Ensembl Compara Perl API

Matthieu Muffato
API workshop - EBI

May 2013
Outline of the course

- Introduction about Compara
  - Resources
  - API

- Base objects
  - Genes, peptides, RNAs
  - Multiple / pairwise alignments

- Data objects
  - Families
  - Gene trees
  - Homologies
Outline of the course

- Introduction about Compara
  - Resources
  - API

- Base objects
  - Genes, peptides, RNAs
  - Multiple / pairwise alignments

- Data objects
  - Families
  - Gene trees
  - Homologies
What is Ensembl Compara?

A single database which contains precalculated comparative genomics data and which is linked to all the Ensembl Species databases.

Access via perl API and mysql

A production system for generating that database

(not in this presentation)
Help & Useful documentation

• **perldoc** – Viewer for inline API documentation
  – `shell> perldoc Bio::EnsEMBL::Compara::GenomeDB`
  – `shell> perldoc Bio::EnsEMBL::Compara::DBSQL::MemberAdaptor`

• **Online documents (website)**

• **CVS**
  – [ensembl-compara/docs/protein_schema.png](http://www.ensembl.org/info/docs/protein_schema.png)
  – [ensembl-compara/docs/genomic_schema.png](http://www.ensembl.org/info/docs/genomic_schema.png)

• **ensembl-dev mailing list:**
  – dev@ensembl.org
  – helpdesk@ensembl.org
Compara data

**Genome level  *(this afternoon)*

- Whole genome alignments (pairwise and multiple)
- Syntenic regions (based on pair-wise align.)

**Gene level  *(now !)*

- Families (clusters of proteins + multiple align.)
- Gene trees (proteins, non-coding RNAs)
- Gene orthology / paralogy predictions
Outline of the course

• Introduction about Compara
  - Resources
  - API

• Base objects
  - Genes, peptides, RNAs
  - Multiple / pairwise alignments

• Data objects
  - Families
  - Gene trees
  - Homologies
The Compara Perl API

- Written in Object-Oriented Perl
- Used to retrieve data from and store data into the Ensembl Compara database
- Links species together for Ensembl website
- Generalized to extend to non-Ensembl genomic data (Uniprot)
- Follows same ‘Data Object’ & ‘Object Adaptor’ DBAdaptor design as the other Ensembl APIs
use strict;
use Bio::EnsEMBL::Registry;
my $reg = "Bio::EnsEMBL::Registry";

# Auto-configure the registry
$reg->load_registry_from_db(
    -host => "ensembldb.ensembl.org",
    -user => "anonymous"
);

# Get the adaptor object for the data type you want
# e.g. GeneTree
my $xx_adaptor = $reg->get_adaptor("Multi", "compara", "XX");

# Fetch the data objects using the adaptor
# e.g. get all the genes in a given gene tree
my $all_interesting_xx = $xx_adaptor->fetch_all_by_YY();

print "All XX objects from E!Compara :
";
foreach my $this_xx (@$all_interesting_xx) {
    # Do some stuff with the data object
    print "\t", $this_xx->stable_id, "\n";
}
Compara object model overview

**Primary Data**
- DnaFrag
  - GenomicAlignBlock
  - SyntenRegion
  - ConservationScore
  - ConstrainedElement

**Results**
- GenomeDB
  - NCBI Taxon
    - species
  - Member
    - genes
    - proteins
    - transcripts

**AlignedMemberSet**
- Family
- GeneTree
- Homology
Outline of the course

● Introduction about Compara
  - Resources
  - API

● Base objects
  - Genes, peptides, RNAs
  - Multiple / pairwise alignments

● Data objects
  - Families
  - Gene trees
  - Homologies
Overview of Member objects

GeneMember: ENSEMBLGENE

get_all_SeqMembers()
gene_member()
get_Gene()
get_Transcript()
get_Translation()

SeqMember: ENSEMBLPEP ENSEMBLTRANS

SeqMember: UNIPROT/SPTREMBL UNIPROT/SWISSPROT

Compara

Core

Gene
Transcript
Translation
(*)Member object / (*)MemberAdaptor

- **GeneMember / GeneMemberAdaptor** for genes
  - ENSEMBLGENE
- **SeqMember / SeqMemberAdaptor** for RNAs and proteins
  - ENSEMBLPEP, ENSEMBLTRANS, Uniprot/SPTREMBL, Uniprot/SWISSPROT

- **Objects share most their properties:**

  ```plaintext
  $seq_member_adaptor->fetch_by_source_stable_id(...)
  $gene_member_adaptor->fetch_all_by_source_taxon(...)
  ```

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stable ID</td>
<td>$member-&gt;stable_id()</td>
</tr>
<tr>
<td>Coordinates</td>
<td>$member-&gt;chr_name()</td>
</tr>
<tr>
<td></td>
<td>$member-&gt;chr_start() …</td>
</tr>
<tr>
<td>Sequence (SeqMember only)</td>
<td>$member-&gt;sequence()</td>
</tr>
<tr>
<td>Function</td>
<td>$member-&gt;description()</td>
</tr>
</tbody>
</table>
HOWTO: get a GeneMember from a gene symbol

- Compara only references genes by their Ensembl stable ID
- From a gene symbol, you first have to use the core API to get the stable id(s)
- Gene symbols may not be unique (for instance: U6)

```perl
# Get the Human gene adaptor
my $hg_adaptor = $reg->get_adaptor("human", "core", "Gene");

# Get all the genes
my $all_genes = $hg_adaptor->fetch_all_by_external_name(XX);

# For each gene
foreach my $gene (@{$all_genes}) {
    do some stuff with $gene->stable_id();
}
```
Exercises - Member

- Print the sequence of the Member corresponding to SwissProt protein O93279

- Find the Member(s) for the human ncRNA gene(s) FAM41C

- Find and print the sequence of all the peptide Members corresponding to the human protein-coding gene(s) FRAS1