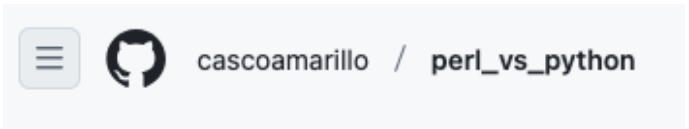
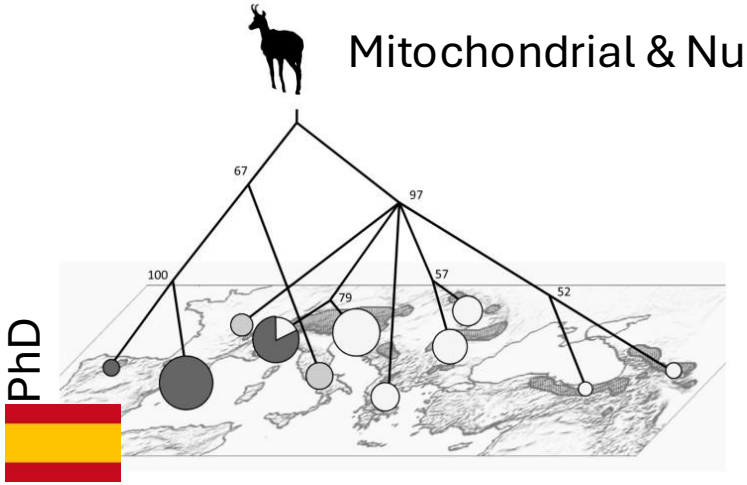
Fernando Rodriguez

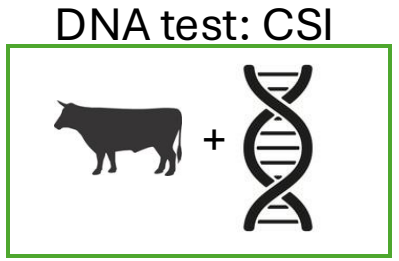
- I am a molecular biologist interested in genomics in eukaryotes

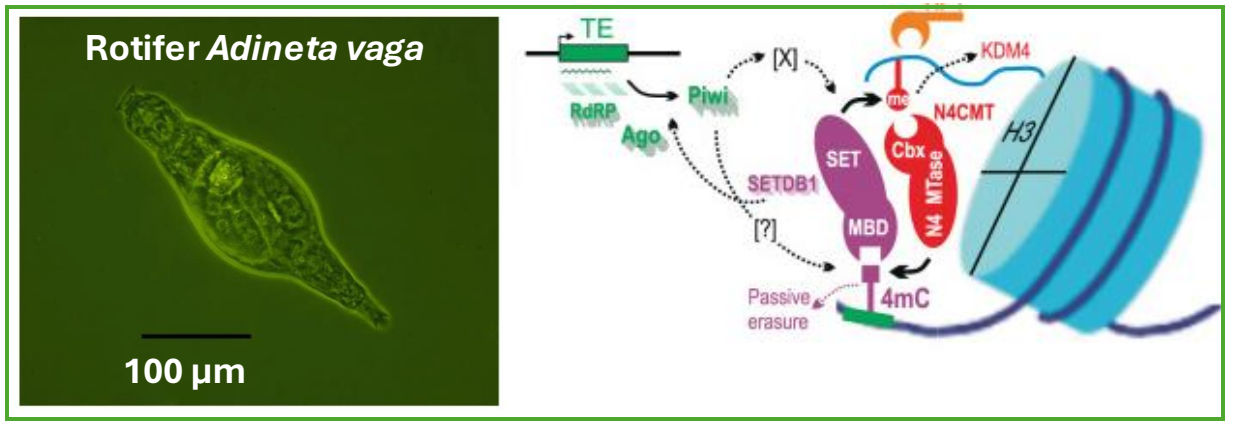
- I have studied genomic structural organization/regulation in different model organisms:
  - Cattle, chamois (mammals)
  - Fungi: Neurospora
  - Rotifers
  - Desert ants
  - Ostracods (crustaceans)



PhD







Epigenetic mechanism to repress transposons (TEs)

Programmatic Access To Biological Databases (Perl)

