Compara API Tutorial

Introduction

This tutorial is an introduction to the EnsEMBL Compara API. Knowledge of the EnsEMBL Core API and of the concepts and conventions in the EnsEMBL Core API tutorial is assumed. Documentation about the Compara database schema is available in ensembl-compara/docs/ from the EnsEMBL CVS repository, and while is not necessary for this tutorial, an understanding of the database tables may help as many of the adaptor modules are table-specific.

Installing the API

API installation and updating is the same as per the core API.

Connecting to an EnsEMBL Compara database

Connection parameters

Starting from rel.48 EnsEMBL is running two public MySQL servers on **host=ensembldb.ensembl.org** with two different port numbers. The server on **port=3306** hosts all databases prior to rel.48 and the server on **port=5306** hosts all newer databases starting from rel.48.

There are two API ways to connect to the EnsEMBL Compara database:

- In most cases you will prefer the implicit way - using Bio::EnsEMBL::Registry module, which can read either a global or a specific configuration file or auto-configure itself.
- However there are cases where you might want more flexibility provided by the explicit creation of a Bio::EnsEMBL::Compara::DBSQL::DBAdaptor.

Implicitly, using the Bio::EnsEMBL::Registry auto-configuration feature (recommended)

For using the auto-configuration feature, you will first need to supply the connection parameters to the Registry loader. For instance, if you want to connect to the the public EnsEMBL databases you can use the following command in your scripts:

```
use Bio::EnsEMBL::Registry;
Bio::EnsEMBL::Registry->load_registry_from_db(
    -host => 'ensembldb.ensembl.org',
    -user => 'anonymous',
    -port => 5306);
```

This will initialize the Registry, from which you will be able to create object-specific adaptors later. Alternatively, you can use a shorter version based on a URL:

```
use Bio::EnsEMBL::Registry;
Bio::EnsEMBL::Registry->load_registry_from_url('mysql://anonymous@ensembldb.ensembl.org:5306/');
```

Implicitly, using the Bio::EnsEMBL::Registry configuration file

You will need to have a registry configuration file set up. By default, it takes the file defined by the ENSEMBL_REGISTRY environment variable or the file named .ensembl_init in your home directory if the former is not found. Additionally, you can use a specific file (see perldoc Bio::EnsEMBL::Registry or later in this document for some examples on how to use a different file). Please, refer to the EnsEMBL registry documentation for details about this option.

Explicitly, using the Bio::EnsEMBL::Compara::DBSQL::DBAdaptor

EnsEMBL Compara data, like core data, is stored in a MySQL relational database. If you want to access a Compara database, you will need to connect to it. This is done in exactly the same way as to connect to an EnsEMBL core database, but using a Compara-specific DBAdaptor. One parameter you have to supply in addition to the ones needed by the Registry is the **-dbname**, which by convention contains the release number:

```
use Bio::EnsEMBL::Compara::DBSQL::DBAdaptor

my $host   = 'ensembldb.ensembl.org';
my $user   = 'anonymous';
my $port   = 5306;
my $dbname = 'ensembl_compara_51';

my $comparadb= new Bio::EnsEMBL::Compara::DBSQL::DBAdaptor(
    -host   => $host,
    -port   => $port,
    -user   => $user,
    -dbname => $dbname,
);
```

EnsEMBL Compara object-specific adaptors

EnsEMBL Compara adaptors are used to fetch data from the database. Data are returned as EnsEMBL objects. For instance, the GenomeDBAdaptor returns Bio::EnsEMBL::Compara::GenomeDB objects.

Below is a non-exhaustive list of EnsEMBL Compara adaptors that are most often used:

- GenomeDBAdaptor to fetch Bio::EnsEMBL::Compara::GenomeDB objects
- DnaFragAdaptor to fetch Bio::EnsEMBL::Compara::DnaFrag objects
- GenomicAlignBlockAdaptor to fetch Bio::EnsEMBL::Compara::GenomicAlignBlock objects
- DnaAlignFeatureAdaptor to fetch Bio::EnsEMBL::DnaDnaAlignFeature objects (note that this adaptor returns an EnsEMBL Core object)
- SyntenyAdaptor to fetch Bio::EnsEMBL::Compara::SyntenyRegion objects
- MemberAdaptor to fetch Bio::EnsEMBL::Compara::Member objects
- HomologyAdaptor to fetch Bio::EnsEMBL::Compara::Homology objects
- FamilyAdaptor to fetch Bio::EnsEMBL::Compara::Family objects

- PeptideAlignFeatureAdaptor to fetch Bio::EnsEMBL::Compara::PeptideAlignFeature objects

Only some of these adaptors will be used for illustration as part of this tutorial through commented perl scripts code.

