UPF Bioinformatics course projects

Students guide 2017

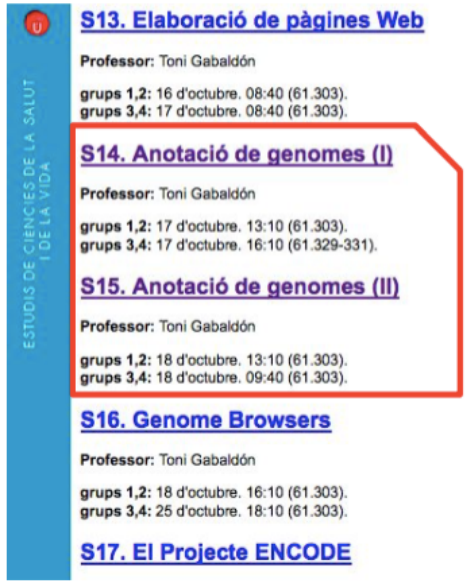
Didac Santesmasses (PhD) didac.santesmasses@crg.eu
Aida Ripoll (Master student) aida.ripoll@crg.eu

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

Protocol overview

Tools:

- **BLAST** - typically **tblastn**
- **Exonerate** - protein2genome mode
- **Genewise**
- **T-coffee**



http://bioinformatica.upf.edu/

- Webserver with **SECISearch3** and **Seblastian**:

  http://seblastian.crg.es/

Protocol overview

1st step: Get selenoprotein sequences

- **SelenoDB 2.0 (and 1.0)** SelenoDB

  http://www.selenodb.org (2.0; automatic annotation)

  http://www1.selenodb.org (1.0; manually curated, less species)

- **Protein databases**

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

  http://www.uniprot.org

- **Past year projects:**

  http://bioinformatica.upf.edu/

Genomes

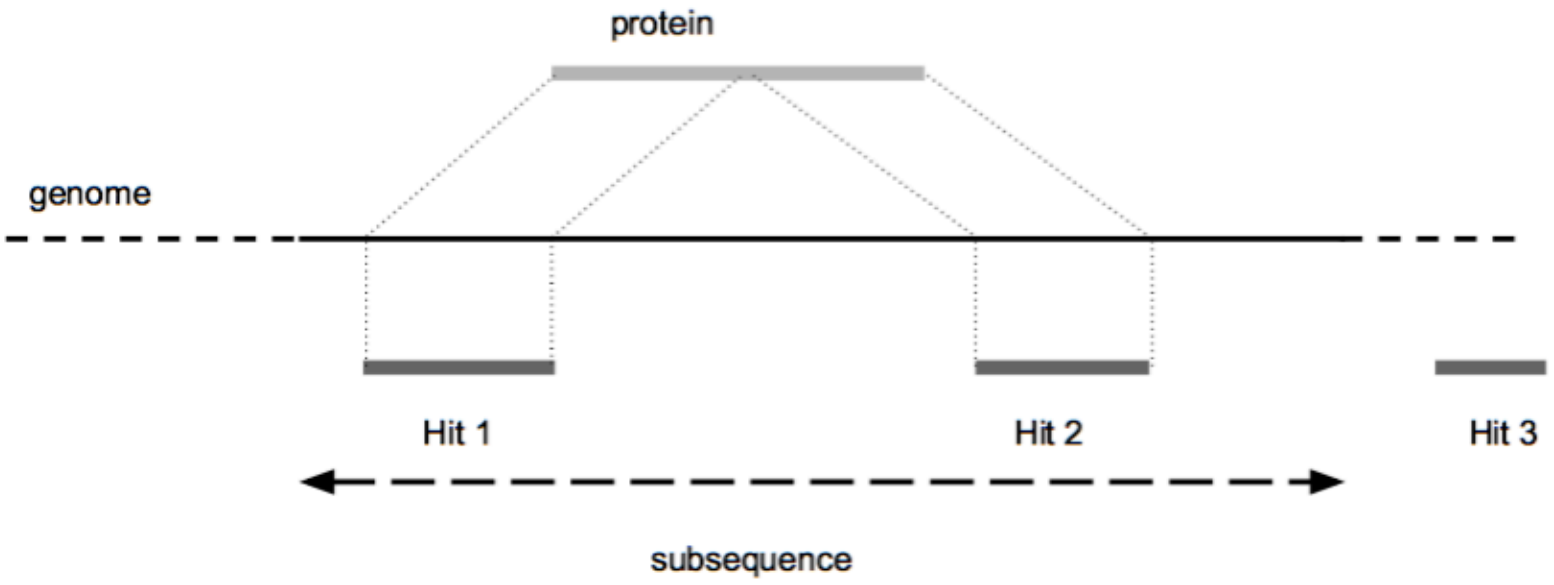
Genomes (blast formatted and indexed)

