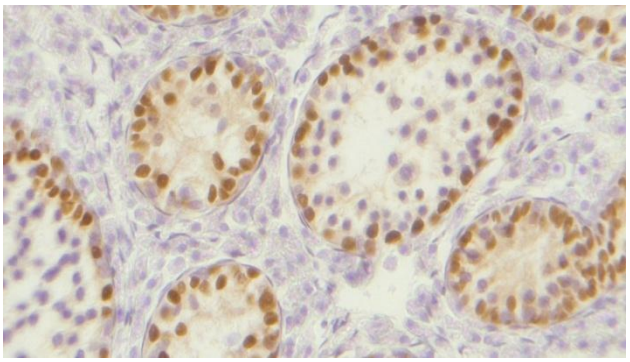


Figure 3: Immunohistochemical labeling of mouse testicular tissue using an anti-SOX9 primary antibody. Only Sertoli cells appear marked with brown. The rest of the cells are shown light blue by counterstaining with haematoxylin.

In this practical session, students will be provided with immunohistochemical preparations, performed by the peroxidase method, for SOX9 protein in mouse embryos at embryonic stage 12.5 (E12.5). Since *Sox9* is a pleiotropic gene, its expression will be detected in different embryonic tissues and organs. The objective of this practical will be to identify the presence or absence of expression of this gene in the different organs and tissues observed in the embryo

sections examined.

## 6.4. WEB RESOURCES

Through the following YouTube link you can access the video-tutorial of the practical.

https://www.youtube.com/playlist?list=PLBa9sJUx0zXWnO2Wu4H6qmJrEOIFNCzaI

## 6.5. QUESTIONS

1. What immunological techniques are currently available to detect the presence of a protein of interest in a tissue?

2. What happens in mammals when the *SRY* gene is mutated? And if it is translocated to the X chromosome?

3. Does a pleiotropic gene have the same function in all tissues where it is expressed? Describe an example that includes SOX9.

*NOTE: The correct nomenclature for mammalian genes is as follows:

Gene names are italicized with capital letters (e.g. *SOX9*), for all species, except for mouse and rat, in which case they are italicized with the first letter capitalized and the others lowercase (e.g. *Sox9*). The names of the corresponding proteins are always written without italics and capitalized (e.g. SOX9).

# 7.- RT-PCR FOR THE STUDY OF GENE EXPRESSION

## 7.1. AIM

After this practice the student should be able to extract RNAs and use them for the study of gene expression via application of the RT-PCR technique.

## 7.2. THEORETICAL BASIS

***Identification of the Anti-Müllerian hormone***

Müllerian ducts (also called paramesonephric ducts) as well as Wolffian ducts (also called mesonephric ducts) are two tubular embryonic structures that show up laterally in the urogenetil primordium during the mammalian embryonic development. In females, Müllerian ducts are differentiated in various structures of the female urogenital tracts: Oviducts (Fallopian tubes in women), the uterus, the cervix and the upper part of the vagina, while the Wolffian ducts atrophy. In males, the Wolffian ducts develop into the epididymis, vas deferens (also called ductus deferens or sperm ducts) and seminal vesicles, while the Müllerian ducts atrophy.

