Fundació "la Caixa"

Computational Analysis of DNA Sequences
Gene Prediction Techniques

http://genome.imim.es/courses/laCaixa05/

Introduction

Genes and Genomes

Mining the Genome

Genes and Disease

Gene prediction

Comparative prediction

Genome Bioinformatics Research Lab
IMIM/UPF/CRG

Robert Castelo
Nuria Lopez
Josep F. Abril
Miguel Pignatelli

Roderic Guigó

Computational Analysis of DNA Sequences
Gene Prediction Techniques

Introduction

Overview

This short course, on the analysis of DNA sequences through internet resources, is aimed at those willing to characterize protein coding genes in eukaryotic genomes. First, we examine basic concepts on genomes and gene structure in eukaryotes and learn how to extract genomic information from widely use online databases. Then, we generate our own annotation of protein coding genes on a real genomic sequence and see current limitations of gene prediction programs. Finally, we make use of a state-of-the-art comparative genomics approach to refine our predictions.

Use blue links on your left to follow the course.

Genes and Genomes

The concepts needed to understand eukaryotic genomes and gene structure are revised. Protein and non-protein coding genes, coding and non-coding exons, forward and reverse annotations...

The basics of mRNA processing steps useful to understand eukaryotic gene structure are also outlined. In addition, the analysis of protein sequences at a functional level are briefly introduced.

Mining the Genome

The sequences of multiple genomes and their annotations are now available. For the biologist it is crucial to be able to access this information without having to rely on programming skills. Additionally, the researcher must be able to query the databases with biologically relevant questions.

In this short practical we introduce the Ensmart system as a way to access genomic data through high-level biological queries. We emphasize the importance of how the biological data is structured.

Genes and Disease

Many diseases are caused by mutations in the DNA. In some cases the disease is hereditary These diseases are usually caused by mutations in a single gene that makes the protein it encodes not to function properly or not to function at all. These are called Mendelian diseases or hereditary diseases, and can have different type of inheritance (Dominant, Recessive or X-linked).

In this short practical we introduce the access to several databases that can help us to find information related to diseases, mutations and polymorphisms, as well as the access to some web-servers that predict whether a given gene could be involved in disease based on the existing set of known disease genes.

Gene Prediction

The finding of protein-coding genes on a genome sequence is a complex task. Within millions of non-coding nucleotides, very short stretches of DNA which actually code for a protein (coding exons) lie scattered. This tiny coding fraction, can be unveiled making use of the biological properties and the particular statistical composition found in these regions. Gene prediction programs are computational tools able to find these dispersed coding exons in a sequence and then, to provide the best tentative gene models.

As we will see, this ab initio gene prediction approach is useful but of a limited accuracy.

Comparative Genomics

Comparative genomics is the analysis and comparison of genomes from different species. The purpose is to gain a better understanding of how species have evolved and to determine the function of genes and noncoding regions of the genome. Researchers have learned a great deal about the function of human genes by examining their counterparts in simpler model organisms such as the mouse. Genome researchers look at many different features when comparing genomes: sequence similarity, gene location, the length and number of coding regions (called exons) within genes, the amount of noncoding DNA in each genome, and highly conserved regions maintained in organisms as simple as bacteria and as complex as humans.

Modern gene prediction programs can integrate these comparative data to improve predicted genes.

Internet Resources

ENSEMBL	http://www.ensembl.org
NCBI	http://www.ncbi.nlm.nigh.gov
Golden Path (UCSC)	http://genome.ucsc.edu
geneid	http://genome.imim.es/software/geneid/geneid.html
genscan	http://genes.mit.edu/GENSCAN.html
fgenesh	http://www.softberry.com/berry.phtml?topic=gfind
sgp2	http://genome.imim.es/software/sgp2/sgp2.html
twinscan	http://genes.cs.wustl.edu
slam	http://baboon.math.berkeley.edu/~syntenic/slam.html
This course as a single PDF document	http://genome.imim.es/courses/laCaixa05/laCaixa05.pdf

Acknowledgements

Materials for this course have contributions from:

Josep F. Abril, Enrique Blanco, Charles Chapple, Sergi Castellano, Robert Castelo, Eduardo Eyras, Roderic Guigó, Núria Lòpez, Genís Parra

Course sponsored by

http://genome.imim.es/courses/laCaixa05/

Genes and Genomes

Written by Sergi Castellano

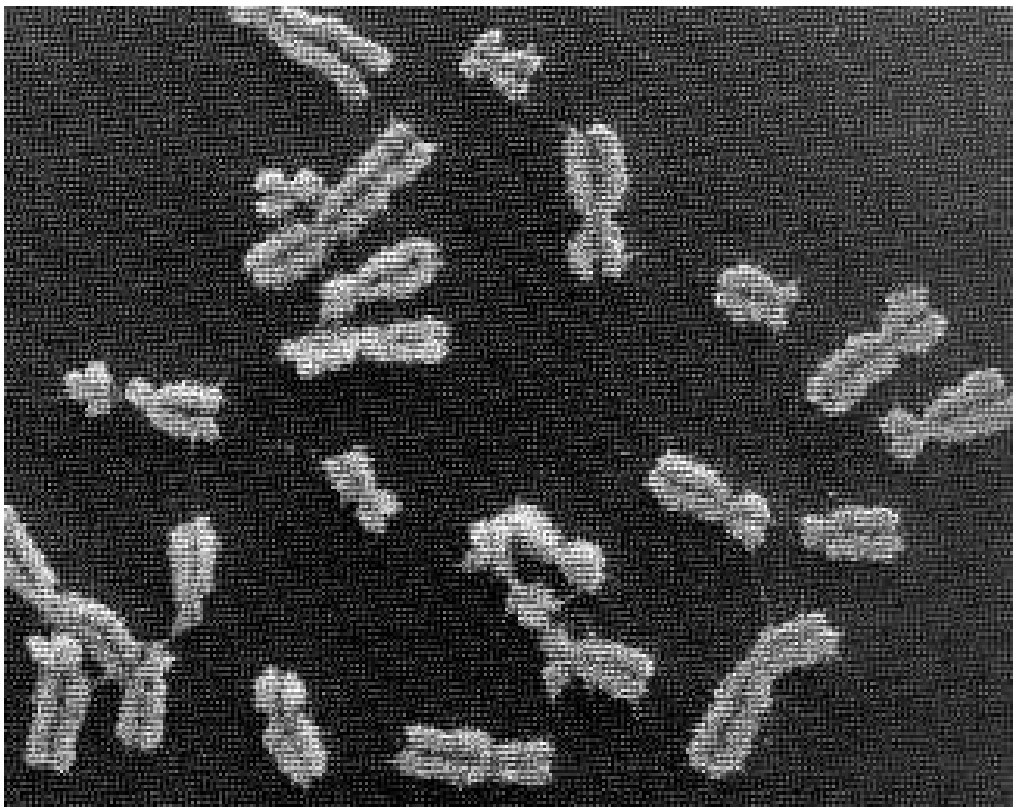
Overview

In what follows, we revise the key concepts regarding the eukaryotic gene and genome structure needed to understand the annotation of genomes. In this context, annotation refers to the description and location of genes and other biologically relevant features of a genomic sequence. Our main goal is to fully comprehend what genome annotation projects offer and, just as important, what they do not yet provide.

In this regard, it's worth noting that current gene prediction programs, among other bioinformatics tools, systematically ignore the complexity of eukaryotic gene structure. Diversity comes from alternatively spliced genes, non-canonical signals (that affect either splicing or translation) and from regions that control gene transcription, promoters, which are not yet well understood. Here, we give and overview to see some of the limitations and future directions in the gene prediction field.

Genomes

The genome is the genetic material of an organism, that is, the total amount of DNA in the cell. In eukaryotes, it is usually organized into a set of chromosomes, which are extremely long chains of DNA that are highly condensed. In the picture below, human DNA is shown packaged into chromosome units (as seen during mitotic metaphase). Note the sister chromatids (that contain identical daughter DNA molecules), centromeres and telomeres.



Human chromosomes

It is time to introduce the three main genome browsers and check which genomes they are serving. In this practical, we will stick to the Ensembl server, but feel free to browse the others later on.

- Ensembl
- Golden Path (UCSC)
- NCBI

Questions:

1. How many species does Ensembl provide annotation for?
2. What is the Trace server? Hint: how are genomes sequenced?
3. What is the EnsMart service? Hint: be patient! we will use in the next lesson
4. What is the difference between the BLAST and SSAHA programs? Hint: think of a strict blast search

Now, select human and make sure you understand the main concepts in this page. Contrary to the view reported in newspapers, the genome sequence and its corresponding annotations are highly dynamic. There are two levels at which data can be updated: 1) sequence level and 2) annotation level.

Questions:

1. How many releases has Ensembl provided annotation for?
2. What are the main reasons that drive the release of genome versions?
3. How many genes, exons and nucleotides does the last genome version have?
4. Do you think that, over the same sequence, Ensembl, NCBI and Golden Path provide the same annotation?
5. What is the Vega annotation? Do you think it is more reliable than the default Ensembl annotation?

DNA

DNA molecules consist of two anti-parallel chains held together by complementary base pairs that form a double helix. This structure is of major importance for computational analysis and, to a great extend, determines how it can be performed:

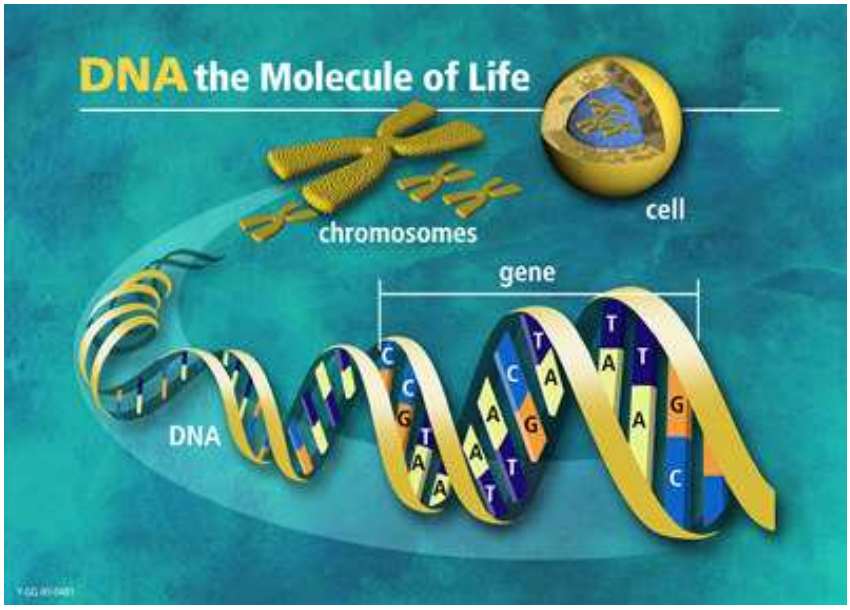
1. The digital nature of the sequence (nucleotides: Adenine, Guanine, Cytosine and Thymine) permits an easy and symbolic computational representation as A, G, C and T letter codes, respectively. It is worth knowing that Uracil (U), which is in place of Thymine in RNA, is also written as T in sequence databases.

2. The double helical nature of DNA, gives us two different sequences to analyze (with distinct encoded information). In order to handle such dual data, the concept of forward or positive (+) and reverse or negative (-) strands and elements (genes, exons, introns...) is introduced. The forward strand, for us, is simply the original sequence we are working on. Note that this concept is meaningless in the cell, so no differences are made between strands in the cell. For example, genes are transcribed from both chains.

