Structural annotations

Exercise I Predicting the gene structure of a genomic fragment

You are going to annotate a genome sequence of *Arabidopsis thaliana.*

*Connect to the NCBI website to retrieve the sequence in FASTA format whose accession number is the following:* AL356332 from base 1 to base 10000 .

1. Detection of splicing signals

First use the Netgene2 software at the following address <https://services.healthtech.dtu.dk/services/NetGene2-2.42/> to detect the position of introns and exons on the sequence to be annotated.

1. Using ab initio prediction software

Use the FGENESH software at the following address

[http://www.softberry.com/gfind](http://www.softberry.com/berry.phtml?topic=fgenesh&group=programs&subgroup=gfind) to predict the genes present.

1. How many genes have been predicted.
2. For each gene, give the strand, the number of exons and their size.
3. Save the sequence of the predicted mRNA as well as that of the corresponding protein.
4. Using protein sequences to predict gene structure
5. Perform a blast between your sequence to be annotated and the swissprot and nr-prot databases to locate the coding regions.
6. What is the benefit of using the swissprot database ?
7. Combine all the information obtained to predict the gene structure of your region.
8. How many genes do you predict?
9. Comparison of results with the official phytozome annotation.

Go to the Phytozome blast site <https://phytozome-next.jgi.doe.gov/>

From each mRNA sequence, search for chromosomal localization by blast against the Arabidopsis thaliana genome.

Compare the structure predicted by FGENESH and the official annotation. Are there any differences? Try to explain them.