Identification of selenoprotein-related genes in a vertebrate genome

- Bioinformatics 2022/23 -

Marco Mariotti, Diego Garrido, Roderic Guigó

Bioinformatics and genomics programme
Roderic Guigó's group
Centre for Genomic Regulation, Barcelona

Selenoproteins as test case

- Selenoproteins have the particular characteristic of including a **UGA codon**, recoded because of the presence of the **SECIS element**.

BIOINFORMATICS PROJECT

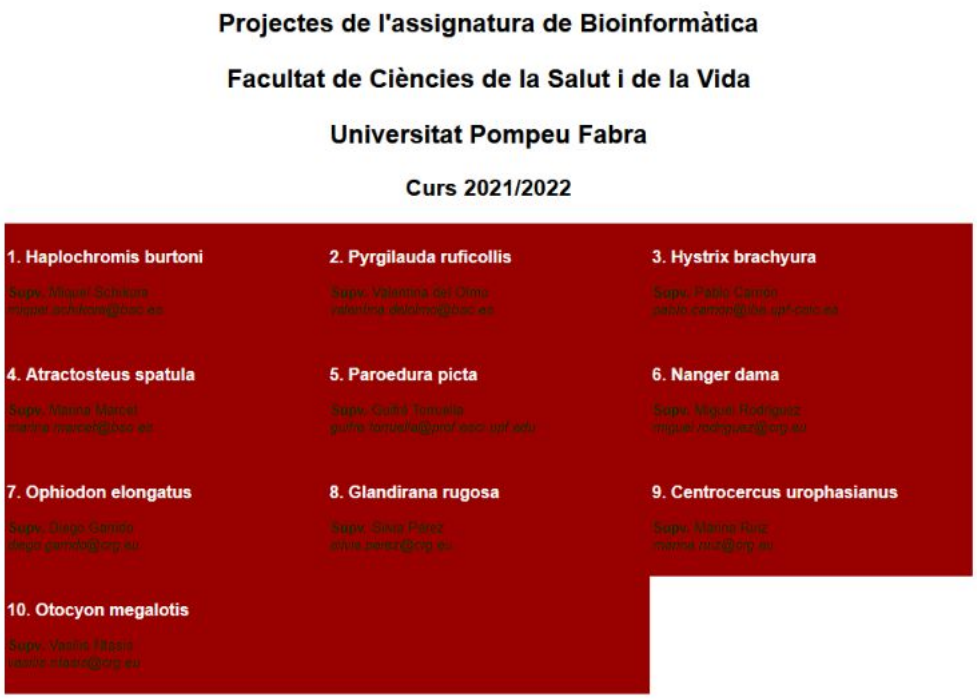
Find all **selenoprotein-related genes**

in a **vertebrate genome**

- If you learn how to predict selenoproteins, you will be able to do the same with any protein family.

UPF Human Biology. Bioinformatics Courses 2007-2022

2007/08 – 2008/09: find all selenoproteins in a given protist genome
2009/10 – 2011/12: find a given selenoprotein family in all protist genomes
2012/13 – **2022/23**: <u>find all selenoproteins in a given **vertebrate** genome</u>



http://bioinformaticaupf.crg.eu

Project 2022-2023 selenoproteins in vertebrates

- **Web page**: Structure of a scientific paper
- **Wikipedia**: Species description

ESTUDIS DE CIÈNCIES DE LA SALUT I DE LA VIDA

Selenoproteins in *Miichthys miiuy*

Selenoproteins are a group of proteins characterized by the presence of, at least, one Selenocysteine (Sec) residue in its chain. Since this residue is codified by UGA, which is normally considered as a stop codon, some of this proteins are dismissed in genome databases.

Moreover, the inclusion of Selenocysteine residue depends on the presence of an element called Selenocystein Insertion Sequence (SECIS), which is a secondary mRNA structure that allows the insertion of a selenocysteine instead of a stop codon.

The aim of our study is to predict the selenoproteins of *Miichthys miiuy*, a Japanese benthic fish, performing an homology-based in silico search. In order to assess the characteristics of the *Miichthys miiuy*'s selenoproteome, we have compared the genome of this species with *Danio rerio*'s and *Homo sapiens*'s selenoproteins annotations obtained from SelenoDB. For the prediction, different bioinformatic tools such as BLAST, Exonerate, Genewise, T_coffee, Seblastian and SECISearch3 were needed. Additionally, we have designed an automatic program to speed up the process.

Our results show a high conservation between Zebrafish' and *Miichthys miiuy*' selenoproteome. We have found 33 selenoproteins, 8 Cys-containing homologous proteins, 5 machinery proteins and 11 proteins related to selenium metabolism.

This study contributes with the identification of selenoproteins in new-sequenciated organisms.

Miichthys miiuy

Miichthys miiuy és una espècie de peix de la família dels escènids i de l'ordre dels perciformes.