3. The complementary nature of the two strands (A-T, G-C base-pairing), permits to work in the computer only with one strand (forward), the other (reverse) being conceptually retrieved when needed. Usually, genome projects only provide one sequence strand (forward) and forward and reverse elements are annotated on it, the latter simply being tagged as reverse. Usually, again, note that the cell has access to both strands at the same time.

4. The anti-parallel nature of the double helix (due to 5' and 3' nucleotide ends), gives polarity to the strands. There is a general agreement to write DNA sequences from 5' to 3' (do not confuse this fact with the forward and reverse concept). The 5' region is also known as the upstream region and, therefore, the 3' region is called the downstream region of the sequence. Be aware that the upstream and downstream concepts have nothing to do with the cis and trans relationship between biological elements, such as trancriptor factors and their DNA binding site (see below).

5. The triplet nature of the genetic code (a codon is 3 nucleotides long), permits to translate any potentially coding sequence in six different ways (frames). Three in the forward strand and 3 more in the reverse sense of the sequence. The finding of the right frame by the ribosome (it already knows the strand) is a challenging process to reproduce computationally.



From DNA to chromosomes

Note that, in this course, we only analyze DNA at the sequence level. So, let's now check it. Please, select the chromosome 21 in the karyotype plot.

Questions:

1. What is the difference between known and novel genes?
2. What does SNP stand for?
3. Why the chromosome short arm (p arm) is shaded? Hint: is there any feature annotated in these region?
4. Is there a correlation between repeats and the GC content?

Now, let's search for a specific gene over the whole genome. In the Find window, lookup the alcohol dehydrogenase gene.

Questions:

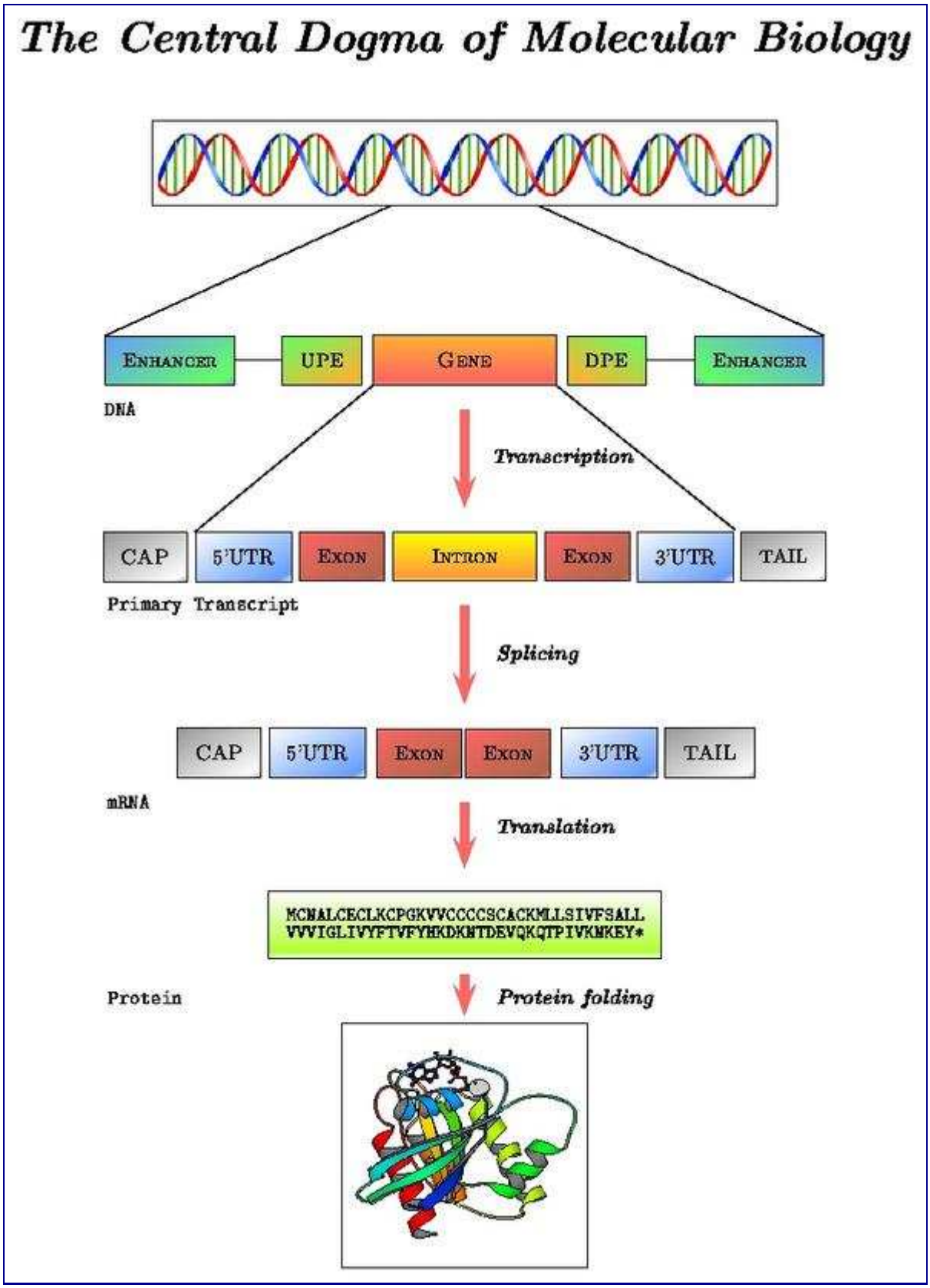
1. How many entries do you get? Why?
2. Do you think it is a good idea to search genes by keywords? hint: try "alcohol dehydrogenase"
3. Select the first entry and display the gene report. In which chromosome is the gene?

However, before analyzing the alcohol dehydrogenase gene, we will take a look at the eukaryotic gene structure, processing and expression (see below).

Gene Expression: from DNA to RNA to Protein

Transcription, splicing and translation are the main processes that account for the expression of protein coding genes. Each step is directed by sequential and structural signals. In what follows, we describe from both the biological and computational point of view, how these sequence motifs are used to go from DNA to RNA to the final protein product.

The schema below, highlights these processing steps:



mRNA processing pathway

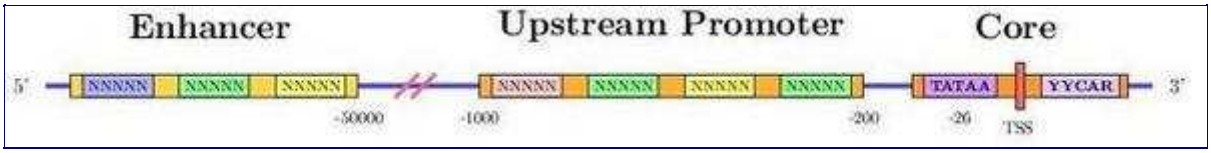
Locate the "Transcript Structure" section in the alcohol dehydrogenase gene report. Note the correspondence between levels of reported data and processing steps:

- Exon information (gene structure on the DNA sequence): sequence before transcription
- Transcript information (mature mRNA sequence): sequence after splicing
- Protein information: sequence after translation

We will follow these links in the order shown, but first, move to the text below for a more precise discussion of the eukaryotic gene structure.

Transcription

Transcription starts when a region upstream of the gene (promoter region) is activated (bound) by transcription factors. These regions, control whether a gene is transcribed from the forward or reverse strand. In any case, the strand which is actually transcribed is called template or sense strand and the other, nonsense or antisense strand.

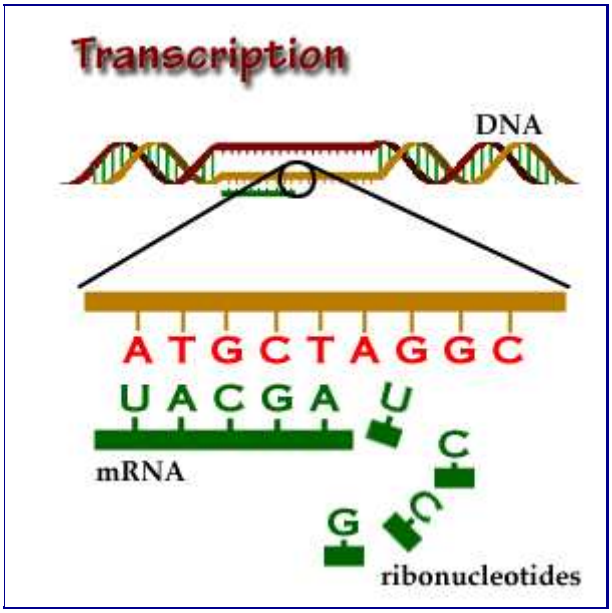


Promoter region

In short, transcription is the copying of DNA (template strand) to RNA (pre-mRNA). However, when analyzing mRNA, cDNA or EST data, bear in mind that the mRNA to be translated is, in sequence, identical to the coding strand (coding here always refers to translation, and not to transcription). That is, the mRNA is transcribed from the strand that has its complementary sequence. In conclusion, when annotating genomes, genes are annotated in relation to their coding strand.

There are three main types of transcript data:

1. mRNA: messenger RNA
2. cDNA: a double-stranded copy, usually a fragment, of an mRNA molecule
3. EST: expressed sequence tag. A short single-pass sequencing of a cDNA clone. It is tipically a fragment from the 5' or the 3' end of the cDNA.



The copy of the template strand

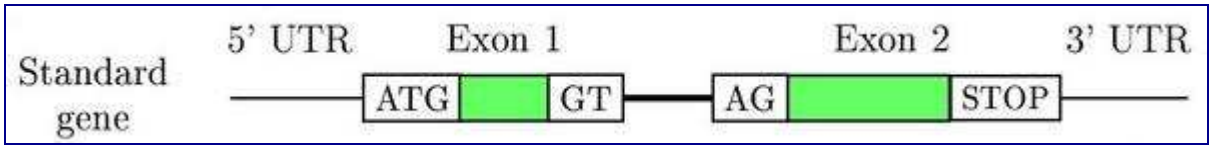
Cis and Trans Elements

The transcription process brings us to two relevant terms in relation to the study of gene regulation: cis and trans elements. A locus is a cis-acting element if it must be on the same DNA molecule in order to have its effect. Transcription factor binding sites are a good example of cis-acting regulatory elements. A locus is trans-acting if it can effect a second locus even when on a different DNA molecule. Transcription factors are a good example of trans-acting regulatory elements. Note, again, that these terms are distinct from the notions of upstream and downstream region explained above.

Another biological example is the so-called trans-splicing, where exons from different transcripts are spliced and joined together. That is, elements from independent sequences end up acting together in the same mature mRNA.

Gene Structure

Eukaryotic genes are short DNA stretches within a genome with a peculiar and discrete structure.