/cursos/20428/BI/genomes/2017/Genus_species/genome.fa

/cursos/20428/BI/genomes/2017/Genus_species/genome.index

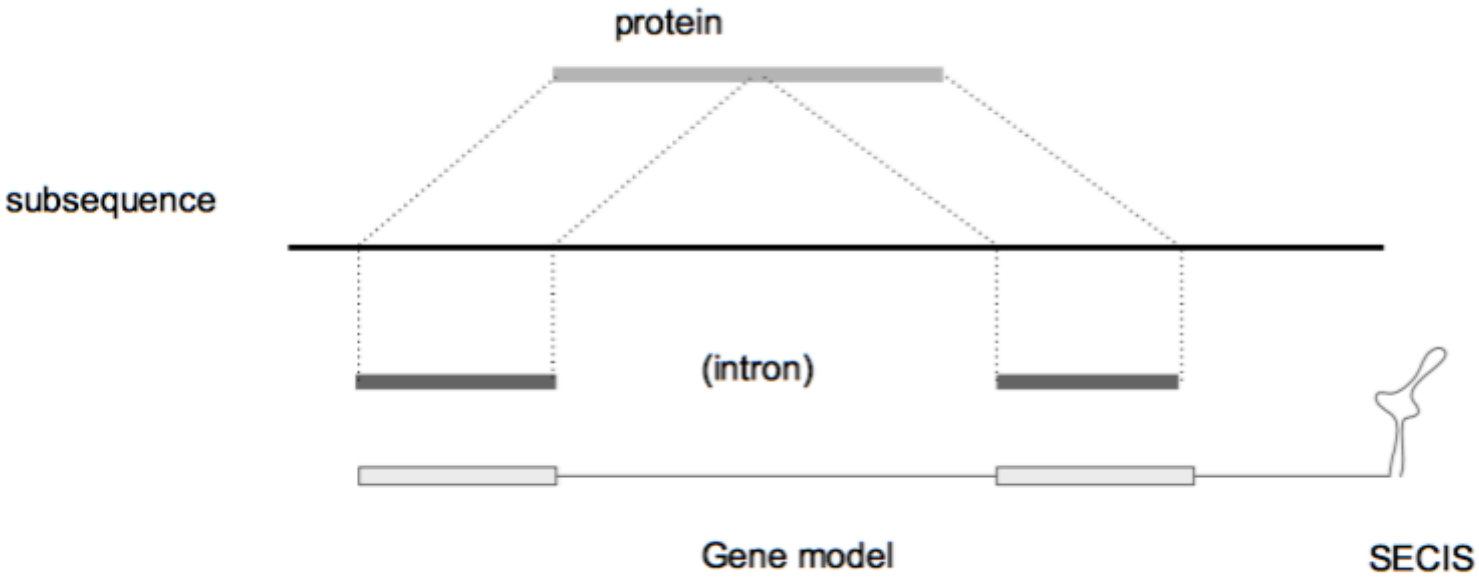
Protocol overview

- **Tblastn:** locate gene exons (independent blast hits)



Protocol overview

- **Exonerate** or **genewise:** multi-exonic gene model
- **Seblastian:** SECIS prediction



Protocol overview

Gene finding tools: fastasuite **(exonerate)**

- **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:** obtain the cDNA sequence that encodes the final protein. We get it from the subsequence and the file that contains the exons.