Date: Monday 1 October 2012

Application opens: Friday 01 June 2012

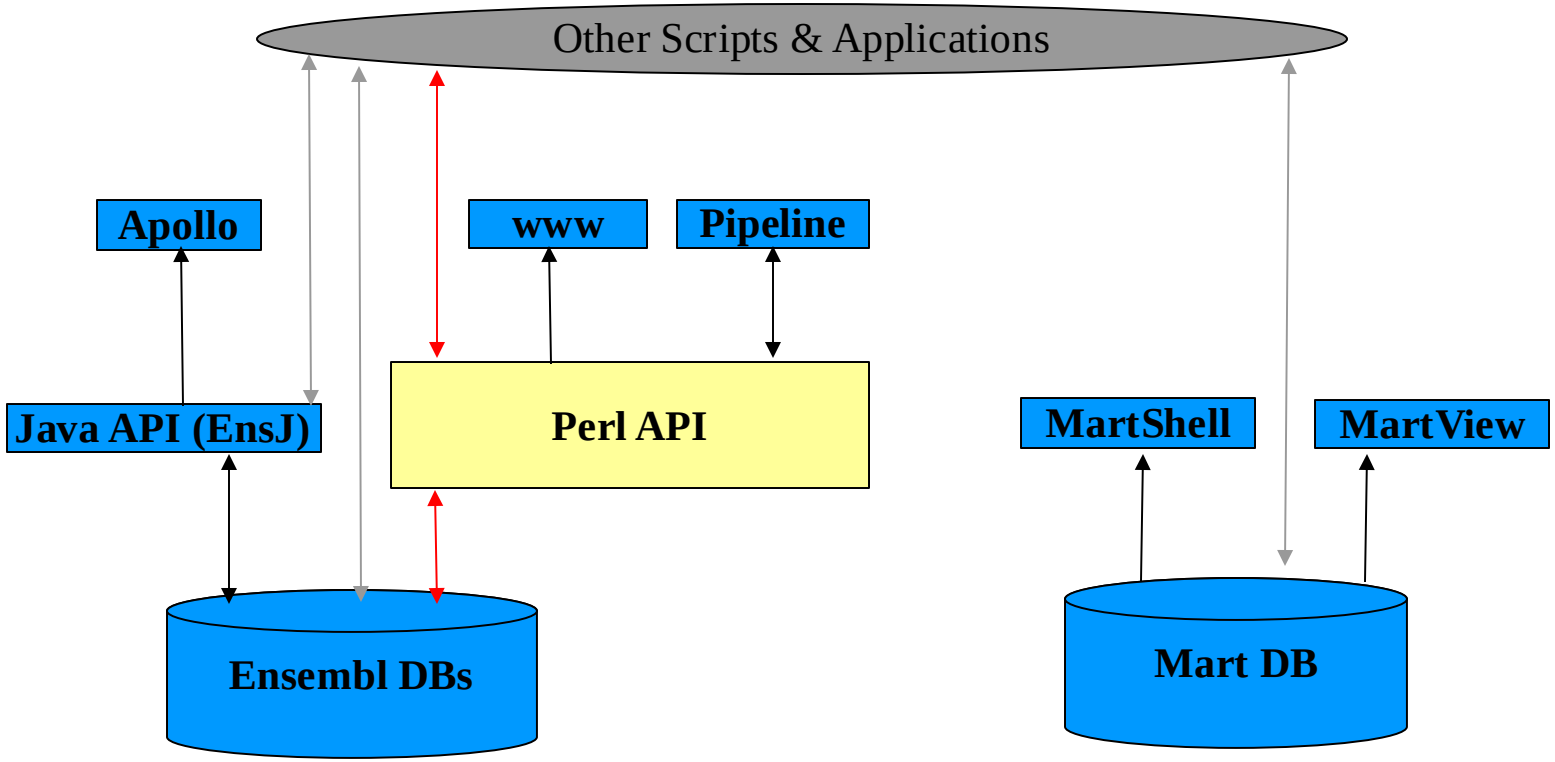
Application deadline: Monday 06 August 2012

Contact: Frank O'Donnell

Registration closed

European Bioinformatics Institute (EBI)
Hinxton, Cambridge, UK

System Context



What is the API?

- The Ensembl API (application programming interface) is a framework for applications that need to access or store data in Ensembl's databases.

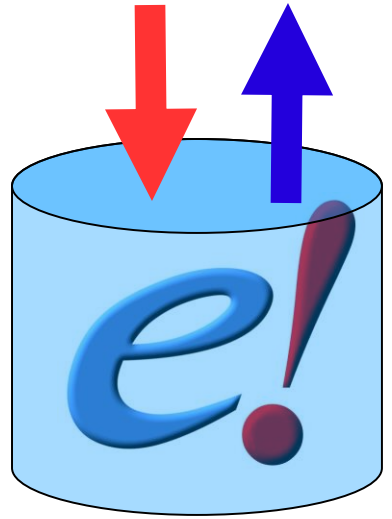
The Ensembl API

The Perl API

Written in Object-Oriented Perl.