Schematic representation of a two exon eukaryotic gene on a DNA sequence

Gene prediction programs make use of this structure to find genes in a genome. The main characteristics are:

- Coding and non coding exons (UTRs)
- Introns
- Translation start site (ATG)
- Splice sites (GT, donor and AG, acceptor)
- Translation termination site (STOPs: TAG, TGA and TAA)

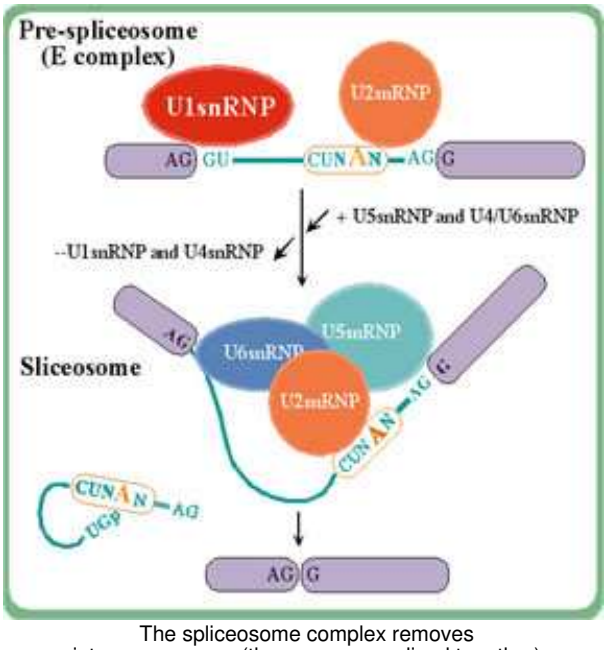
Follow the Exon information link.

Questions:

1. How many exons has this gene?
2. Are the exons completely coding?
3. What is the difference between upstream/downstream, UTRs and intron sequences?
4. Can you spot any common pattern on the start/end of introns?
5. What is the supporting evidence for this gene? Would you trust it?

Splicing

Splicing is an RNA-processing step in which introns in the primary transcript are removed. Splicing signals, GT (donor) and AG (acceptor) in the intron region, are used to delimit exon-intron boundaries, so that exons (coding and non-coding ones) are joined together. In this way, the open reading frame sequence along with the 5' and 3' Untranslated Regions (UTRs) are ready to be processed by the ribosome.



The spliceosome complex removes
intron sequences (the exons are spliced together)

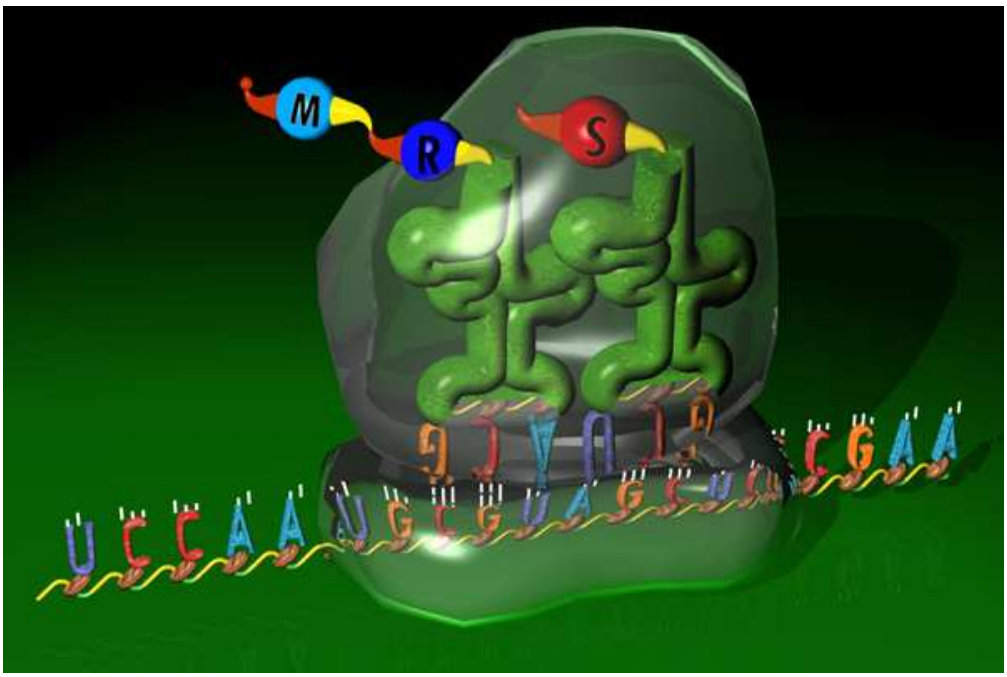
Follow the Transcript information link.

Questions:

1. How many different transcripts does this gene have?
2. How many different proteins does this gene produce?
3. Find out how to highlight the coding and non-coding regions (UTR) in the transcript
4. Can you think of a biological role for UTRs?
5. Check that start and end of the coding region have the right signals.
6. Which of the three stop codons does this mRNA have?
7. How many SNPs are annotated? in the UTR? in the CDS? Is it a synonymous change?

Translation

In translation the mature mRNA sequence is translated into a protein. Again, the ribosomal machinery is guided by several signals along the mRNA sequence to find the right open reading frame (ORF) and to determine where the translation should terminate.



Translation maps RNA to proteins, from a 3 letter code to a 1 letter code

Follow the Protein information link.

Questions:

1. What is the length of the protein?
2. Are all exons of similar length?
3. In which chromosomes are other members of this family?
4. Is there any known protein domain annotated?
5. Can you get the signature for this domain?
6. Take a look at other genes with this domain.
7. What is the function of this protein? Do you think this gene is essential for you?
8. Look for the Gene Ontology in the Gene Report page.

Let's now try to get a deeper insight into the biological role of this gene. We will connect to a couple of web-based resources to learn more about our gene and protein. GeneCards is a database of human genes, their products and their involvement in diseases. Connect to this database and search GeneCards by symbol using the HUGO id corresponding to the adh gene: AKR1A1 (as shown in the Ensembl gene report).

Questions:

1. Check the gene expression in several human tissues. What is an electronic northern?
2. Do you think that, given the number of clones, the electronic northern is significant?
3. Can you list any disorders and mutations in which this gene is involved?
4. What is its cellular location? Hint: go to the GenAtlas database

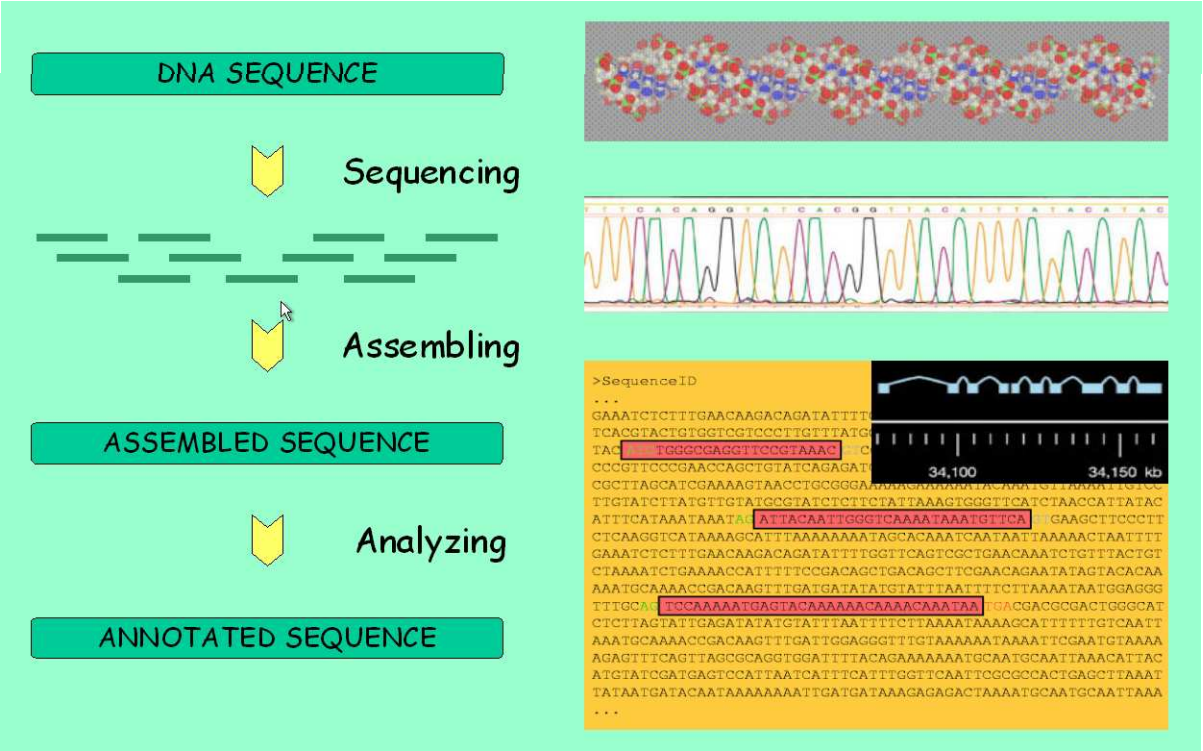
Finally, we will try to get other possible identifiers for this gene in several databases. Connect to GeneLynx and do a quick search in human for the HUGO ID: AKR1A1.

Questions:

1. What is the SwissProt ID?
2. What is the PDB ID? Which method was used to characterize this structure? at which resolution?
3. Can you get a picture of the assumed biological protein at PDB?
4. Can you get the metabolic pathway in which this gene is involved? Hint: go to KEGG pathway

Data Integration

It is about time to fully use the main feature of genome browsers: the ability to display all available information along a genomic region of interest. However, first take a look at the picture below to make sure you understand the pipeline behind these data.



Genome analysis: from sequencing to annotation

We will browse the genomic region of the adh gene that we are working with. From the Gene Report page, follow the link with the genomic location. There are four levels of resolution:

1. Chromosome view
    - In which arm of the chromosome 1 is the adh gene?
2. Overview
    - Is there any novel gene in this region? And pseudogene?
    - What are the mouse and rat synteny tracks?
3. Detailed view
    - Make sure you understand each track.
4. Basepair view
    - Take a look at the region translated in the 3 possible frames for each strand (forward and reverse).

Synteny

The mouse and rat synteny tracks above call for a brief discussion of this concept. Although, historically, synteny means "in the same strand" and syntenic genes are those in such a disposition, but syntenic regions between genomes are understood as regions in which gene order is conserved and, therefore, syntenic genes are putative homologues that have an orthologous, paralogous or even xenologous relation. See below for a short definition of these key but often misinterpreted concepts.

Go to Chromoview in the Ensembl web-site by clicking on a chromosome. From here we can link to Synteny view, which offers a chromosomal view of the synteny between genomes.

Homology: Orthology, Parology and Xenology

Homologous genes are genes that are related through a common evolutionary ancestor. Homology is usually inferred on the basis of sequence similarity but bear in mind that, through random and convergent evolutionary processes, biological sequences can share a reasonable degree of similarity without a true evolutionary relationship. In addition, it is incorrect to say that a pair of related genes are, for example, 80% homologous, because genes are either evolutionarily related or not. On the other hand, one can speak of a percentage of sequence similarity between genes.

This is of importance, for instance, when reading a blast output and deriving evolutionary implications. The score and the e-value are based on the similarity between sequences. However, this does not necessarily ensure a close phylogenetic relationship, although it suggests one.

Orthology, paralogy and xenology are homology subtypes. That is, they define a specific type of relationship between genes over space and time. Read these definitions carefully:

- Orthologous genes are those homologues that are present in different organisms and have evolved from a common ancestral gene by speciation.

- Paralogous genes are present in the same organism or in different organisms and have evolved from a common ancestral gene by a gene duplication event. If this gene duplication event took place before a speciation event, there are paralogous genes in different genomes.
- Xenologues are homologues that originated by an interspecies (horizontal) transfer of the genetic material for one of the homologues.