You can get the adaptors from the Registry with the get_adaptor command. You need to specify three arguments: the species name, the type of database and the type of object. Therefore, in order to get the GenomeDBAdaptor for the Compara database, you will need the following command:

```
my $genome_db_adaptor = Bio::EnsEMBL::Registry->get_adaptor(
    'Multi', 'compara', 'GenomeDB');
```

NB: As the EnsEMBL Compara DB is a multi-species database, the standard species name is 'Multi'. The type of the database is 'compara'.

Code Conventions

Refer to the EnsEMBL core tutorial for a good description of the coding conventions normally used in EnsEMBL.

We can divide the fetching methods of the ObjectAdaptors into two categories: the **fetch_by** and **fetch_all_by**. The former return one single object while the latter return a reference to an array of objects.

```
my $this_genome_db = $genome_db_adaptor->fetch_by_name_assembly("Homo sapiens", "NCBI36");
```

```
my $all_genome_dbs = $genome_db_adaptor->fetch_all();
foreach my $this_genome_db (@{$all_genome_dbs}) {
  print $this_genome_db->name, "\n";
}
```

Whole Genome Alignments

The Compara database contains a number of different types of whole genome alignments. A listing about what are these different types can be found in the ensembl-compara/docs/schema_doc.html document in method_link section.

GenomicAlignBlock objects

GenomicAlignBlocks are the preferred way to store and fetch genomic alignments. A GenomicAlignBlock contains several GenomicAlign objects. Every GenomicAlign object corresponds to a piece of genomic sequence aligned with the other GenomicAlign in the same GenomicAlignBlock. A GenomicAlign object is always in relation with other GenomicAlign objects and this relation is defined through the GenomicAlignBlock object. Therefore the usual way to fetch genomic alignments is by fetching GenomicAlignBlock objects. We have to start by getting the corresponding adaptor:

```
# Getting the GenomicAlignBlock adaptor:
my $genomic_align_block_adaptor = Bio::EnsEMBL::Registry->get_adaptor(
    'Multi', 'compara', 'GenomicAlignBlock');
```

In order to fetch the right alignments we need to specify a couple of data: the type of alignment and the piece of genomic sequence in which we are looking for alignments. The type of alignment is a more tricky now: you need to specify both the alignment method and the set of genomes. In order to simply this task, you could use the new Bio::EnsEMBL::Compara::MethodLinkSpeciesSet object. The best way to use them is by fetching them from the database:

```
# Getting the GenomeDB adaptor:
my $genome_db_adaptor = Bio::EnsEMBL::Registry->get_adaptor(
    'Multi', 'compara', 'GenomeDB');

# Fetching GenomeDB objects for human and mouse:
my $human_genome_db = $genome_db_adaptor->fetch_by_name_assembly('homo_sapiens');
my $mouse_genome_db = $genome_db_adaptor->fetch_by_name_assembly('mus_musculus');

# Getting the MethodLinkSpeciesSet adaptor:
my $method_link_species_set_adaptor = Bio::EnsEMBL::Registry->get_adaptor(
    'Multi', 'compara', 'MethodLinkSpeciesSet');

# Fetching the MethodLinkSpeciesSet object corresponding to LASTZ_NET alignments between human and mouse genomic sequences:
my $human_mouse_blastz_net_mlss =
    $method_link_species_set_adaptor->fetch_by_method_link_type_GenomeDBs(
        "LASTZ_NET",
        [$human_genome_db, $mouse_genome_db]
    );
```

There are two ways to fetch GenomicAlignBlocks. One uses Bio::EnsEMBL::Slice objects while the second one is based on Bio::EnsEMBL::Compara::DnaFrag objects for specifying the piece of genomic sequence in which we are looking for alignments.

```
my $query_species = 'human';
my $seq_region = '14';
my $seq_region_start = 75000000;
my $seq_region_end   = 75010000;

# Getting the Slice adaptor:
my $slice_adaptor = Bio::EnsEMBL::Registry->get_adaptor(
    $query_species, 'core', 'Slice');

# Fetching a Slice object:
my $query_slice = $slice_adaptor->fetch_by_region(
    'toplevel',
    $seq_region,
    $seq_region_start,
    $seq_region_end);

# Fetching all the GenomicAlignBlock corresponding to this Slice:
my $genomic_align_blocks =
```

```
    $genomic_align_block_adaptor->fetch_all_by_MethodLinkSpeciesSet_Slice(
      $human_mouse_blastz_net_mlss,
      $query_slice);
```