The first evidence on the molecular mechanism responsible of the atrophy of the Müllerian ducts was obtained during the decade 1940-1950 in the works of Alfred José, who xenografted testicular tissue in rabbit foeti (or foetuses) that were previously castrated and observed that the Wolffian ducts differentiated into epididymis, vas deferens and seminal vesicles, while the Müllerian ducts atrophy. He later observed that a testosterone propionate crystal was able to induce the differentiation of the Wolffian ducts in castrated mouse foeti, but did not affect the development of the Müllerian ducts, which produce the oviducts, the uterus, the cervix and the upper vagina. From these experiments he deduced the existence of a diffusible molecule, produced by the testicle, that is not testosterone, and that was responsible for the atrophy of the Müllerian ducts in the male foetus. This molecule was initially called The Müllerian ducts inhibitory substance. However, the identification of such a substance was not easy, and it was not until 1984 when it was characterized. It is currently called the Anti-Müllerian Hormone (AMH), or the Müllerian ducts inhibitory substance (MIS). Further experiments confirmed that the AMH was responsible for the atrophy of the Müllerian ducts. One such experiment identified AMH as the causing agent of *freemartinism*, a phenomenon that had been described in mammals since the beginning of the 20$^{th}$ century. A *freemartin* is an XX individual that has non-functioning ovaries and a reproductive anatomy characterized by female external genitalia and internal genitalia that show a varying degree of phenotypically masculine structures. Cases of *freemartinism* are always produced when an XX individual has an XY twin. Because of this, it was hypothesized that some masculinizing factor travelled from the male foetus to its female twin sister. Thus, several researchers discovered that in cases of *freemartinism*, the female foetus in the uterus has its chorion fused with the chorion of its male twin brother, which causes interconnection between the blood vessels of both foeti. A 1984 study confirmed AMH as the diffusible substance that, in cases of *freemartinism*, travelled from the male foetus to its female twin sister through the interconnected blood vessels**.**

***AMH and testicular development***

In mammals, the expression of the *SRY* gene in pre-Sertoli cells of the primordial XY gonad is

responsible for the differentiation of the up-to-then bipotent gonads into testicles. Several other genes involved in the masculine pathway, such as the *SOX9* and *SF1* genes, will later be activated in the pre-Sertoli cells, which will be differentiated into Sertoli cells. These will undergo a mesenchymo-epithelial transition to produce the testicular ducts. Leydig cells will then differentiate in the mesenchyme surrounding the testicular ducts. Sertoli cells will secrete the AMH that will be transported to the mesenchyme that surrounds the Müllerian ducts. There, the AMH will bind to its receptor, the Anti-Müllerian Hormone Receptor 2 (AMHR2). The union AMH-AMHR2 quick starts a gene activation and repression cascade leading to the atrophy of the Müllerian ducts by cell apoptosis. For their part, the Leydig cells produce testosterone, which is needed for the development of the Wolffian ducts. Sertoli cells do not differentiate in mammalian females. Thus, no AMH is produced and the Müllerian ducts do not atrophy. Like Sertoli cells, Leydig cells do not develop in mammalian females, so no testosterone is produced and the consequent absence of the Wolffian ducts does not allow the development of the male sexual organs (Figure 1).