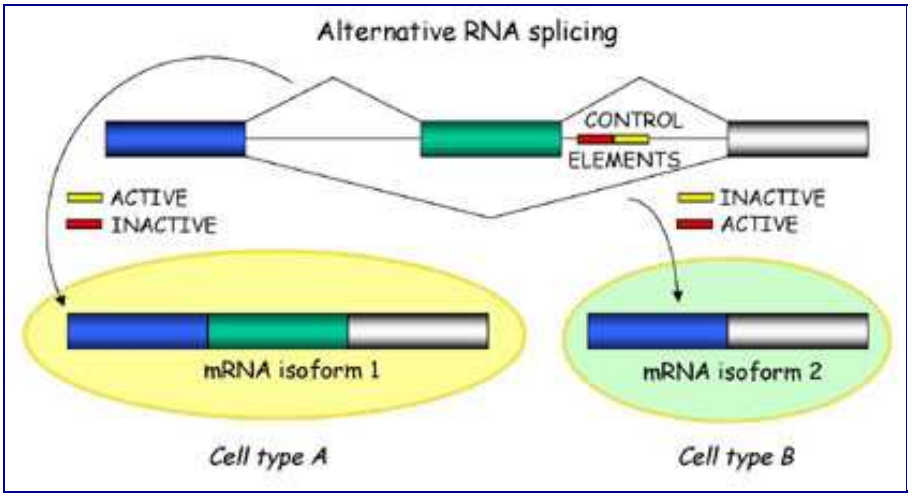
This paper discusses these concepts in more detail.

Alternative Transcription

The transcription start site, can vary in the same gene depending on how the promoter region is activated. This results in a different pre-mRNA and, potentially, a differentially expressed mRNA or even a distinct protein may be achieved.

Alternative Splicing

Briefly, alternative splicing is an important cellular mechanism that leads to temporal and tissue specific expression of unique mRNA products. This is accomplished by the usage of alternative splice sites that result in the differential inclusion of RNA sequences (exons) in the mature mRNA.



Alternative splicing produces unique mRNA products

In general, current gene prediction programs cannot predict alternative mRNAs in a reliable way, unless transcriptional data (mRNAs, cDNAs and ESTs) are available.

Alternative Translation

Translation by the ribosome is a complex process. Sources of variability are:

1. In the same mRNA, alternative translation start sites (ATG) can be used. Furthermore, translation can even start at GTG (valine), TTG (leucine), ATT (isoleucine) but they still code for methionine when they function as an initiator codon.
2. Alternative decoding (recoding):
    - In the ribosome, the meaning of specific codons can be redefined
    - The ribosome can alter the reading frame by switching from one overlapping reading frame to another
    - The ribosome can bypass a stretch of sequence, with or without a change in the reading frame.
3. In the same mRNA, alternative poly-A sites modify the 3' UTR region.

LINUX and Genome Annotation

Try to complete this practical.

Mining the Genome

Written by Eduardo Eyras

Introduction

There is an increasing amount of information available in databases that can be useful to unveil important biological facts. One of the problems in Bioinformatics is how to structure and store this data in such a way that it can be readily used for answering biologically relevant questions. We will explore one of the methods used to make the information closer to the questions posed by the biologists.

We know now how to find information about our favourite gene, like the number of exons, transcripts, the functional domains, etc. We also know how to view some of the properties of the genome in a region of interest, e.g.: sequence conservation with other genomes, synteny, etc. However, can we somehow link all this to ask questions like:

Give me all genes in human with a given Pfam domain and an ortholog gene in mouse and rat, with a sequence similarity greater than 70%.

Ensmart

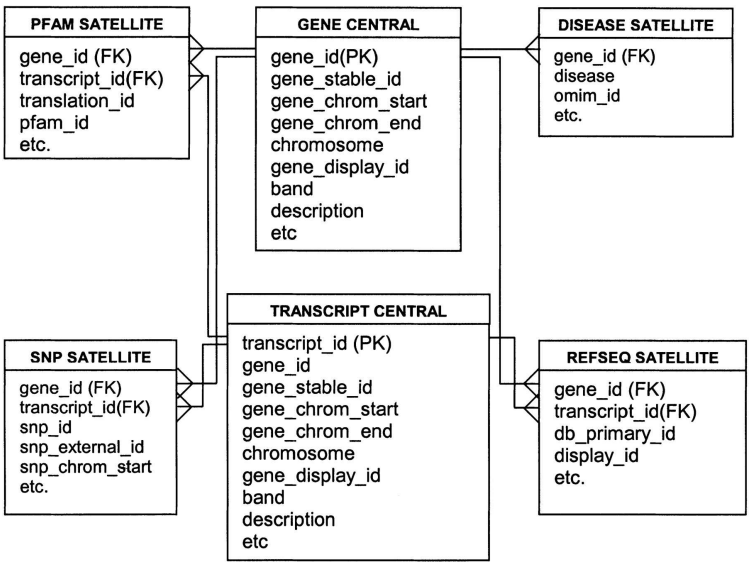
We will use the Ensmart browser:

- www.ensembl.org/EnsMart

Biological databases are usually designed such that the storage and update of information is optimized, and have structures (tables, files) that are very specific to the data at hand. Complex queries are therefore computationally expensive as they require a large amount of analyses and computations. This also often requires specialized software (e.g. Ensembl and UCSC browsers).

Ensmart provides a way of organizing the data such that it is optimized for querying. The computations are still necessary, but the results are stored in such a way, that we are able to obtain a very fast answer to queries that involve a non-direct relationships between attributes. Ensmart is built with pre-computed data and it is in the way the results are stored that we gain usability.

The Ensmart database structure is adapted from a star schema. There are one or more central tables (gene, transcript, SNP) and these are linked to a number of satellite tables containing the attributes. The central table is the source for the constraints and the satellite is the source for the attributes:



You can read more about it here: http://www.genome.org/cgi/content/full/14/1/160

We are going to illustrate the use of Ensmart with an example. We have seen previously (see previous section) that the ADH gene contains a domain from the Aldo/Keto reductase family with Pfam ID PF00248.

Let's imagine that we are interested in promoters and we want to study the upstream region of human genes with this domain. Additionally we would like to concentrate on genes that have an ortholog in mouse. However, we do not want all of them, we only want to look at those that have non-synonymous SNPs in the coding region of the gene.

We would like to obtain two pieces of information for these genes:

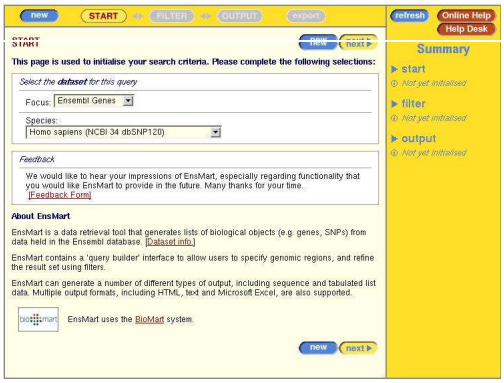
- The upstream genomic sequence.
- The pattern of expression.

Ensmart is based on a focus, which is the entity about which we ask the questions, and a series of satellites sets of data, which hold the attributes and which are used to do to operations:

- Filtering: Eliminates or includes entities according to some attribute filter.
- Output: It selects a set of attributes from the select items for output.

Data Mining

Open the browser: www.ensembl.org/EnsMart



How many foci can you see?

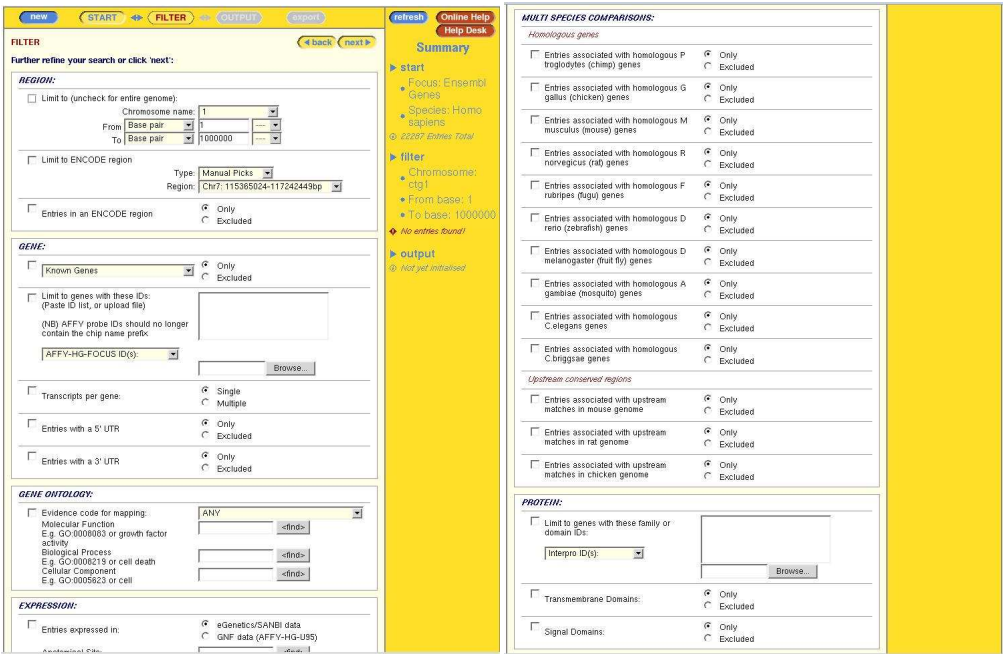
How many species can we choose from? Is there anything different in 'species' with respect to the species you can see in www.ensembl.org?

Select focus: human. Note the versions: NCBI34 dbSNP120.

Once we have selected the focus, we need to specify the filters. A wide range of filters can be applied in any combination.

Question: Explore the possible filters

- Region
- Gene
- Gene Ontology
- Expression
- Multispecies comparisons
- Protein
- SNP



Question: Select the filters according to what we want to answer:

- Make sure we search all human genes
- Select for mouse orthology
- How do you specify the Pfam domain PF00248?
- How do you specify the type of SNP?

Once we have applied the filter, we can choose which output to generate. Press the 'next' button. Note the update on the right-hand side.

Question: Explore the four possible types of 'output':

- Features
- SNPs
- Structures
- Sequences

Select the 'sequence' output. Choose 5' end sequence.

Question: Why can we choose between gene and transcript to select the sequences?

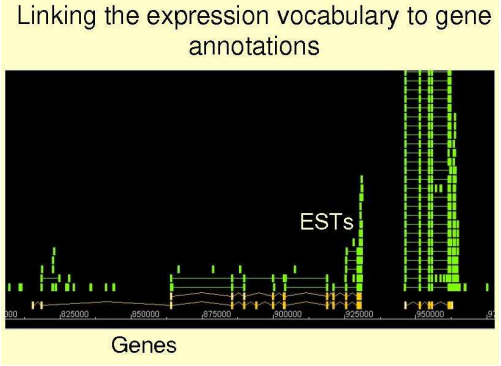
Select the upstream region of the gene. You should obtain an output like this:

```
>ENSG00000165568.4 assembly=NCBI34|chr=10|strand=forward|bases 4821426 to 4822425|region upstream of gene only
CTCCCCTGATGGCCAGCACTGAAGACCCAGGCAAGGAACCTAGAAACAAAGCCCTCATCTGGGTGTGGGTGTCCTCAAGG
CAGTAGGACTCCCAGGGCTGAGGGGGGCAATGAAGGGGGAGCTGTAAGCTCCAGGAGAGATAAGAGGGGCGTCGGAAGGC
TCCCTTGACCCCTCTTTCCCTCCACTGGCCCTGGGGGAGCCCAGTCCACTCATAAGGGGGGTGTCCAGTCCACCCCATCC
...
```

We can go back in the browser to select the expression patter of these genes. In the Features section select one of the sources of expression.