Used to retrieve data from and to store data in Ensembl databases.

Foundation for the Ensembl Pipeline and Ensembl Web interface.



European Bioinformatics Institute (EBI)
Hinxton, Cambridge, UK

RECORDED WEBINAR

Programmatic access to UniProt using Python

https://www.ebi.ac.uk/training/events/programmatic-access-uniprot-using-python/

LIVE DEMO
Using collab Notebook

https://bit.ly/up-colab-2022

Ways to get UniProt data

FTP

Big one-off download, post-processing needed

API

Medium-size download, customisable

One-off, or workflow integration, scripts, etc

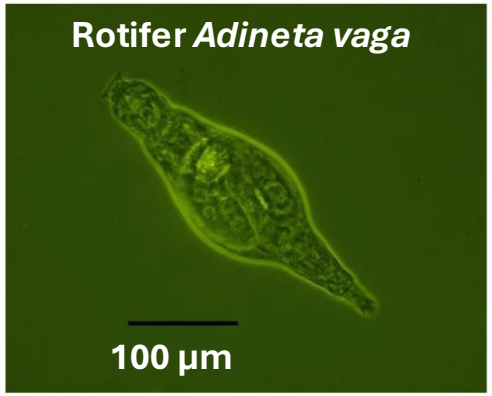
Website download

Small one-off download, customisable

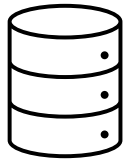
EMBL-EBI PIR SIB

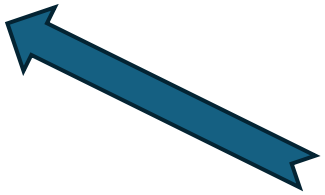
Ways to get data from your custom database

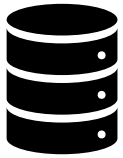
- Genome contigs
- Genes
- Protein
- BLAST output

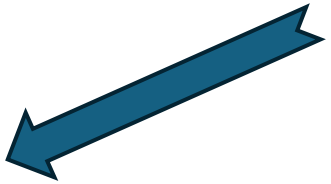


Perl
Python









grep_fasta.pl

```
#!/usr/bin/perl

use Bio::Perl;
use Bio::SeqIO;
use IO::String;
use Bio::SearchIO;
```