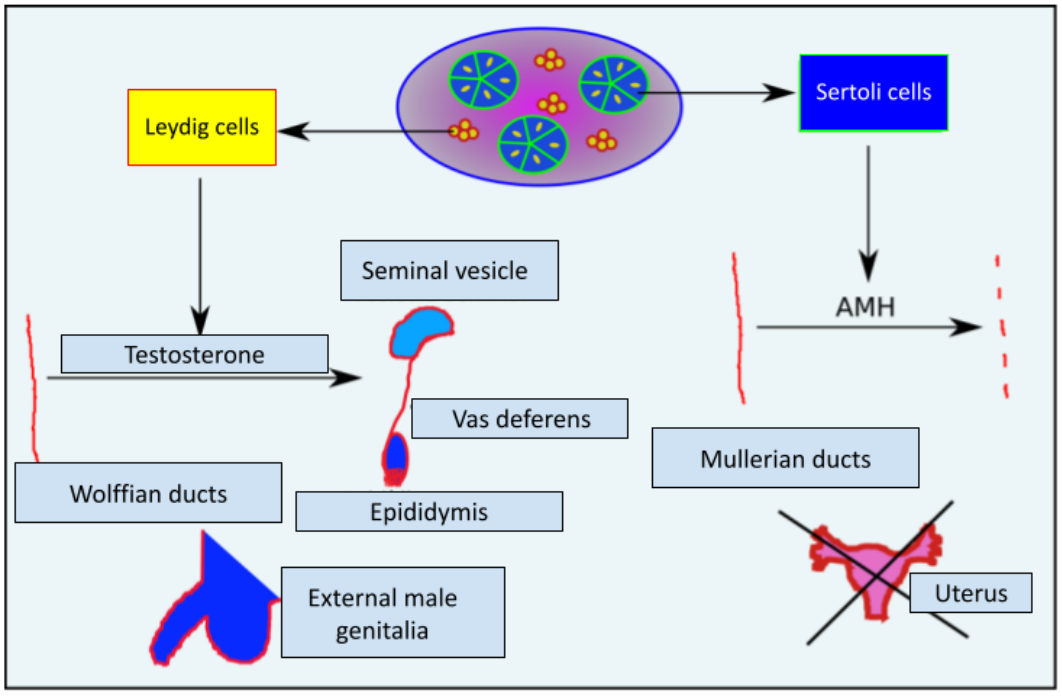



Figure 1: Schematic representation of testicular development. Hormones produced by the testicle include testosterone and AMH

A while after these events have taken place, in mammalian males, the function of Sertoli cells changes during puberty, when they undergo a morphological and functional transformation that prepares them to reinforce the spermatogenic cycle. During this process, known as Sertoli cell maturation, these cells change morphology and enter their mature non-proliferative state. If the maturation process does not take place, germ cells will not enter meiosis and spermatogenesis will not happen.

Expression and production of AMH in the testicle lasts until puberty, coinciding with Sertoli cell maturation, to which its inactivation seems to be associated with the beginning of maturation —although the mechanism behind this process is still unknown. AMH is also expressed in the ovarian granulosa cells, where it has a role in the maturation of the ovarian follicles during the post-natal to the menopause period.

***The AMH gene***

The AMH protein is a gluco-proteic homodimer, of around 140 KD, and which is highly

conserved between species. The carboxy-terminal region of the monomeric AMH protein shows high homology to members of the Transforming Growth Factors TGFβ super-family. Human AMH is encoded by a 2.75 kb gene containing 5 exons that are characterized by their high GC content. The 5' untranslated region contains around 10 nucleotides, whereas the polyadenylation signal is located 90 nucleotides downstream from the stop codon TGA. Two *AMH* mRNA types —differentiated by the polyA tail length— were described in rats. A 2 kb mRNA was detected in the testicular region during the testicular differentiation period. The amount of such mRNA latter keeps decreasing during posterior gestation stadia. Finally, only a 1.8 kb transcript is detected after birth. The promoter of the bovine, mice and rat *AMH* gene has a TATA box and a single transcription start site, located 10 pb downstream from the start codon ATG. The human *AMH* gene, however, has no TATA box or CCAAT sequence, instead showing a functional initiating element (Inr) to which the transcription factor TFII-I binds. The human *AMH* gene promoter contains binding sites for the transcription factors *SOX9*, *SF1* and *GATA* (Figure 2)

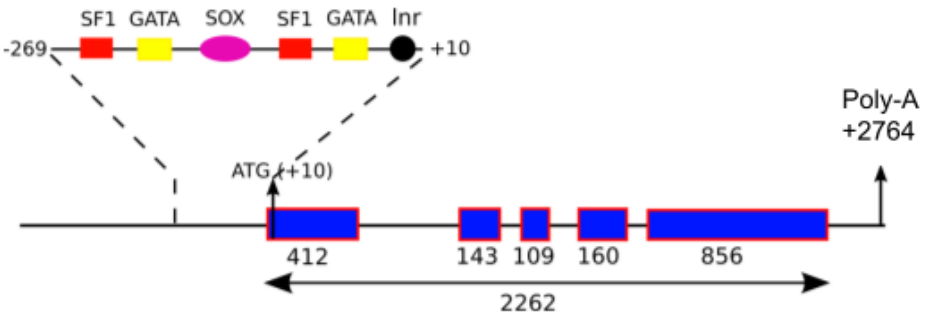



Figure 2: Structure of the human *AMH* gene.