Question:

- Look at how the expression information is presented. It is in a tree strucuture.
- Do you know how this information is obtained?
- How are the external links obtained in general?



Exercise

Carry out the following searches:

- Rat orthologs of human genes annotated as involved in disease(s) and expressed in brain.
- All validated human SNPs on chromosome 2 between 100-200Mb with a minor allele frequency >10% in East Asia population samples.
- Genomic location and of all mouse and Fugu homologs of all the human genes that have transmembrane domains, are expressed in cardiovascular system and have non-synonymous SNPs.

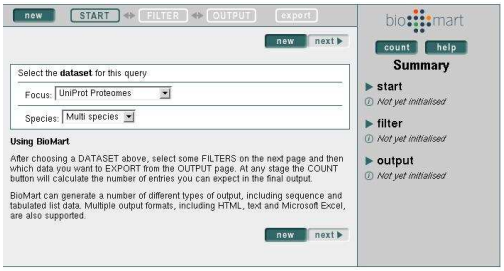
Other Examples

Other examples of data you can retrieve with Ensmart:

- Coding SNPs for all novel kinases
- Genes on chromosome 5 expressed in liver
- Sequences for all Ensembl genes mapped to some microarray probe (e.g. U95A).
- Disease related genes between two markers (e.g. D10S255 and D10S259).
- Transmembrane proteins with an Ig-MHC domain (IPR003006) on chromosome 2.
- Genes with associated coding SNPs on chromosomal band 5q35.3.

Biomart

As the idea behind Ensmart is data-driven rather than data-dependent, the same approach is extensible to other data sources. This is achieved by collecting the raw data from a database and generating the tables in a star schema. Once the satellite tables are defined the structure and dependencies can be automatically generated. Any additional domain specific information can be added in the form of external lookup tables.

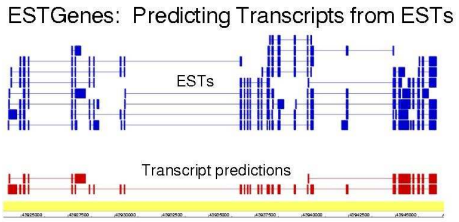


Go to the web page of Biomart

We can see that so far it has been applied to other databases:

- Vega genes (manually annotated human genes)
- Uniprot (all known proteins)
- Molecular Structure Database,
- ESTGenes (Genes built from ESTs by Ensembl)
- dbSNP (Database of single nucleotide polymorphysms)

This picture illustrates what ESTGenes are:



Merge ESTs according to splicing structure compatibility

Click on the BioMart logo.

Exercise

- Get the publication information from those proteins in Arabidopsis thaliana that have an entry in the Protein Data Bank (PDB).

Hint: Obtain first from Uniprot the "Uniprot Accession Identifiers" of those proteins with an associated PDB ID. Then search for the entries in the MSD database corresponding to these accession IDs.

Questions:

- Do you see in Uniprot or MSD the same possible outputs as before with Ensembl? Why?
- Note that in some cases the Focus is a multispecies database. Can you see any relation to the way the data is originally stored/generated?
- Can you give an example of other type of data (Biological or not) that could be put into this system?

To Sum Up

Some take-home messages:

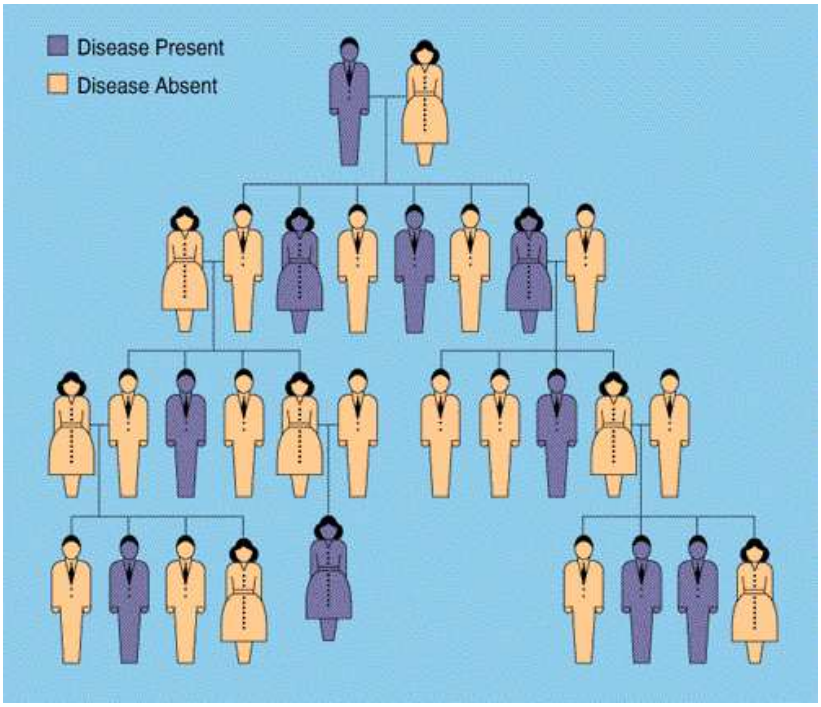
- It is very important how you structure and organize the data to make it useful for querying.
- Use formats and terminologies as standard as possible. It is will make your tool/method more user-friendly.
- It is all actually based on pre-computed data. If the data is bad or is wrongly computed, all the rest is useless.

Genes and Disease

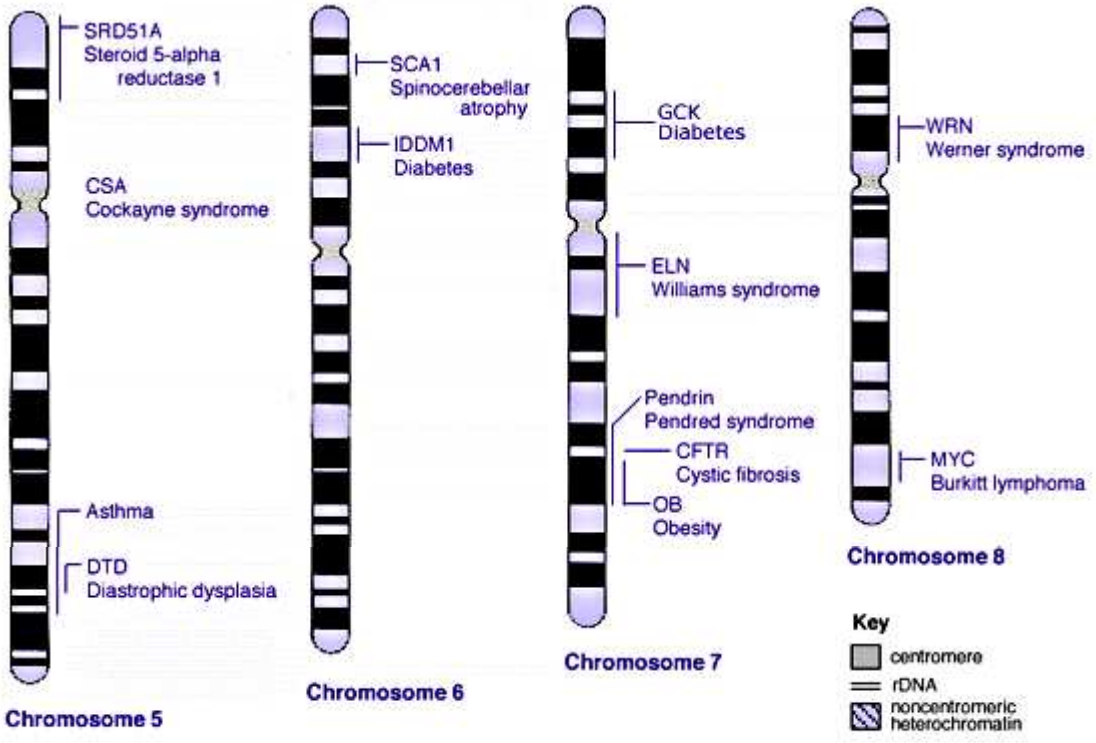
Written by Núria Lòpez

Introduction

Many diseases are caused by mutations in the DNA. In some cases the disease is hereditary (i.e. it is inherited through different generations in a family). These diseases are usually caused by mutations in a single gene that makes the protein it encodes not to function properly or not to function at all. These are called Mendelian diseases or hereditary diseases, and can have different type of inheritance (Dominant, Recessive or X-linked).



Some examples of these diseases are: Cystic fibrosis, neurofibromatosis, some types of deafness, Duchene muscular dystrophy, some types of diabetes...
Other types of diseases are also due to mutations or alterations in the DNA but are not Mendelian diseases, for example, cancer is caused by mutations in some key genes (onco-genes and tumour suppressor genes), although this mutations are usually not inherited but acquired during the life time (somatic mutations). Complex diseases, for example, diabetes, atherosclerosis, cancer, and neurodegeneration, are also due to variations in the DNA, but in that cases is not a single mutation in a particular gene that cause the disease but a number of factors, including several modifications in the DNA in conjunction with environmental factors that determine the risk to suffer the disease.
It is important to know the genes implicated in hereditary diseases, since this knowledge can lead to improvements in disease prevention, diagnosis and treatment. The usual process followed to identify disease genes and mutations is the linkage analysis in DNA samples of affected and non-affected families members with the disease, that is to analyse markers along the genome to identify a region linked to the disease. Once the region is identified, it will be necessary to do mutation analysis to detect the causative mutation leading to the disease.



There is big amount of information available in databases that can be useful to find genes and mutations involved in diseases.

We may want to respond questions like:

Do we know the causative gene for a particular disease?

Which disease mutations have been found in a particular gene?

What genes are in a particular region that we have found linked to the disease we are studying?

And which of all these genes in the linked region is the more likely to be causing the disease of interest?

Databases and Catalogs

Several databases help us to find information related to diseases, mutations and polymorphisms.

- Online Mendelian Inheritance in Man (OMIM)
- The Human Gene Mutation Database (HGMD)
- Single Nucleotide Polymorphisms Database (dbSNP)
- NCBI Entrez Gene

We can find information about a particular disease for which we already know the gene:

Question: Which type of inheritance follows Cystic Fibrosis? Which symptoms characterize this disease?

Note: use OMIM

Question: Which is the causative gene of Cystic Fibrosis?

Note: you can use OMIM or NCBI Entrez Gene

Question: How many mutations have been found in this gene?

Note: use HGMD

For some disease we don't know yet the causative gene, but the linkage analysis has been done and we know the region of the genome that is linked to that disease.

Question: Find the genomic localization of Nephropathy-hypertension.

Note: use NCBI Entrez Gene

Prediction of disease genes

Imagine we can do some computational analysis that will predict which is the causative gene for the disease we are studying. We would then skip a lot of work on mutation analysis. There have been some recent works on that direction. These works exploit the big amount of data we have currently on genes and diseases to predict which other genes are the more likely to cause other diseases.