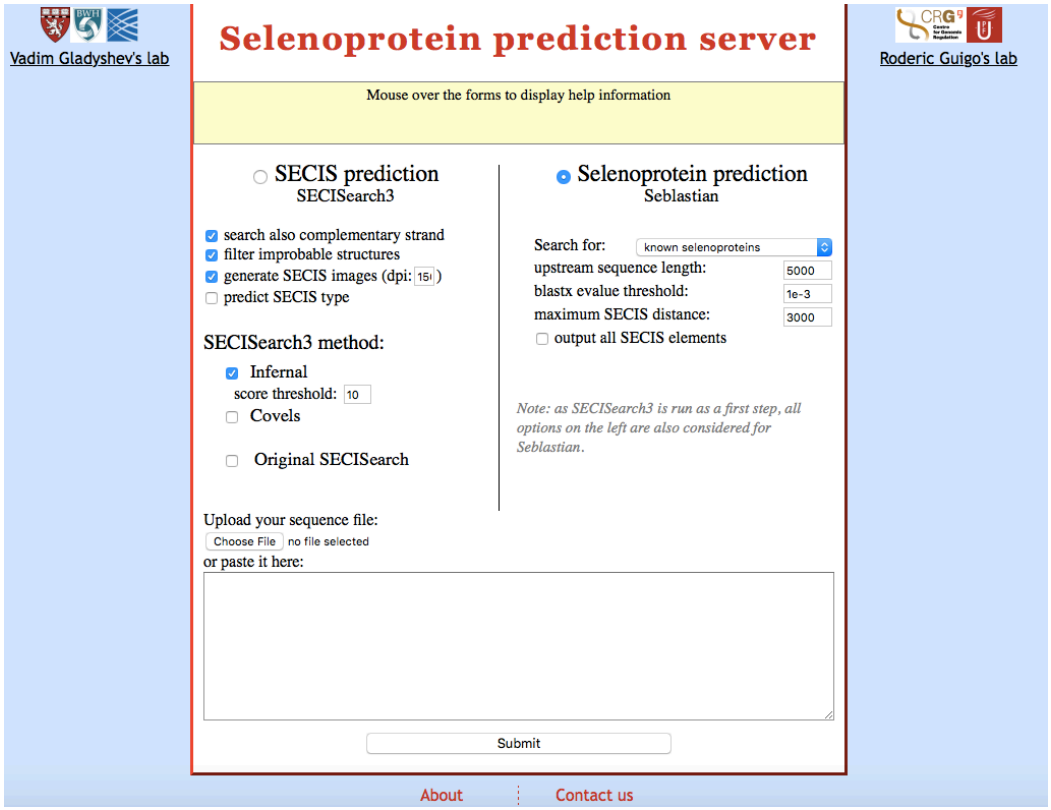
- **Fastatranslate:** translate coding sequences careful with the selenocysteine codon character! It is a good idea to substitute the "*" with "X" or "U" multiple sequence alignment programs just ignore "*"

Protocol overview

- **Tcoffe:** compare two sequences, in this case we compare the known sequence (query protein) with the homologous sequence of the the genome (predicted protein).

Protocol overview

Seblastian: Predict SECIS in the 3'UTR, then upstream for selenoprotein coding sequences



UPF Human Biology.
Bioinformatics Courses 2007-18

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 – **2017/18**: find all selenoproteins in a given **vertebrate** genome

http://bioinformatica.upf.edu/

Projectes de l'assignatura de Bioinformàtica

Facultat de Ciències de la Salut i de la Vida

Universitat Pompeu Fabra

Curs 2012/2013

1A: Ailuropoda melanoleuca *AM. Barrios, A. Bellot,* *S. Castany, M. De Manuel*	**1B: Cricetulus griseus** *J. Fernandez, J. Gomez,* *FD. Jurquiza, A. Lopez*	**1C: Mustela putorius furo** *M. Perez, L. Taberner,* *G. Vilajosana, I. Villate*
2A: Nomascus leucogenys *M. Alemany, H. Costa,* *A. Escrig, I. Gafarot*	**2B: Saimiri boliviensis** *P. Garcia, J. Latorre,* *R. Martinez, H. Palma*	**2C: Sarcophilus harrisii** *G. Rodriguez, E. Ros,* *AM. Saludes, H. Xicoy*
3A: Chrysemys picta bellii *C. Bitlloch, G. Clua,* *J. Domingo, P. Gelabert*	**3B: Meleagris gallopavo** *J. Jancyte, L. Mateo,* *A. Olle, M. Perera, C. Perez*	
4A: Pelodiscus sinensis *SU. Abad, A. Almeyda,* *A. Azagra, R. Bartomeus*	**4B: Gadus morhua** *O. Bover, N. Cortell,* *B. Grau, E. March*	**4C: Latimeria chalumnae** *A. Martinez, A. Perlas,* *T. Robert, S. Walsh*