Contingut [amaga]

1 Morfologia
2 Hàbitat
3 Distribució geogràfica
4 Ús comercial
5 Observacions
6 Referències
7 Bibliografia
8 Enllaços externs

Morfologia [modifica | modifica el codi]

Els mascles poden assolir 70 cm de longitud total.[5][6] Com la resta de peixos de la família Sciaenidae, *M. miiuy* és conegut per tenir uns otòlits excepcionalment grans que els doten d'un sistema auditiu molt desenvolupat.[7] Aquests peixos s'anomenen sovint peixos tambors o corballs a causa dels sons que produeixen amb les seves bufetes natatòries.

Hàbitat [modifica | modifica el codi]

És un peix de clima temperat i demersal que viu entre 15-100 m de fondària.[5][8] Eviten les aigües clares, prefereixen viure en estuaris, badies i riberes de rius fangosos. Són organismes carnívors bentònics.[7]

Miichthys miiuy	
Taxonomia	
Super-regne	Eukaryota
Regne	Animalia
Fílum	Chordata
Classe	Actinopterygii
Ordre	Perciformes
Família	Sciaenidae
Gènere	*Miichthys*
Espècie	***Miichthys miiuy*** (Basilewsky, 1855)[1][2][3]
Nomenclatura	
Sinònims	• *Argyrosomus miiuy* (Basilewsky, 1855) • *Miichthys imbricatus* (Matsubara, 1937) • *Nibea imbricata* (Matsubara, 1937) • *Otolithus fauvelii* (Peters, 1881) • *Sciaena miiuy* (Basilewsky, 1855)[4]

Bioinformatics methods for selenoprotein prediction

- ***De novo***: Selenogeneid (Castellano et al. 2001)

- **Homology-based approaches:**
  - UGA/Sec or UGA/Cys alignments (e.g. Kryukov et al. 2003)
  - Selenoprofiles (Mariotti and Guigó 2010)
  - Seblastian (Mariotti et al. 2013)

- **SECIS prediction:**
  - SECISearch (Kryukov et al. 2003)
  - SECISearch3 (Mariotti et al. 2013)

- **tRNA-Sec prediction:**
  - Secmarker

Protocol steps

- **SelenoDB 2.0**

  http://selenodb.crg.es

- **BLAST** (tblastn)
- **Exonerate** - protein2genome mode
- **Genewise**
- **T-coffee**