- Disease Gene Prediction
- Prospect
- Candidate Gene to Inherited Diseases (G2D)
- The GeneSeeker

Question: Find the candidate genes for Nephropathy-hypertension using the four servers.

Gene Prediction

Written by Enrique Blanco

Overview

In this section we use several gene prediction programs on a particular genomic DNA sequence. For each of these programs we obtain a prediction of a candidate gene and we will analyze the differences between predictions and the annotation of the real gene.

The programs we are going to use are `geneid`, `genscan` and `fgenesh`, which are available through a web interface. In these, and in many other tools in the web, we access a form where we can paste, or submit, the sequence we want to analyze, and then we press a button in the form that starts the computing process in some computer where the program runs. Once this process is finished, we get a new page in our browser with the results, which in this case should be a predicted gene.

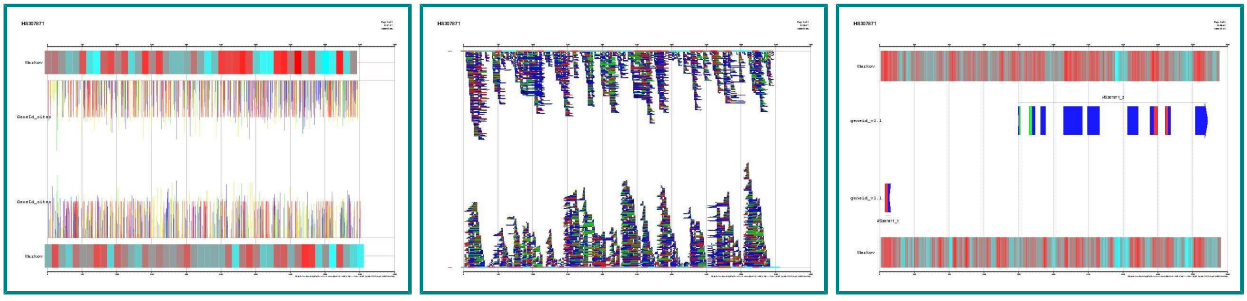
A genomic DNA sequence

We are going to work with the sequence HS307871, which is stored in FASTA format. This sequence contains one gene, annotated in the following EMBL and NCBI records. Try to identify in these records the different pieces of information related to the annotation of the gene.

`geneid`

In order to use `geneid` follow these steps:

1. Connect to the `geneid` server by following this link.
2. Paste the DNA sequence.
3. Select organism (human).
4. Run `geneid` with different (output) parameters:
   - Searching signals: Select acceptors, donors, start and stop codons. For each type of signal, try to find the real ones.
   - Searching exons: Select All exons and try to find the real ones.
   - Finding genes: You do not need to select any option (default behavior).
5. Compare the prediction with the real annotation.
   - By taking a look to the graphical representation of the predicted sites and exons.
   - By inspection of the output and the EMBL/NCBI record.
   - By taking a look to the graphical representation of both, the output and the EMBL/NCBI annotation in this link.
6. Improve the prediction from some confirmed evidence.
   - Below the text box where we pasted the DNA sequence, we can find a text box where we can paste evidences, which should consist of one or more exons (in GFF format) that are, e.g., experimentally confirmed.
   - In this case we are going to paste as evidence the first exon which has not been predicted. Select and copy the GFF line corresponding to this exon contained in this file.
   - Paste the line into the evidences text box, and run again geneid on the sequence.
   - Compare the result with the real annotation. What has changed from the previous prediction?



Signal, exons and genes predicted by `geneid` in the sequence HS307871

genscan

In order to use genscan follow these steps:

1. Connect to the genscan server by following this link.
2. Paste the DNA sequence.
3. Select organism (vertebrate).
4. Compare the prediction with the real annotation.
   - By inspection of the output and the EMBL/NCBI record.
   - By taking a look to the graphical representation of both, the output and the EMBL/NCBI annotation in this link.

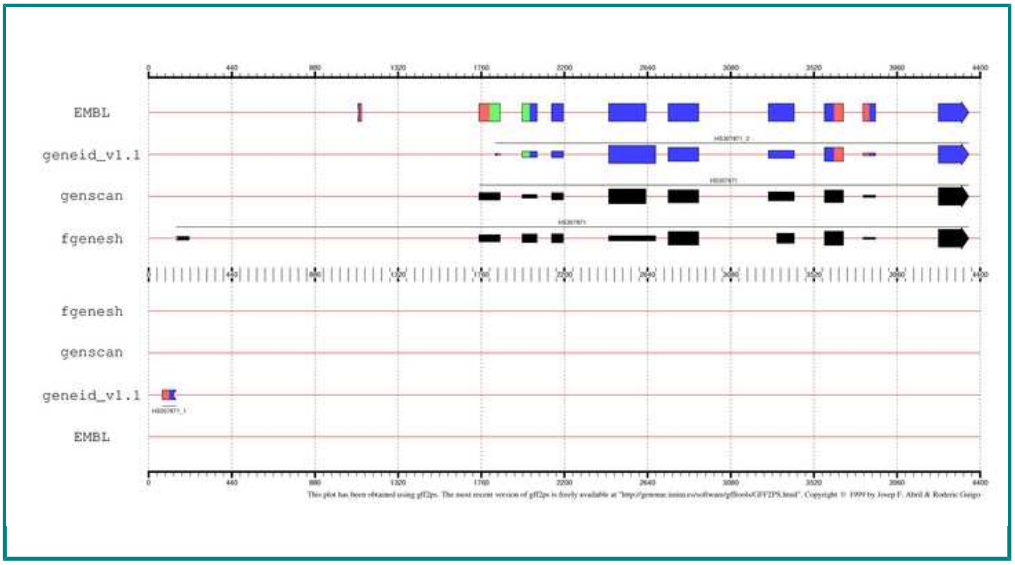
fgenesh

In order to use fgenesh follow these steps:

1. Connect to the fgenesh server by following this link.
2. Paste the DNA sequence.
3. Select organism (human).
4. Compare the prediction with the real annotation.
   - By inspection of the output and the EMBL/NCBI record.
   - By taking a look to the graphical representation of both, the output and the EMBL/NCBI annotation in this link.

Current annotations in the genomic DNA sequence

We can see the annotation of the gene together with the three predicted genes by geneid, genscan and fgenesh by following this link.

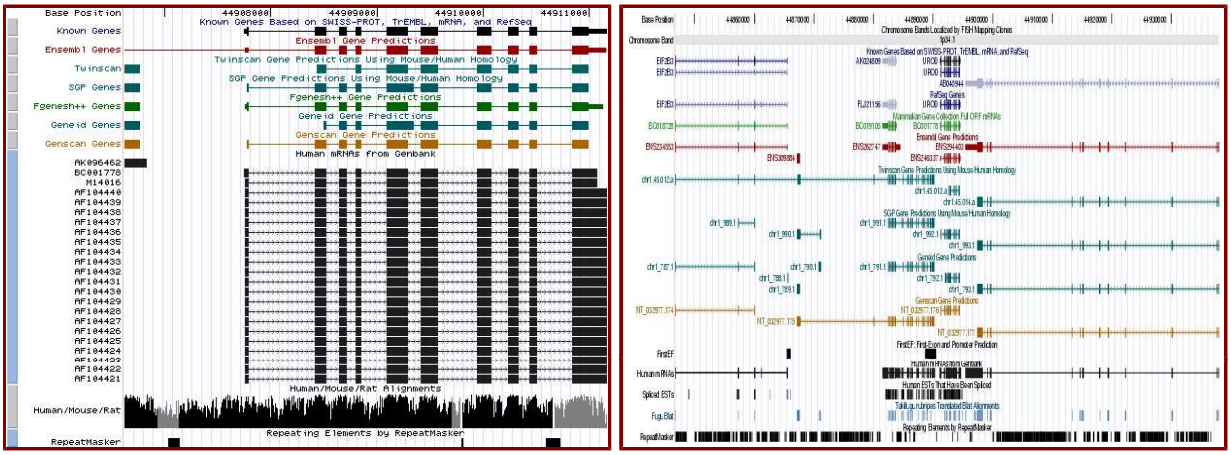


EMBL annotation and genes predicted by geneid, GENSCAN and FGENESH in the sequence HS307871

Go to the page where we saw the NCBI record, click on the link CDS, and, next to the Display button, unroll the menu box and select the display option FASTA. Now press the button Display, and we will obtain the protein-coding DNA sequence of this gene in FASTA format.

Select the entire sequence (first line is not necessary) and go to the UCSC genome BLAT search by following this link. In the big text box, paste the coding sequence we just copied, and press the Submit button on the top-right corner of this page.

The result is a single match, where we find two links, browser and details. Visit first the details link and try to understand the the information provided there. Then go backwards and visit the browser link where we will see where this gene is located within the Human genome, as well as other annotated information as EST spliced alignments, etc.



Left) UCSC genome browser representation of the region containing the gene *uroporphyrinogen decarboxylase (URO-D)*

Right) UCSC genome browser representation of the contex (100Kbps) region around the gene *uroporphyrinogen decarboxylase (URO-D).*

Comparative Genomics and Gene Finding

Written by Genís Parra and Josep Francesc Abril

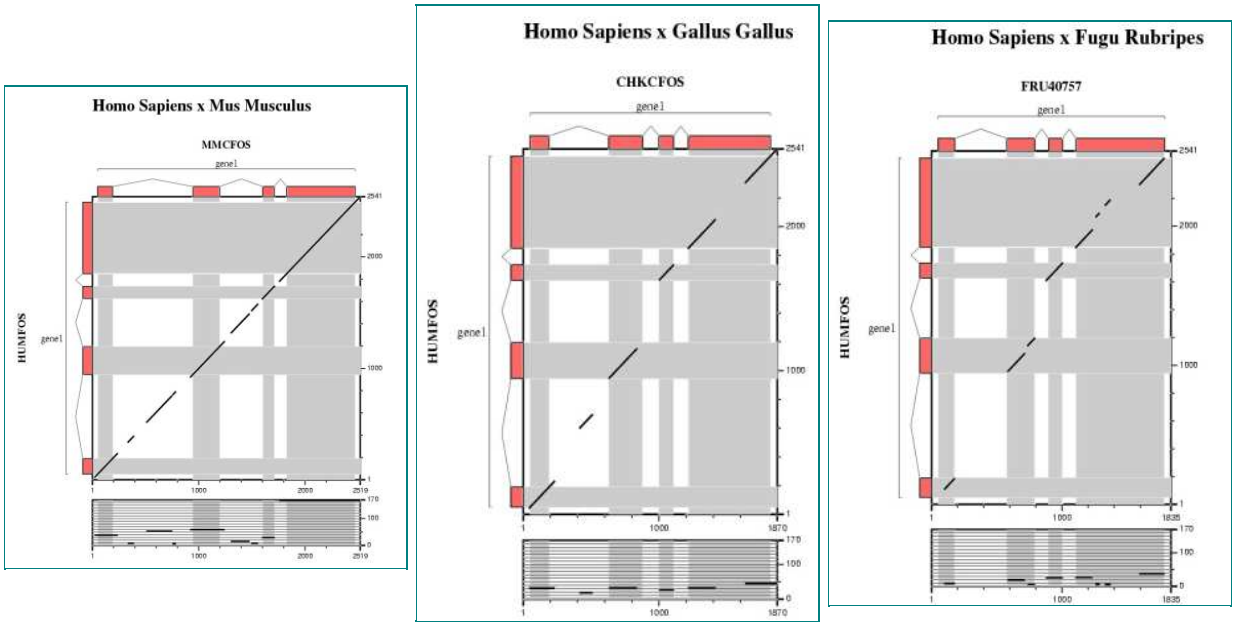
Introduction

"Everything is the result of comparisons"
J-F. Champollion at the time in Grenoble in a letter to his brother -- April 1818.