### *AMH Mutations*

Persistent Müllerian ducts syndrome (PMDS) is a rare genetic condition characterized by abnormalities of the reproductive tract. Foeti develop testicles and, at birth, are unambiguously identified as males. However, a more detailed observation shows abnormalities in the genitalia, among which are cryptorchidism, which is a condition whereby one testicle (unilateral cryptorchidism) or both testicles (bilateral cryptorchidism) do not descend from the abdomen to the scrotal sac. The descent of the testicles to the scrotal sac is essential for male fertility, in all likelihood due to the lower temperature in the scrotum than inside the rest of the body —such low temperature seems to be needed for spermatogenesis. In addition, PMDS babies retain some Müllerian ducts-derived structures; such as the uterus and the Fallopian tubes. Because these are internal structures, and unless an older brother is diagnosed with PMDS, surgery is needed to correctly diagnose PMDS. The testicles are normally differentiated and developed in PMDS and, when the cryptorchidism is not prolonged in time, they usually contain germ cells. Nevertheless, the excretory ducts are usually not correctly connected, as they frequently develop aplastic epididymis and vas deferens.

Genetic screening of over 100 families showing PMDS revealed that the mutations responsible for the condition were of the *AMH* gene in 45% of the cases. In 40% of the cases the gene coding for the AMH receptor, *AMHR2*, was what mutated. In both cases, the condition is inherited following an autosomal recessive pattern, and is symptomatic only in males. The cause of the PMDS of the remaining 5% of the cases is still unknown.

## 7.3. METHODOLOGY

In this practical we will see how the *AMH* gene is expressed in testicular tissue. To do so, we will extract total RNAs from testicles and ovaries of neonate mice —the ovaries will be used as negative control. The RNA will be retro-transcribed and the resulting cDNA will be used for a type of polymerase chain reaction (PCR) known as RT-PCR, in order to detect the presence of *AMH* transcripts.

### *RNA extraction*

An Eppendorf tube containing a small sample of testicular or ovarian tissue —previously extracted from mice neonates and frozen at -80ºC— will be provided to each student for RNA extraction. Silica membrane-based RNA extraction columns will be used. Samples will be homogenized then lysed in the presence of a highly denaturing buffer that also contains guanidine thiocyanate, which inactivates RNases and thus prevents RNA degradation. Ethanol will then be added in order to generate the physical-chemical conditions needed for the union of the RNA to the silica membrane of the extraction column. The rest of the cell components of the extraction solution (lysate + ethanol) will flow down through the column, made possible by centrifuging the extraction solution-containing RNA extraction columns. After washing the column using the wash solution, the silica membrane of the RNA extraction column will contain high purity RNAs, while the rest of the components of the lysate and the ethanol will have been filtered and eliminated—as they are not retained by the silica membrane. Hydration of the silica membrane of the RNA extraction column, using RNAse-free water, will allow recovery of the RNA, as this will dissolve in the water and flow down the RNA extraction column into a new tube following centrifugation.

### *Protocol*

- Add 350 µl of the lysis buffer (10 µl β-ME per 1 ml Buffer RLT).
- Homogenize by moving up and down through a 0.8 mm diameter syringe.
- Centrifuge at maximum speed for 5 min.
- Move the upper phase to a new clean Eppendorf tube.
- Add 350 µl of EtOH 70% and mix by inverting the tube a couple of times.
- Transfer the extraction solution (i.e., upper phase + ethanol) to the RNA extraction column.
- Centrifuge for 1 minute at maximum speed.
- Digest the DNA by adding 80 µl of the Dnase I solution and leaving at room temperature for 15 minutes.
- Prepare the agarose gel for checking the RNA quantity and quality —see the agarose gel electrophoresis practice.
- Add 700 µl of the buffer RW1 to the column.
- Centrifuge for 1 minute at maximum speed.
- Discard the flow through.
- Add 500 µl of the RPE buffer to the column.

- Centrifuge for 1 minute at maximum speed.
- Discard the flow through.
- Add 500 µl of the buffer RPE.
- Centrifuge for 2 minutes at maximum speed.
- Discard the flow through.
- Place the column in an RNAse-free clean Eppendorf tube.
- Add 30 µl of RNAse-free water.
- Wait for 1 minute.
- Centrifuge for 1 minute at maximum speed.
- The total RNA will be in the water that should have flowed through the column after centrifugation.