http://bioinformaticaupf.crg.eu

- **SECISearch3** and **Seblastian**

  http://seblastian.crg.es

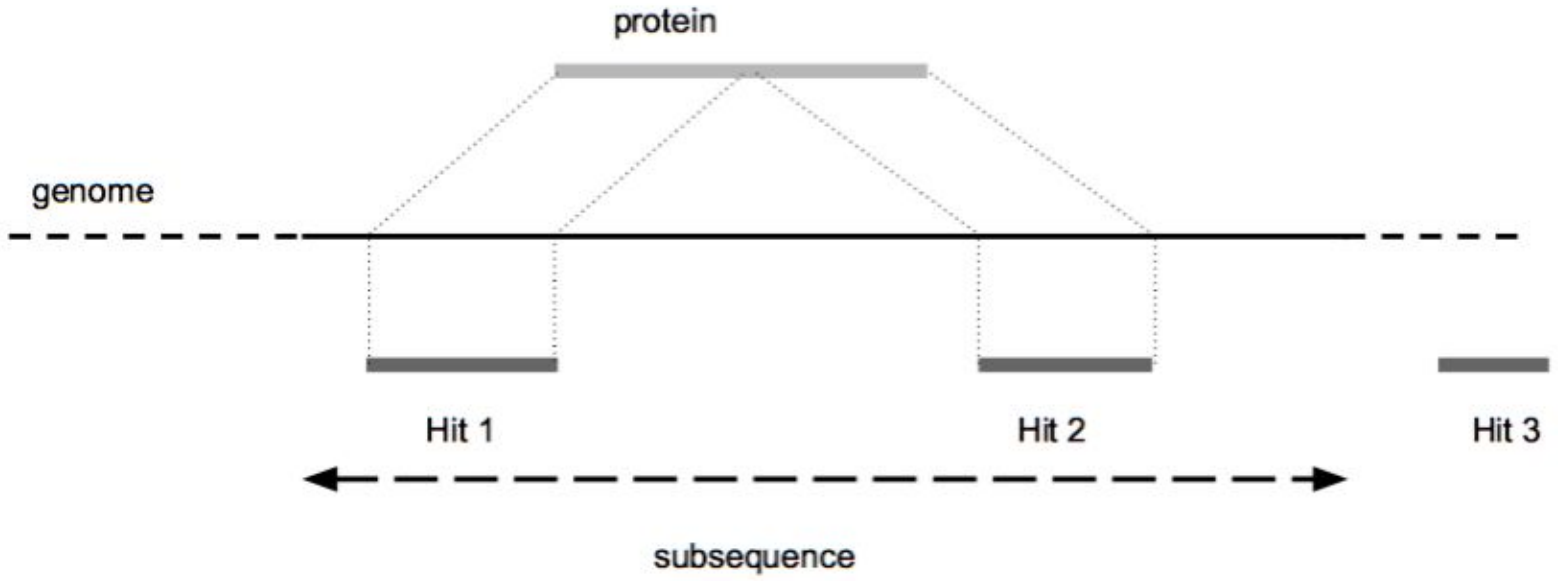
Protocol steps

Get selenoprotein sequences from a related organism

- **SelenoDB 2.0** SelenoDB
  http://selenodb.crg.es (automatic annotation)

Protocol steps

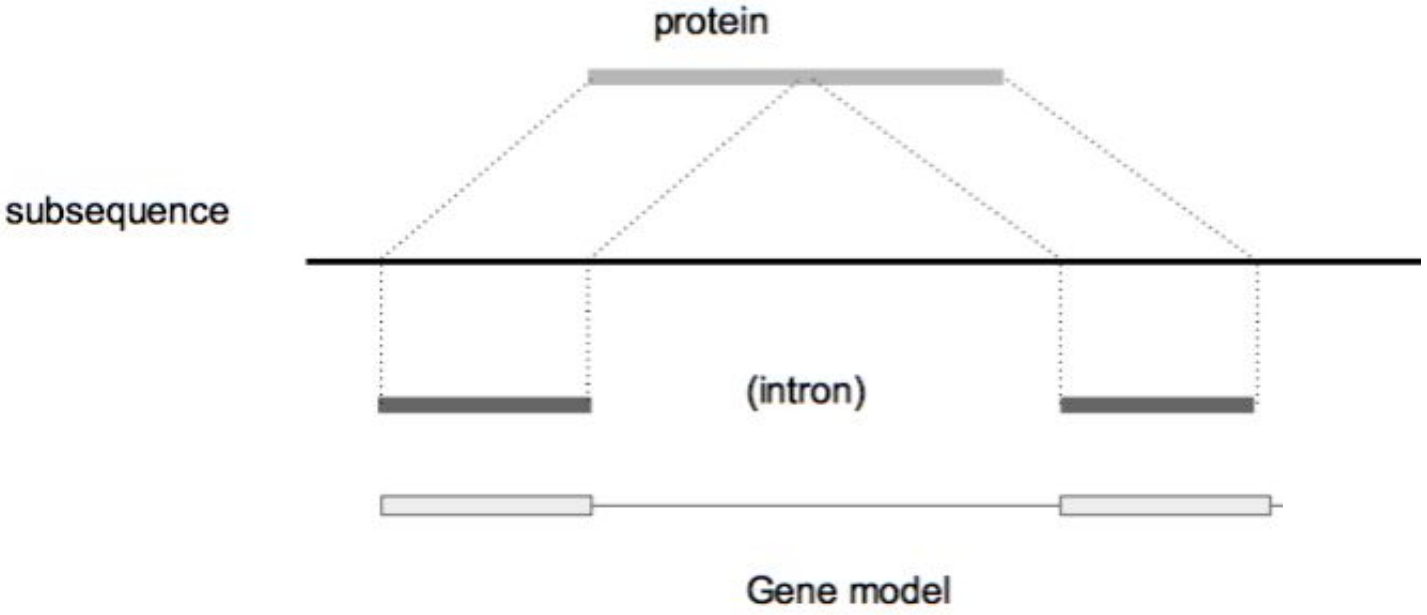
Locate gene exons (independent **tblastn** hits) in the genome of your organism



Search Name	Query Type	Database Type	Translation
tblastn	Peptide	Nucleotide	Database

Protocol steps

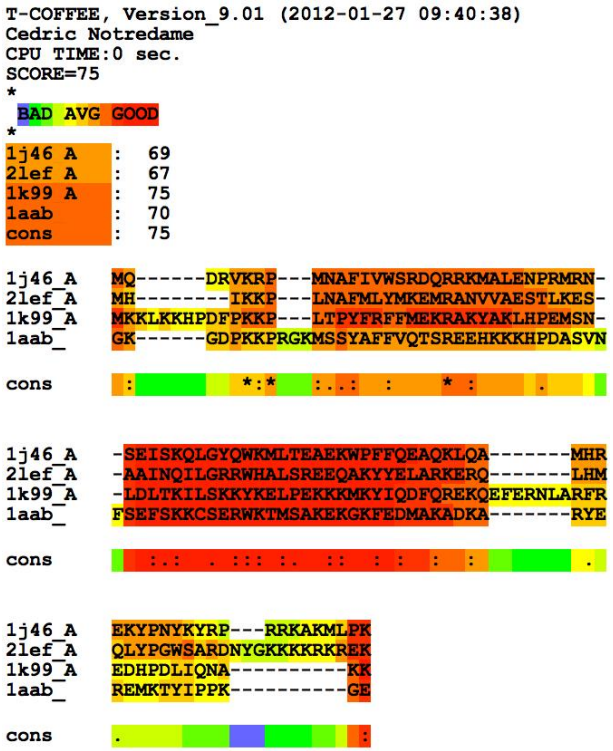
Build a multi-exonic gene model (**exonerate** or **genewise**)