The Rosetta Stone (left figure) was a key piece to decipher the ancient Egyptian hieroglyphics. The text appears in form of hieroglyphs (script of the official and religious texts), of Demotic (everyday Egyptian script), and in Greek.

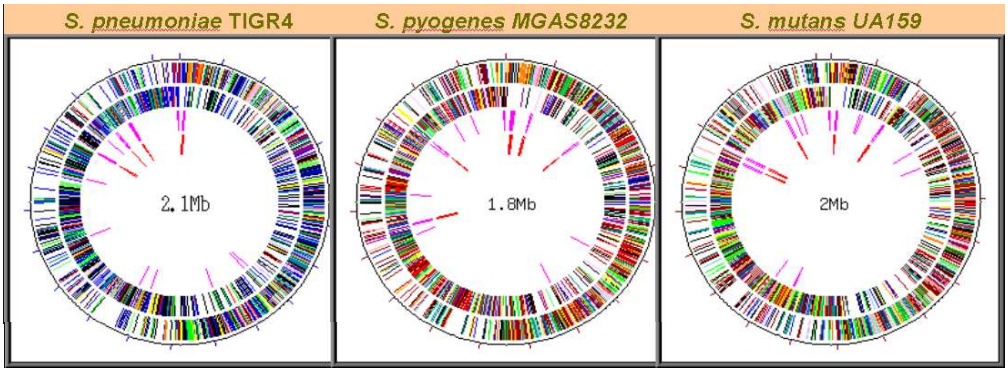
Comparative genomics is the analysis and comparison of genomes from different species. The purpose is to gain a better understanding of how species have evolved and to determine the function of genes and non-coding regions of the genome.

Researchers have learned a great deal about the function of human genes by examining their counterparts in simpler model organisms such as the mouse. Genome researchers look at many different features when comparing genomes: sequence similarity, gene location, the length and number of coding regions (called exons) within genes, the amount of non-coding DNA in each genome, and highly conserved regions maintained in organisms as simple as bacteria and as complex as humans.



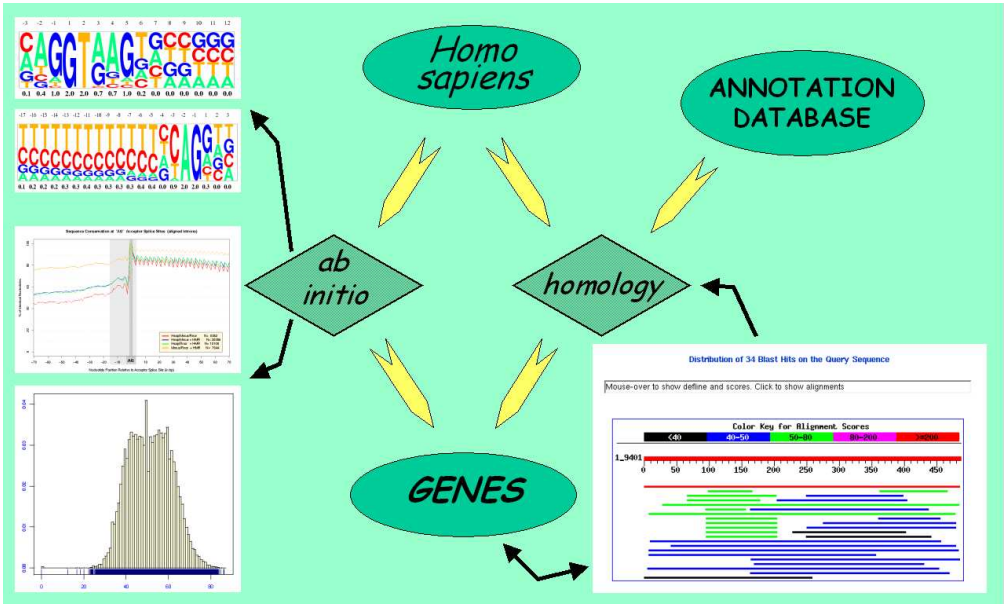
Comparative analysis of the mouse, chicken and fugu orthologs for the human FOS gene.

On the other hand, finding similarities is not as much important as finding differences. The comparative approach also points out those features which are unique for a given phylogenetic group or particularly a species. Species specific functions can be involved in, for instance, pathogenicity, resistance to antibiotics, and so on, but also will result on more complex phenotypic characters such as the human ability to speak.



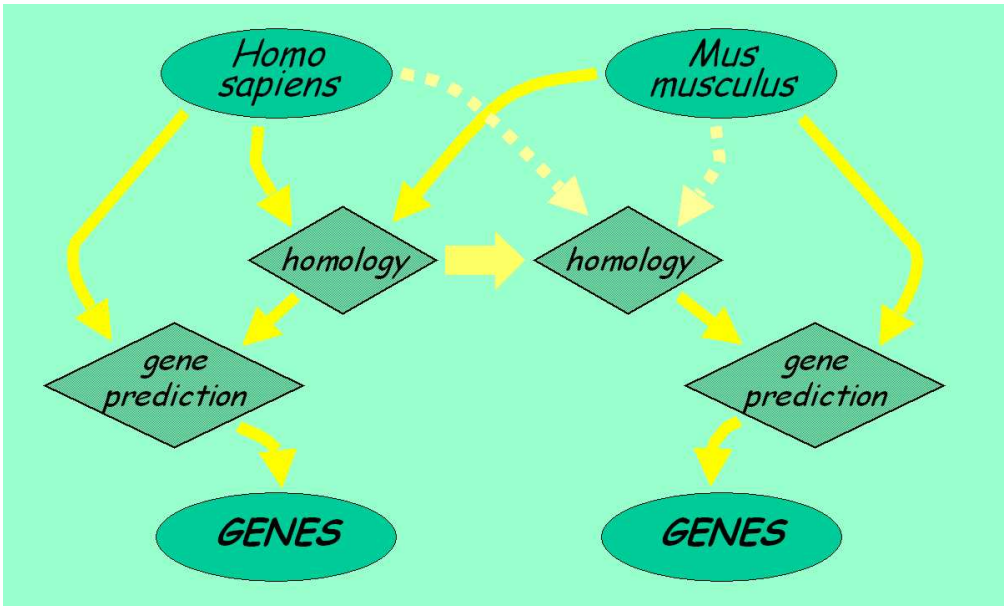
Visualizing the complete catalog of genes for three Streptococcus species.

Ab initio gene finding programs integrated different measures obtained from the raw genomic sequences, such as G+C content, periodicity of coding regions, exon bounds signal detection, etc. The obvious next step was to include homology from the growing annotation databases like SWISSPROT and EMBL/GenBank.



General schema of an ab initio gene prediction tool.

Modern gene prediction programs can integrate the data obtained from the comparison of two genomes to improve the exonic structure of already predicted genes. Furthermore, novel genes not represented in the annotation databases can be found as well.



The comparative gene prediction approach can produce a complete gene set for both species.

Before we start this practical on comparative gene-finding tools lets review some important concepts. Any of the following links will help us to illustrate some of them: Ensembl Synteny View of Chromosome 7, NCBI Homology Map of Chromosome 22.

Conserved
: Derived from a common ancestor and retained in contemporary related species.

Homologs
: Features in species being compared that are similar because they are ancestrally related.

Orthologs
: Homologous features that separated because of a speciation event, they derive from the same gene in the last common ancestor.

Paralogs
: Homologous features that separated because of duplication events.

Synteny
: The property of being on the same chromosome.

Homology Blocks
: Also defined as Conserved Synteny, occurs when the orthologs of genes that are on the same chromosome in one species are also on the same chromosome in the comparison species.

Conserved Segments
: Also known as Conserved Linkages, is a special case of the conserved synteny in which the order of multiple orthologous genes is the same in the compared species.

Overview

In this section we will run several ab initio gene prediction programs on a particular genomic DNA sequence and we will compare the results against predicted genes from a gene finding program that uses genomic homology. For each of these programs we will obtain a prediction of a candidate gene and we will analyze the differences between predictions and the annotation of the real gene both in human and mouse.

The programs we are going to use are `geneid`, `genscan` and `fgenesh`, which have been used in the previous practical exercise. `blast` will be used to compare human and mouse sequences. Then, `sgp2` (syntenic gene prediction tool) will predict genes taking into account the homology found between these two species. Finally, we will take a look at comparative tools that are based on the sequence alignment rather than on the gene prediction paradigm.

A genomic DNA sequence

We are going to work with this Human sequence, which is stored in FASTA format. We also provide the homologous region in the mouse genome in this Mouse sequence.

Ab initio gene finding

In the first approach, we will use all the ab initio tools from the Gene Prediction section and compare the result of the three programs. You could open a simple word processor and paste the results of each gene-finding program in order to compare the coordinates of the predicted exons.

Step 1.- Analyzing the Human sequence.

In order to use `geneid` follow these steps:

1. Connect to the `geneid` server by following this link.
2. Paste the DNA sequence.
3. Select organism (human).
4. Finding genes: You do not need to select any option (default behavior).

In order to use `genscan` follow these steps:

1. Connect to the `genscan` server by following this link.
2. Paste the DNA sequence.
3. Select organism (vertebrate).
4. Run gene predictions.

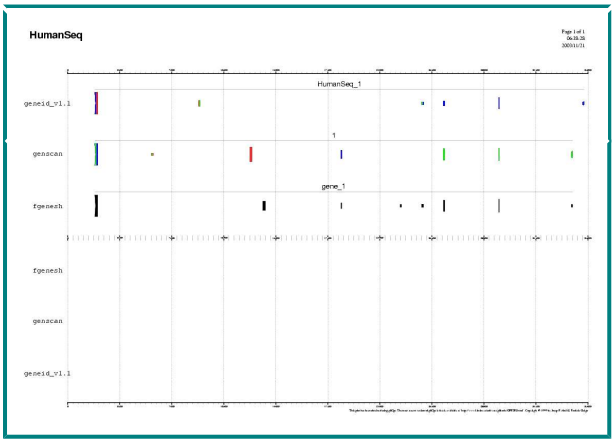
In order to use `fgenesh` follow these steps:

1. Connect to the `fgenesh` server by following this link.
2. Paste the DNA sequence.
3. Select organism (human).
4. Run gene prediction.

Some questions:

- Do the ab initio gene finding programs predict the same exonic structure ?
- How many common exons the different programs have predicted ?

Here you can find a plot with the predictions of the ab initio gene finding tools in the human genome.

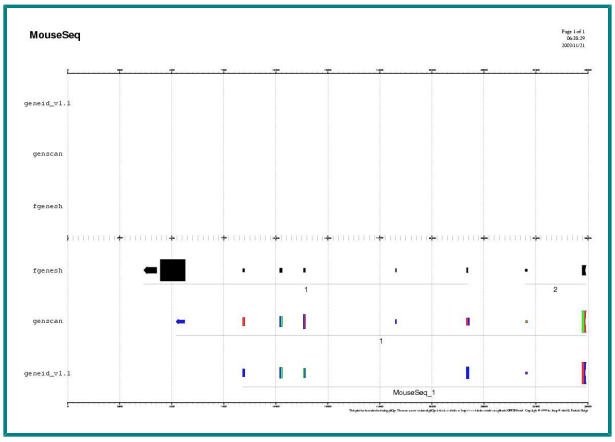


Step 2.- Now, make the prediction in the Mouse sequence, with all the ab initio programs.