Here is an example script with all of this:

```
use strict;
use Bio::EnsEMBL::Registry;
use Bio::EnsEMBL::Utils::Exception qw(throw);
use Bio::SimpleAlign;
use Bio::AlignIO;
use Bio::LocatableSeq;
use Getopt::Long;

my $usage = qq{
perl DumpMultiAlign.pl
  Getting help:
    [--help]

  For the query slice:
    [--species species]
        Query species. Default is "human"
    [--coord_system coordinates_name]
        Query coordinate system. Default is "chromosome"
    --seq_region region_name
        Query region name, i.e. the chromosome name
    --seq_region_start start
    --seq_region_end end

  For the alignments:
    [--alignment_type method_link_name]
        The type of alignment. Default is "BLASTZ_NET"
    [--set_of_species species1:species2:species3:...]
        The list of species used to get those alignments. Default is
        "human:mouse". The names should correspond to the name of the
        core database in the registry_configuration_file or any of its
        aliases

  Ouput:
    [--output_format clustalw|fasta|...]
        The type of output you want. "clustalw" is the default.
    [--output_file filename]
        The name of the output file. By default the output is the
        standard output
};

my $species = "human";
my $coord_system = "chromosome";
my $seq_region = "14";
my $seq_region_start = 75000000;
my $seq_region_end = 75010000;
my $alignment_type = "BLASTZ_NET";
my $set_of_species = "human:mouse";
my $output_file = undef;
my $output_format = "clustalw";
my $help;

GetOptions(
    "help" => \$help,
    "species=s" => \$species,
    "coord_system=s" => \$coord_system,
    "seq_region=s" => \$seq_region,
    "seq_region_start=i" => \$seq_region_start,
    "seq_region_end=i" => \$seq_region_end,
    "alignment_type=s" => \$alignment_type,
    "set_of_species=s" => \$set_of_species,
    "output_format=s" => \$output_format,
    "output_file=s" => \$output_file);

# Print Help and exit
if ($help) {
    print $usage;
    exit(0);
}

if ($output_file) {
    open(STDOUT, ">$output_file") or die("Cannot open $output_file");
}

Bio::EnsEMBL::Registry->load_registry_from_db(
    -host => 'ensembldb.ensembl.org', -user => 'anonymous');

# Getting all the Bio::EnsEMBL::Compara::GenomeDB objects
my $genome_dbs;
my $genome_db_adaptor = Bio::EnsEMBL::Registry->get_adaptor(
    'Multi', 'compara', 'GenomeDB');

throw("Cannot connect to Compara") if (!$genome_db_adaptor);

foreach my $this_species (split(":", $set_of_species)) {
    my $this_meta_container_adaptor = Bio::EnsEMBL::Registry->get_adaptor(
        $this_species, 'core', 'MetaContainer');

    throw("Registry configuration file has no data for connecting to <$this_species>")
```

```
        if (!$this_meta_container_adaptor);

    my $this_production_name = $this_meta_container_adaptor->get_production_name;

    # Fetch Bio::EnsEMBL::Compara::GenomeDB object
    my $genome_db = $genome_db_adaptor->fetch_by_name_assembly($this_production_name);

    # Add Bio::EnsEMBL::Compara::GenomeDB object to the list
    push(@$genome_dbs, $genome_db);
}

# Getting Bio::EnsEMBL::Compara::MethodLinkSpeciesSet object
my $method_link_species_set_adaptor = Bio::EnsEMBL::Registry->get_adaptor(
    'Multi', 'compara', 'MethodLinkSpeciesSet');

my $method_link_species_set =
    $method_link_species_set_adaptor->fetch_by_method_link_type_GenomeDBs(
      $alignment_type,
      $genome_dbs);

throw("The database do not contain any $alignment_type data for $set_of_species!")
    if (!$method_link_species_set);

# Fetching the query Slice:
my $slice_adaptor = Bio::EnsEMBL::Registry->get_adaptor($species, 'core', 'Slice');

throw("Registry configuration file has no data for connecting to <$species>")
    if (!$slice_adaptor);

my $query_slice = $slice_adaptor->fetch_by_region('toplevel', $seq_region, $seq_region_start, $seq_region_end);

throw("No Slice can be created with coordinates $seq_region:$seq_region_start-".
    "$seq_region_end") if (!$query_slice);

# Fetching all the GenomicAlignBlock corresponding to this Slice:
my $genomic_align_block_adaptor = Bio::EnsEMBL::Registry->get_adaptor(
    'Multi', 'compara', 'GenomicAlignBlock');

my $genomic_align_blocks =
    $genomic_align_block_adaptor->fetch_all_by_MethodLinkSpeciesSet_Slice(
      $method_link_species_set,
      $query_slice);

my $all_aligns;

# Get a Bio::SimpleAlign object from every GenomicAlignBlock
foreach my $this_genomic_align_block (@$genomic_align_blocks) {
    my $simple_align = $this_genomic_align_block->get_SimpleAlign;
    push(@$all_aligns, $simple_align);
}

# print all the genomic alignments using a Bio::AlignIO object
my $alignIO = Bio::AlignIO->newFh(
    -interleaved => 0,
    -fh => \*STDOUT,
    -format => $output_format,
    -idlength => 10
);

foreach my $this_align (@$all_aligns) {
    print $alignIO $this_align;
}

exit;
```

Homologies and Protein clusters

All the homologies and families refer to Members. Homology objects store orthologous and paralogous relationships between Members and Family objects are clusters of Members.