and translate it into a protein

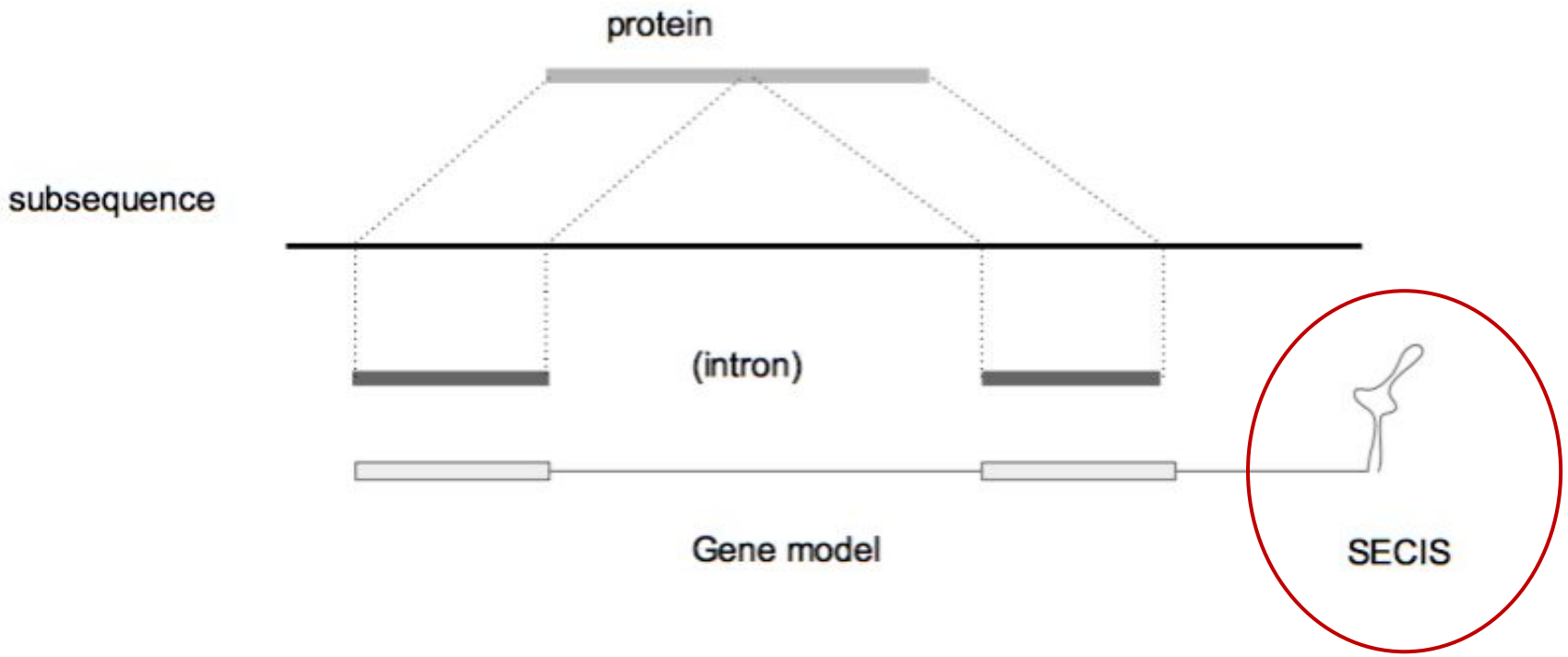
Protocol steps

Compare the known sequence that you obtained from the database (query protein) with the homologous sequence of the genome of your organism (predicted protein) (**t-coffee**)