Some questions:

- Do the ab initio gene finding programs predict the same exonic structure ?
- How many common exons the different programs have predicted ?

Here you can find a plot with the predictions of the ab initio gene finding tools in the mouse genome.



Do you find any common pattern between human and mouse prediction ?

Comparing human and mouse sequences

In this section we will compare the human and the homologous mouse sequence using `blastn` and `tblastx` on the NCBI's server. `Blastn` compares a nucleotide query sequence against a nucleotide sequence database and `tblastx` compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database.

Some questions:

- Where do you expect to find more similarity regions hits ?
- Which program do you think is more sensitive ?

```
Met Tyr Iso Ser Pro Asp
ATG TAT ATC TCT CCC GAC
||| | |  ||     ||  | |
ATG TTT CTC AGC CCT GCC
Met Phe Leu Ser Pro Ala
                        Amino Acid Level Score
+6  +3  +2  +4  +9  -2     Blosum45    :  +22
+++ +-+ -++ --- ++- +-+  Match/Mismatch :   +4
                        Nucleotide Level Score
```

In order to use `blastn` follow these steps:

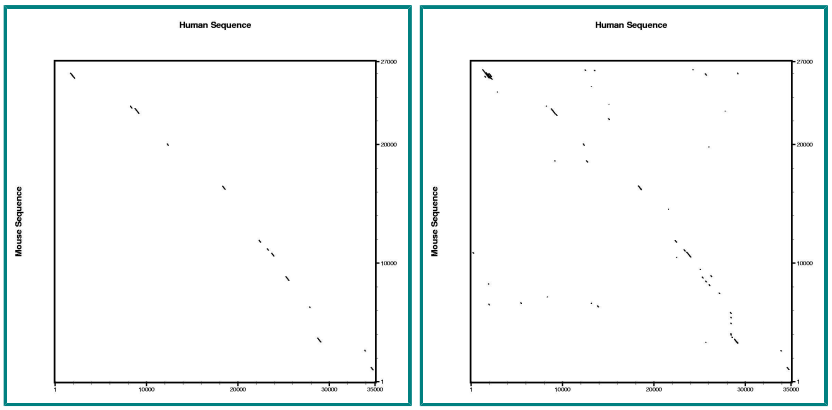
1. Connect to the `Blast2Sequences` server by following this link.
2. Select blastn in the program box.
3. Paste the Human sequence in the "Sequence 1".
4. Paste the Mouse sequence in the "Sequence 2".
5. Align.

In order to use `tblastx` follow these steps:

1. Connect to the `Blast2Sequences` server by following this link.
2. Select `tblastx` in the program box.
3. Paste the Human sequence in the "Sequence 1".
4. Paste the Mouse sequence in the "Sequence 2".
5. Align.

Are all the predicted exons supported by conserved regions ?

Here you can find a plot with the alignment results of the `blastn` and the `tblastx` alignments.



There are several programs to align and visualize pairs of large genomic sequences, for instance: `gff2aplot`, `Vista` and `Pipmaker`.

Using comparative gene finding tools

In this section we will use `sgp2` to make the predictions using the conservation pattern between human and mouse.

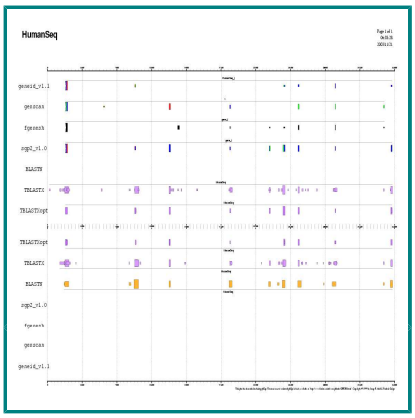
In order to use `SGP2` follow these steps:

1. Connect to the `sgp2` server by following this link.
2. Paste the Human sequence in the "Sequence 1".
3. Paste the Mouse sequence in the "Sequence 2".
4. Select Homo sapiens vs Mus musculus parameters.
5. Select Prediction in both sequences.
6. Select `geneid` output format.

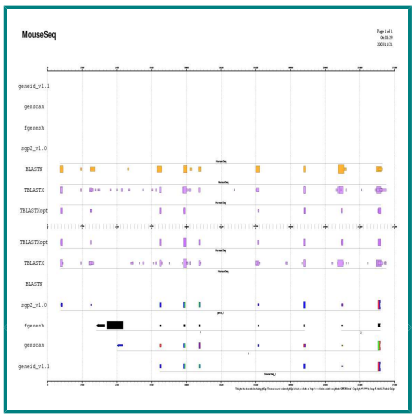
Some questions:

- Are human and mouse predictions similar using `sgp2` ?
- Does any of the ab initio predicted exons match `sgp2` predictions ?
- Does the similarity regions found by tblastx match `sgp2` predictions ?

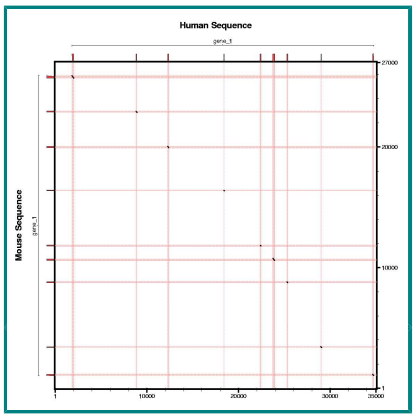
Here you can find the human predictions, the mouse predictions and the human and mouse predictions with the `tblastx` similarity regions.



Human (using mouse)



Mouse (using human)



Human and mouse (using `tblastx`)

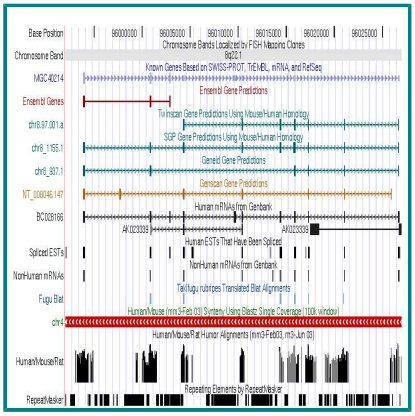
There are other program that uses genomic comparison to improve gene prediction: `twinscan` and `slam`.

Current annotations in the genomic DNA sequence

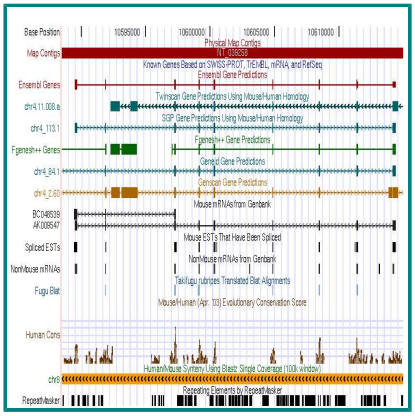
Go to the UCSC genome browser , and look for the annotation of this region in the human genome. Open another web browser window and look for the annotation of the mouse sequence in the mouse genome annotation.

The predictions we have obtained, are they consistent with the annotation of the UCSC genome browser ?

Here you can find a plot sumarazing annotations from human and mouse.



Human annotations

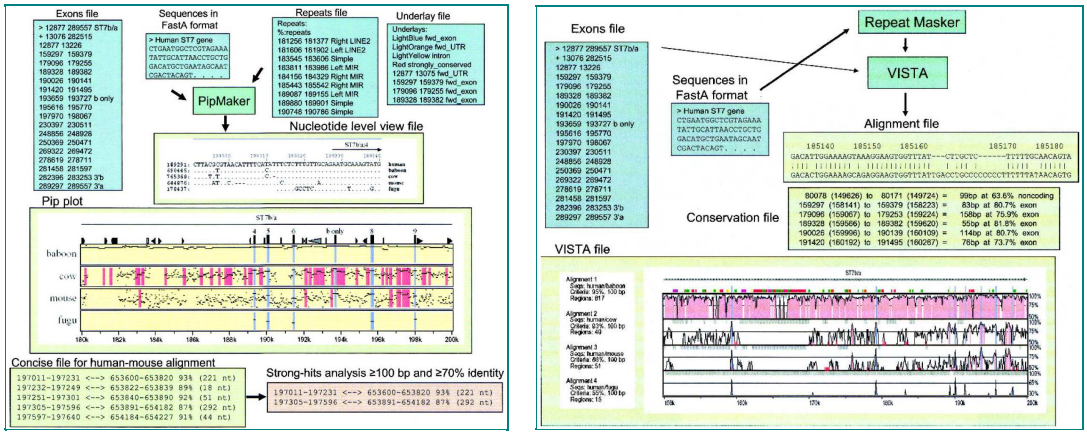


Mouse annotations

Sequence Alignment and Comparative Genomics

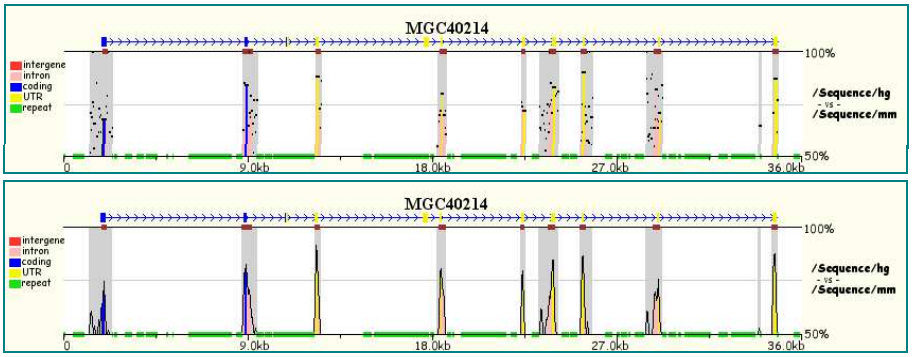
Comparative analysis of DNA sequences from multiple species at varying evolutionary distances is a powerful approach for identifying coding and functional non-coding sequences, as well as sequences that are unique for a given organism. Here we will survey few of such tools.

Two well established programs, `Vista` and `Pipmaker`, will provide the best results. They are not easy to use as they require several input files, each having different formats, to obtain highly customized outputs, as it is illustrated in the figure below (from Frazer et al, Genome Research, 13(1):1-12, 2002; by the way, a must read review).



Input and output files for Pipmaker (left) and Vista (right).

More intuitive and user-friendly tools with similar capabilities have appeared recently, the most remarkable ones being the ECR-Browser, zPicture and eShadow (all three from the Comparative Genomics Center at Lawrence Livermore National Laboratory). The following figure illustrates the differences between pip- and smooth-plots.



Output from zPicture: Human hipotetical protein MGC40214 (RefSeq Id NM152416) genomic region from human chromosome 8, compared against its mouse orthologous region. Upper panel shows a pip-plot (PipMaker like), while the bottom panel visualizes the same data with a smooth-plot (VISTA like).

Finally, we will see two web browsers that have been developed from the comparative genomics standpoint. You can follow the link to the K-Browser and the MultiContigView from Ensembl browser.

http://genome.imim.es/courses/laCaixa05/

Last modified: Thu, 23 Jun 2005 14:09:29 GMT