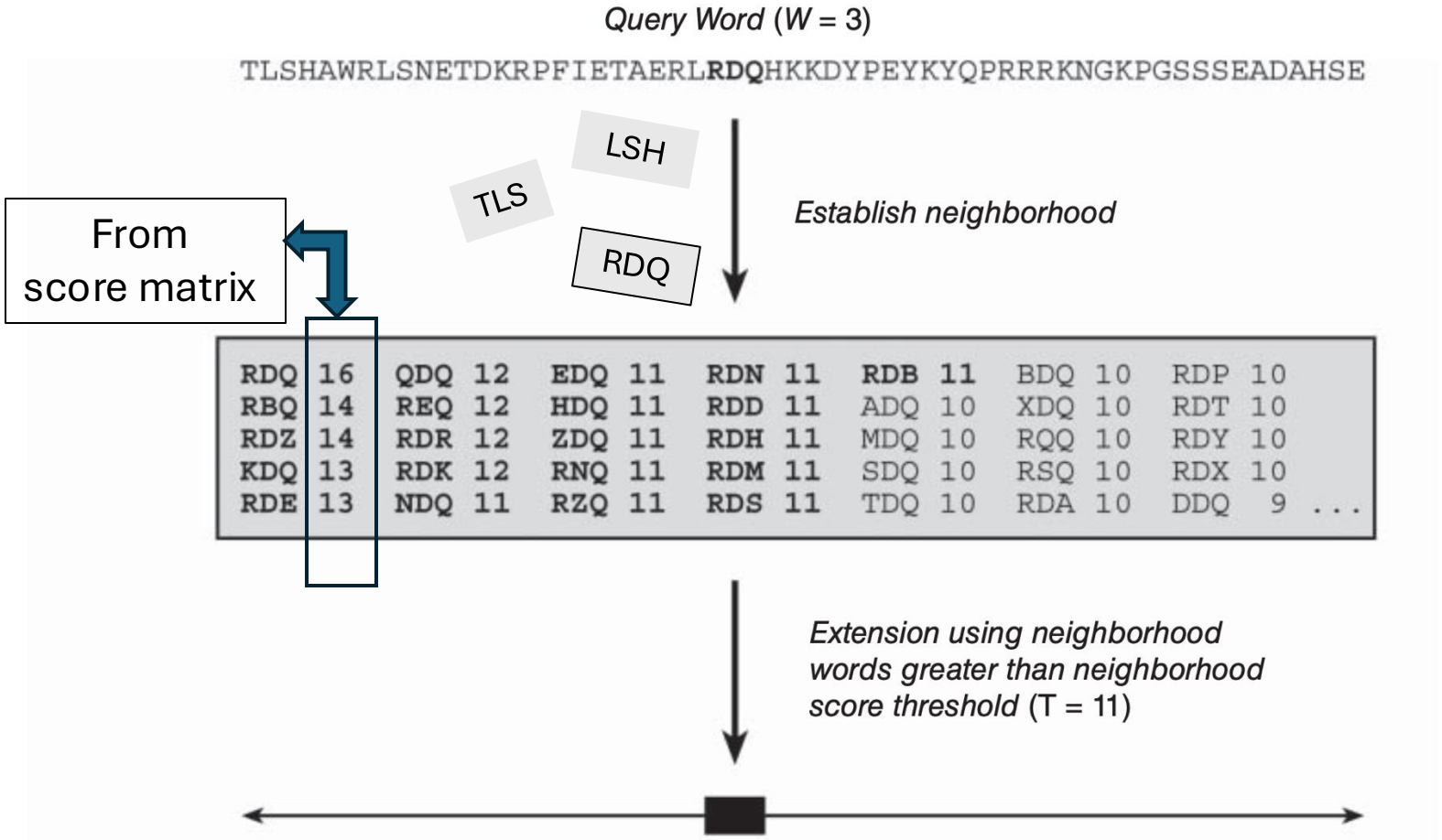
BLAST

- Basic Local Alignment Search Tool (Altschul et al. 1990) in Perusal.
- BLAST heuristically finds ***high-scoring segment pairs*** (**HSPs**):
  - Identical length segments from 2 sequences with statistically significant match scores
- Not only the best region of local alignment, but also whether there are other plausible alignments.
- BLAST method begins by seeding the search with a small subset of letters (query **word**).

BLAST algorithms

Program	Query	Database
BLASTN	Nucleotide	Nucleotide
BLASTP	Protein	Protein
BLASTX	Nucleotide, six-frame translation	Protein
TBLASTN	Protein	Nucleotide, six-frame translation
TBLASTX	Nucleotide, six-frame translation	Nucleotide, six-frame translation

The initiation of a BLAST search



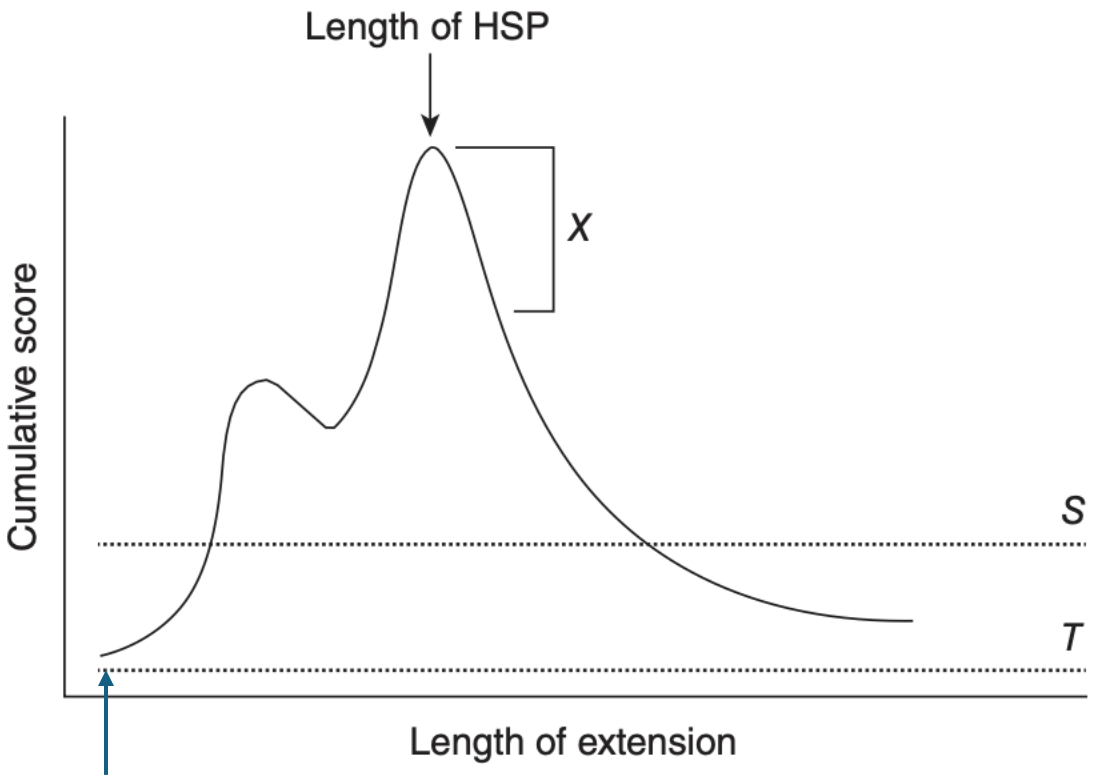
- **Query word** (w=3) is RDQ
- TLS, LSH, …RDQ