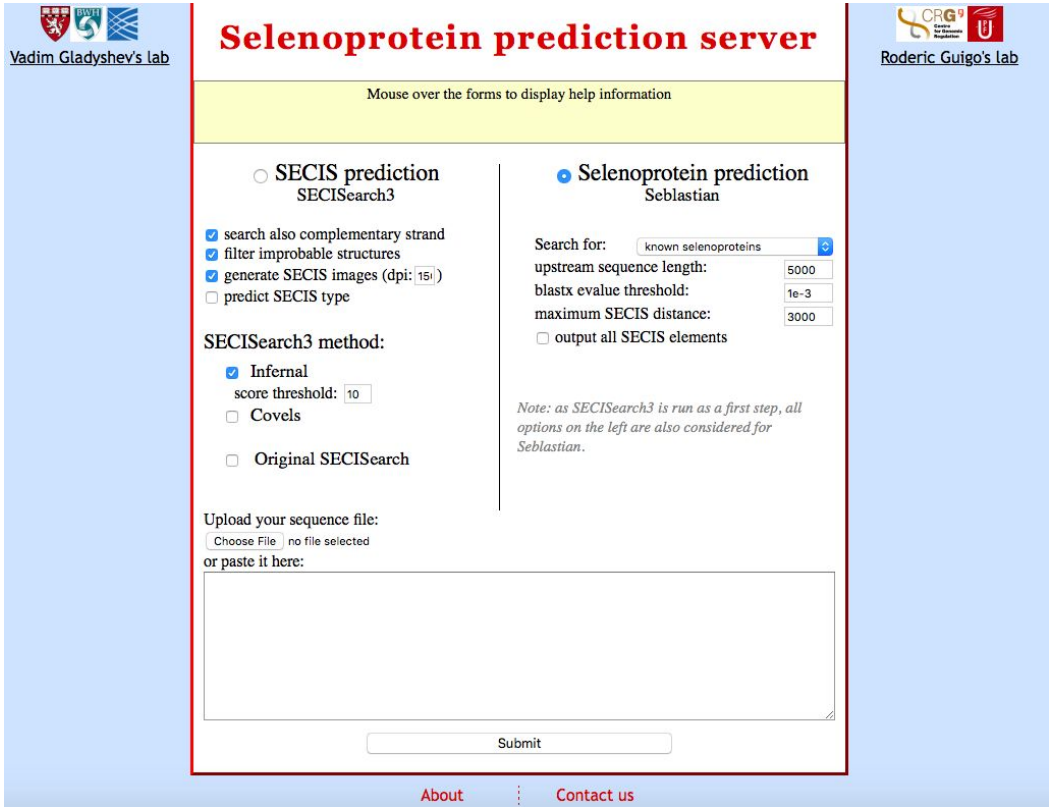
Protocol steps

SECIS and selenoprotein prediction (**SECISearch3** and **Seblastian**)