***One-step RT-PCR***

The RT-PCR can be carried out in two ways. One way, known as the two-step RT-PCR, includes an initial retro-transcription of the RNA using universal random hexamers or oligo-dT as primers in order to generate complementary DNA (cDNA) from all the transcripts available in the RNA —the synthesized cDNA will thus be representative of the whole RNA in the tissue from which it was extracted. A PCR reaction is then carried out in a second, new tube using part of the cDNA and the specific primers for the part of the gene to amplify and study (of course plus the PCR buffer, Taq-polymerase, dNTPs and water). Another way of doing RT-PCR, known as the one-step RT-PCR, consists of carrying out both the retro-transcription and the posterior PCR reactions in the same tube and in one go. The retro-transcription in this case uses the specific primer to the sequence of the gene to study —hence that part of the RNA of that gene is the only one to get retro-transcribed into cDNA. Immediately following this, the PCR —which also uses the gene-specific primers— will amplify the cDNA of that part of the gene whose RNA was retro-transcribed earlier.

We will carry out a one-step RT-PCR on part of the *AMH* gene. To do so we will use the following primers, which are specific to two different exons of the *AMH* gene:

AMH-F: 5'-ACC CTT CAA CCA AGC AGA GA-3'

AMH-R: 5'-CCT CAG GCT CCA GGG ACA-3'

We will also use an enzyme mixture, the One step RT-PCR mix, containing both the retro-transcriptase and the DNA polymerase.

We will need to program the thermocycler as follows: 93 ºC for 3 minutes + 35 x (91ºC for 45 seconds + 60 ºC for 60 seconds + 72ºC for 45 seconds) + 72ºC for 5 minutes + 4ºC □.

For the reaction, which we will carry out in a 200 µl microtube, it will be necessary to add the following reagents following the same order as below:

| | | | |
|---|---|---|---|
| 1. | RNase-free $H_20$ | 14 | µl |
| 2. | 5x RT-PCR buffer | 5 | µl |
| 3. | 10 mM AMH-F primer | 1 | µl |
| 4. | 10 mM AMH-R primer | 1 | µl |
| 5. | 10 mM dNTPs | 1 | µl |
| 6. | RNase inhibitor | 1 | µl |
| 7. | Total RNA | 1 | µl |
| 8. | One step RT-PCR mix | 1 | µl |

We then will have to place the tube in a thermocycler and run the cycling One-step AMH RT-PCR program that we prepared and saved beforehand.

Once the reaction has finished, we will be necessary to load 2 µl of the 25 µl RT-PCR reaction (+ 2 µl 10x loading buffer for agarose gel electrophoresis and 16 µl water) in the agarose gel that we have prepared earlier. We can then run the electrophoresis at 100 volts for 20 to 30 minutes —see the agarose gel electrophoresis protocol in the PCR practice part of this handbook.

In our case, and due to the available time-limitation issue, we will run electrophoresis gels using the RT-PCR product made by students of the previous group of this practical—the first group having the RT-PCR product of the last group from last year and the last group leaving the product to be run in agarose gel electrophoresis by the first group of the following year.

If the tissue from which the total RNA was extracted expresses AMH, then, after electrophoresis using the 2 µl RT-PCR reaction sample, a 200 bp DNA band should be visible in the gel —after exposing the gel to UV or blue light.

## 7.4. WEB RESOURCES

A web tutorial of this practice can be accessed on YouTube at:

https://www.youtube.com/playlist?list=PLBa9sJUx0zXWnO2Wu4H6qmJrEOIFNCzaI

## 7.5. EXERCISES AND QUESTIONS

**1.** What is the key factor for successful RNA extraction? Explain why.

**2.** What enzymes does the *One step RT-PCR mix* contain? What is the role of each one in the reaction?

**3.** Why do we use pre-puberty testicular tissue in this practice instead of using adult testicular tissue?

**4.** Why is the retro-transcription needed for a study of gene expression like the one carried out in this practice?