- BLAST search RDQ in the DB, but also related words (with conservative substitutions)
- **Neighborhood**: related words using scoring matrix

- We use a cut-off in the neighborhood: **score threshold** (T)

- Once the query (RDQ) word is aligned with another word (REQ) from DB (with T > 11).
- Now BLAST attempt to extend the alignment in both directions.

How to determine the maximal length of extension

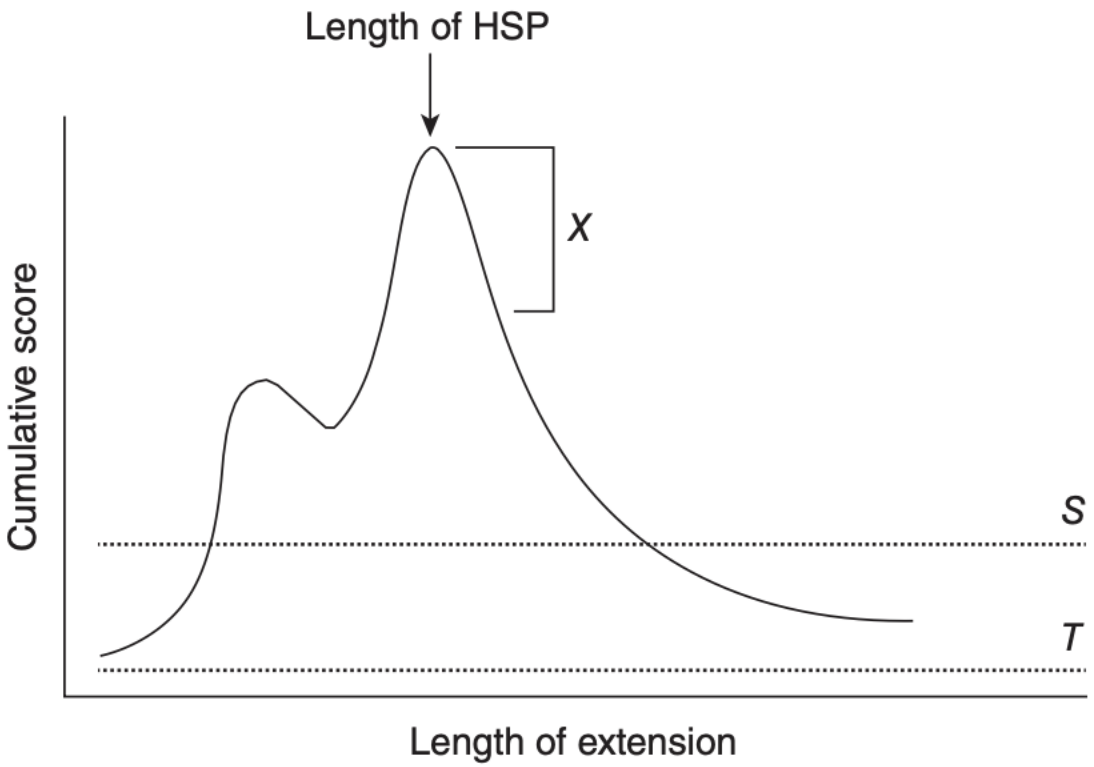


RDQ
| | |
REQ

- The number of residues (**extension**) is plotted vs. the **cumulative score** from the alignment.
- As the alignment gets extended:
  - matches with a positive score (conservative substitutions): score increase
  - Mismatches and gaps: score decrease
- As soon as the cumulative score breaks the **score threshold S**, the alignment is reported in the BLAST output.
- The resulting alignment is called a high-scoring segment pair, or HSP.

T: neighborhood score threshold
S: minimum score to return a BLAST hit
X: significance decay

Simple script implementation of BLAST



cascoamarillo / perl_vs_python

perl

python