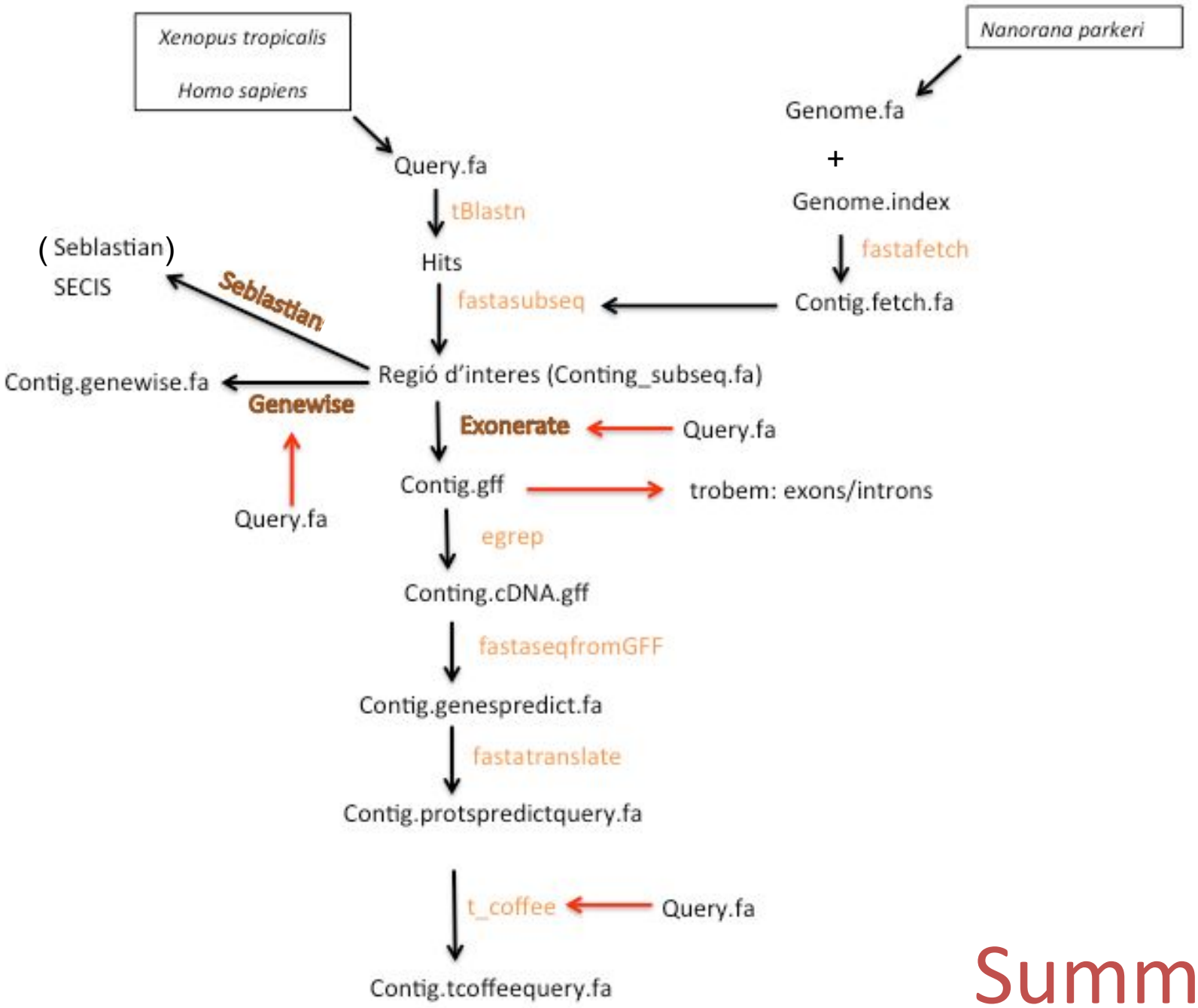


Protocol steps

Seblastian: Predict SECIS in the 3'UTR (using SECISearch3), and then searches upstream for selenoprotein coding sequences.



http://seblastian.crg.es/


Xenopus tropicalis
Homo sapiens
Nanorana parkeri
Genome.fa
+
Genome.index
Query.fa
tBlastn
Hits
fastafetch
Contig.fetch.fa
fastasubseq
(Seblastian)
SECIS
Seblastian
Regió d'interes (Conting_subseq.fa)
Contig.genewise.fa
Genewise
Query.fa
Exonerate
Query.fa
Contig.gff
trobem: exons/introns
egrep
Conting.cDNA.gff
fastaseqfromGFF
Contig.genespredict.fa
fastatranslate
Contig.protspredictquery.fa
t_coffee
Query.fa
Contig.tcoffeequery.fa

Summary

Protocol steps

Gene finding tools:

- **fastafetch:** extracting a single sequence from a multifasta (requires previous run of fastaindex)
- **fastasubseq:** getting a subsequence of a single sequence, careful with indexes, 0-based! Transform gene positions to absolute coordinates.
- **exonerate/genewise:** predict the gene and align it with the sequence of the selenoprotein that encodes, and also recognizes the exons.
- **fastaseqFromGFF.pl:** obtain the cDNA sequence that encodes the final protein. We get it from the subsequence and the file that contains the exons.
- **fastatranslate:** translate coding sequences careful with the selenocysteine codon character! It is a good idea to substitute the "*" with "X" or "U" as multiple sequence alignment programs just ignore "*"

Notes for the project

- Results must be presented in a **web page** with the **structure of a scientific paper**
  - ✔ Aminoacid sequences + SECIS sequences
  - ✔ Genes in GFF format (absolute coordinates)

- All **genes** should be **as complete as possible**: starting with a AUG, ending with a STOP codon, and with an identified SECIS element downstream.

- **Ignore alternative isoforms** (if any), just choose one

- Report also the **genes** of:
  - ✔ **selenoprotein machinery**: SEPSECS, EEFSEC, PSTK, SBP2, SECP43, SEPHS1, SEPHS2, etc.
  - ✔ **Cys-containing homologs**

- **Other helpful resources** to biologically interpret and visualize the results (**phylogenetic trees**):
  - ✔ Phylogeny.fr: http://www.phylogeny.fr/simple_phylogeny.cgi (.mfa)
  - ✔ phyloT: https://phylot.biobyte.de/ (from NCBI taxonomy .nw)
  - ✔ iTOL: https://itol.embl.de/ (.nw)
  - ✔ Etetoolkit: http://etetoolkit.org/treeview/ (.nw or .msa)

- In some cases, the predicted protein can be **located** in **more than one contigs/scaffolds**. You will notice this if you try to predict the protein in both of them, and you pay attention at both sequence alignments performed by T-coffee.

- **HTML language**
  - ✔ https://www.w3schools.com/html/default.asp
  - ✔ https://getbootstrap.com/

Notes for the project

- We **already provide** you, together with the genome:

  /mnt/NFS_UPF/soft/genomes/2022/**Genus_specie**

  ✔ BLAST database for the genome

  ✔ Indexed genome

- **Scaffolds/Contigs lengths** can be found in the *genomes.lengths* file

- **fastatranslate** (option -F 1) to consider only the 1st ORF.