Groups and species

	Treball	Genoma	Contacte	Supervisor
Grup 101	1	*Sporophila hypoxantha*	laia.carrete@crg.eu	Laia Carrate
	2	*Ammotragus lervia*	fernando.cid@crg.eu	Fernando Cid
	3	*Pogona vitticeps*	miki.s.t@hotmail.com / marina.lleal01@estudiant.upf.edu	Miquel Angel Schikora / Marina Lleal
Grup 102	4	*Oncorhynchus mykiss*	Marina.Marcet-Houben@crg.eu	Marina Marcet-Houben
	5	*Monopterus albus*	diego.garrido@crg.eu	Diego Garrido
	6	*Astyanax mexicanus*	aida.ripoll@crg.eu / marta.badia.mb@gmail.com	Aida Ripoll / Marta Badia
Grup 103	7	*Meriones unguiculatus*	miguelangel.naranjo@crg.eu	Miguel A. Naranjo
	8	*Nannopterum harrisi*	irene.julca@crg.eu	Irene Julca
	9	*Taeniopygia guttata*	tonidedios94@gmail.com / ramon.massoni01@estudiant.upf.edu	Toni de Dios / Ramon Massoni
Grup 104	10	*Spermophilus dauricus*	andres.lanzos@crg.eu	Andres Lanzos
	11	*Odocoileus virginianus texanus*	cecilia.klein@crg.eu	Cecilia Klein
	12	*Anas zonorhyncha*	rosamaria.fernandez@crg.eu	Rosa M. Fernandez

Supervisors Sessions

https://docs.google.com/spreadsheets/d/1D1jkz1pQD8NyF_OPpUesdEEw1KA27qQz9iaetCBgmJE/edit#gid=0

Project 2017-18
selenoproteins in vertebrates

http://bioinformatica.upf.edu/

- **Web page**: Structure of a scientific paper

- **Wikipedia**: Species description (not english entry)
- **SelenoDB**: Insert your selenoprotein genes predictions into a real world database. Available to the scientific community.

Technical issues

Genomes (blast formatted and indexed)
/cursos/20428/BI/genomes/2017/Genus_species/genome.fa

Access to the cluster (connect through ssh)

- From informatic rooms (ip):
ssh UXXXXXX@10.200.127.240 -X
(alumnes) password: dd/mm/yyyy (birth date)

- Outside UPF (**VPN**):
ssh UXXXXXX@sitdoc.s.upf.edu -X
(alumnes) password: dd/mm/yyyy (birth date)

Modules

```
module load modulepath/goolf-1.7.20
module load BLAST+/2.2.30-goolf-1.7.20
module load Exonerate/2.2.0-goolf-1.7.20
module load T-Coffee/11.00.8cbe486-goolf-1.7.20
export PATH=/cursos/20428/BI/bin:$PATH
export PATH=/cursos/20428/BI/soft/genewise/x86_64/bin:$PATH
export WISECONFIGDIR=/cursos/20428/BI/soft/genewise/x86_64/wise2.2.0/wisecfg/
```

Cluster

- **VPN connection** https://www.upf.edu/bibtic/en/guiesiajudes/recinfo/vpn/
  ssh UXXXXX@sitdoc.s.upf.edu (campus global password)

- File **blastJob.sh**

```
#!/bin/bash
#$ -o tblastn.stdout
#$ -e tblastn.stderr
#$ -q all.q
#$ -N blastJob
#$ -cwd
module load modulepath/goolf-1.7.20
module load BLAST+/2.2.30-goolf-1.7.20
tblastn -query fitxerquery.fa -db nombbddBLAST -out fitxerdesortida
```

Cluster

- **Queue system**

```
$ qsub blastJob.sh
$ qstat
job-ID  prior   name       user         state submit/start at     queue
    slots ja-task-ID
    -----------------------------------------------------------------------------------------
    -----------
       231 0.55500 blastJob   UXXXXX       r     02/10/2010 11:42:09 llicen.q@luke
    1
```

- **To kill a running job**

```
$ qdel 231
```

Adding genes to SelenoDB

Before adding genes to selenodb, you need to **fill the information:**
http://selenodb.crg.cat/selenodb_barcelona/add_author.html

Then you can use your **email as author** in:
http://selenodb.crg.cat/selenodb_barcelona/add_gene.html.mako