- Before performing the **sequence alignment with T-Coffee**, substitute the "*" with "X" or "U" as multiple sequence alignment programs just ignore "*"

- **Seblastian** and **SECISearch3 web servers:**

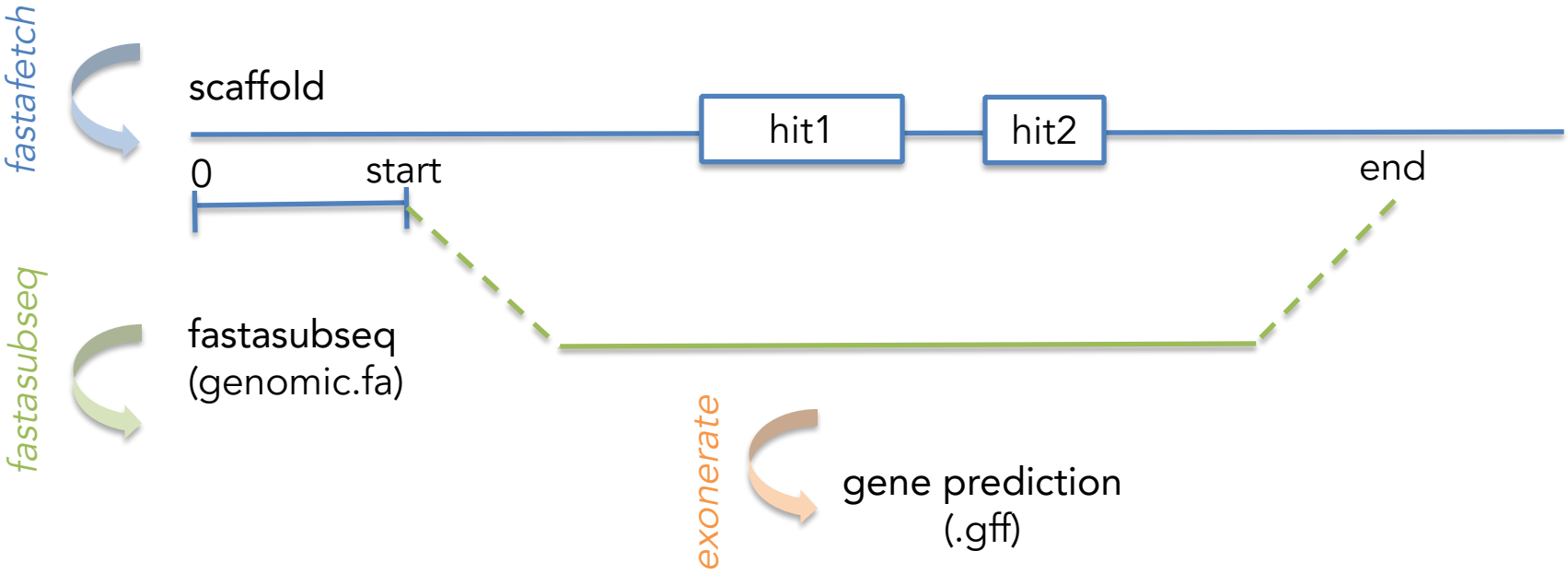
  Input: Nucleotide sequence (*fastasubseq* file)

  * DO NOT take into account other nucleotide bases different than A, C, G, T, a, c, g, t, or N. Then, in case you have one of the other symbols from ambiguity code, one solution could be substituting them by an N.

Johnson A.D. An extended IUPAC nomenclature code for polymorphic nucleic acids. *Bioinformatics.* 2010; 26(10): 1386-1389.

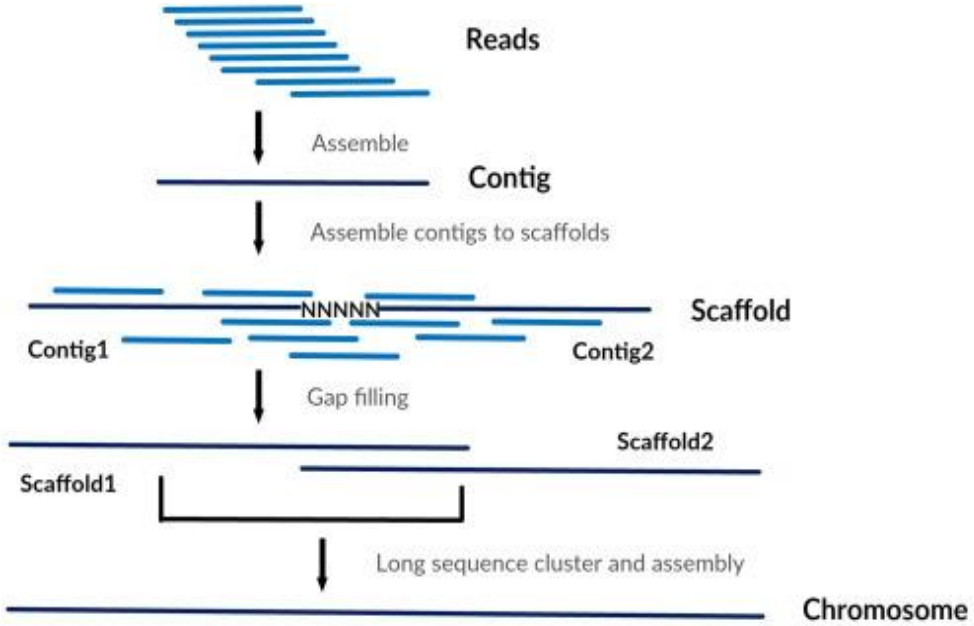
Notes for the project

- **Genes prediction (GFF format):** Conversion of **relative** to **absolute coordinates**
  - Apart from obtaining the protein sequence predictions, you should obtain the gene predictions in .gff format considering the absolute coordinates.
  - Remember that, as you made your prediction using the *fastasubseq* file, you will be predicting the genes (.gff file from exonerate) with the relative coordinates instead of the absolute coordinates. Then, to generate the .gff files with absolute coordinates, you will have to convert the your .gff files with relatives coordinates (.gff file from exonerate) considering the **start** you decided to give to the **fastasubseq program.** [In this case, **start**: start_hit1 nt - 50.000 nt]



Notes for the project

- **Contigs** and **Scaffolds**
  - ✔ **Contig**: a contiguous stretch of nucleotides resulting from the assembly of several reads
  - ✔ **Scaffold**: several contigs stitched together wit NNNs in between



Notes for the project
Automation: BASH scripting

- Basics of Bash scripting
  - ✔ slides

- Bash documentation
  - ✔ https://www.gnu.org/software/bash/manual/bash.pdf

- Bash cheatsheets
  - ✔ https://devhints.io/bash
  - ✔ https://github.com/LeCoupa/awesome-cheatsheets/blob/master/languages/bash.sh

Technical info

- Access to the shared folders in *fs-aules* and VPN usage -

https://bioinformaticaupf.crg.eu/accedir_carpeta_compartida_fs-aules.pdf

Evaluation

The projects will be **evaluated** based on:

- ✔ **Methods:** scripting is encouraged (different levels of automation)
- ✔ **Results:** you are expected to find all selenoprotein-related genes in your assembly
- ✔ **Discussion:** interpret your results logically
- ✔ **Presentation:** the web page should present the work as clearly as possible (including the Wikipedia entry)

Evaluation

Abstract		**0.25**
Introduction	Selenium, selenoprotein biosynthesis, evolution/phylogeny, families (including selenoproteins, cys-homologues and machinery), links to wikipedia entry, schemes and diagrams	**0.50**
Materials and methods	Queries selection, description of each step of the annotation pipeline, automatization, SECISearch3/Seblastian server	**3.00**
Results	Summary table/plot with protein query (+specie/species), Sec/Cys-homologue, tblastn, exonerate/genewise, scaffold, gene prediction, protein prediction, sequence alignment, SECIS and Seblastian predictions. Description of each prediction: protein location (scaffold and coordinates + strand), exons, SECIS elements and Seblastian predictions	**2.00**
Discussion	Discuss each family, highlight interesting cases (duplications, losses and conversions from Sec-to-Cys events), contrast hypotheses with the literature, use specific phyloge netic terms	**2.50**
Conclusions	Wrap up the results, further directions, pros and cons of the project, limitations	**0.50**
References	Citations along the text, proper numeration	**0.25**
Web page format	Structure of scientific paper, creativity, links working properly	**0.50**
Wikipedia entry	Extend the content in different languages	**0.50**

Groups, supervisors and species

Group	Subgroup	Supervisor	E-mail	Species
Grup 101	1	Giovanni Asole	giovanni.asole@crg.eu	*Bettongia penicillata*
	2	Sadoia Manzano	manzano.saioa@gmail.com	*Sceloporus occidentalis*
	3	Guifré Torruella	guiffttc@gmail.com	*Clinocottus analis*
Grup 102	4	Hannah Benisty	hannah.benisty@crg.eu	*Hipposideros pendleburyi*
	5	Sergio Sánchez	sergio.sanchez.moragues@gmail.com	*Artemisiospiza belli*
	6	Miquel Schikora	miquelangel.schikora@irbbarcelona.org	*Bovichtus variegatus*
	7	Tamara Perteghella	tamara.perteghella@crg.eu	*Symphalangus syndactylus*
Group 103	8	Nadia Makarova	na96mk@gmail.com	*Hyperoodon ampullatus*
	9	Xavier Hernández	xavier.hernandez@crg.eu	*Actenoides hombroni*
	10	Martina Cardinale	martina.cardinali@irbbarcelona.org	*Darevskia valentini*
	11	Jaume Reig	jaume.reig@crg.eu	*Tragulus kanchil*