Member objects

A Member represent either a gene or a protein. Most of them are defined in the corresponding EnsEMBL core database. For instance, the sequence for the human gene ENSG00000004059 is stored in the human core database.

The fetch_by_source_stable_id method of the MemberAdaptor takes two arguments. The first one is the source of the Member and can be:

- ENSEMBLPEP, derived from an EnsEMBL translation
- ENSEMBLGENE, derived from an EnsEMBL gene
- Uniprot/SWISSPROT, derived from a Uniprot/Swissprot entry
- Uniprot/SPTREMBL, derived from a Uniprot/SP-TrEMBL entry

The second argument is the identifier for the Member. Here is a simple example:

```
# get the MemberAdaptor
my $member_adaptor = Bio::EnsEMBL::Registry->get_adaptor(
    'Multi','compara','Member');

# fetch a Member
my $member = $member_adaptor->fetch_by_source_stable_id(
    'ENSEMBLGENE','ENSG00000004059');

# print out some information about the Member
print $member->chr_name, " ( ", $member->chr_start, " - ", $member->chr_end,
```

```
    " ): ", $member->description, "\n";
```

The Member object has several attributes:

- **source_name** and **stable_id** define this Member.
- **chr_name**, **chr_start**, **chr_end**, **chr_strand** locate this Member on the genome but are only available for ENSEMBLGENE and ENSEMBLPEP.
- **taxon_id** corresponds to the NCBI taxonomy identifier (see http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/ for more details).
- **taxon** returns a Bio::EnsEMBL::Compara::NCBITaxon object. From this object you can get additional information about the species.

```
my $taxon = $member->taxon;
print "common_name ", $taxon->common_name,"\n";
print "genus ", $taxon->genus,"\n";
print "species ", $taxon->species,"\n";
print "binomial ", $taxon->binomial,"\n";
print "classification ", $taxon->classification,"\n";
```

In our example the species is human, so the output will look like this:

```
common_name:    human
genus:          Homo
species:        sapiens
binomial:       Homo sapiens
classification: sapiens Homo Hominidae Catarrhini Haplorrhini Primates Euarchontoglires Eutheria Mammalia Euteleostomi Vertebrata Cra
```

Homology Objects

A Homology object represents either an orthologous or paralogous relationships between two or more Members.

Typically you want to get homologies for a given gene. The HomologyAdaptor has a fetching method called fetch_all_by_Member(). You will need the Member object for your query gene, therefore you will fetch the Member first like in this example:

```
# first you have to get a Member object. In case of homology is a gene, in
# case of family it can be a gene or a protein

my $member_adaptor = Bio::EnsEMBL::Registry->get_adaptor('Multi', 'compara', 'Member');
my $member = $member_adaptor->fetch_by_source_stable_id('ENSEMBLGENE','ENSG00000004059');

# then you get the homologies where the member is involved

my $homology_adaptor = Bio::EnsEMBL::Registry->get_adaptor('Multi', 'compara', 'Homology');
my $homologies = $homology_adaptor->fetch_all_by_Member($member);

# That will return a reference to an array with all homologies (orthologues in
# other species and paralogues in the same one)
# Then for each homology, you can get all the Members implicated

foreach my $homology (@{$homologies}) {
  # You will find different kind of description
  # UBRH, MBRH, RHS, YoungParalogues
  # see ensembl-compara/docs/docs/schema_doc.html for more details

  print $homology->description," ", $homology->subtype,"\n";

  # And if they are defined dN and dS related values

  print " dn ", $homology->dn,"\n";
  print " ds ", $homology->ds,"\n";
  print " dnds_ratio ", $homology->dnds_ratio,"\n";
}
```

Each homology relation has 2 or more members, you should find there the initial member used as a query. The get_all_Members method returns an array of Member objects. The Member is actually an AlignedMember (for the underlying protein) and contains information about how this Member has been aligned.

```
my $homology = $homologies->[0]; # take one of the homologies and look into it

foreach my $member (@{$homology->get_all_Members}) {

  # each AlignedMember contains both the information on the Member and in
  # relation to the homology

  print (join " ", map { $member->$_ } qw(stable_id taxon_id))."\n";
  print (join " ", map { $member->$_ } qw(perc_id perc_pos perc_cov))."\n";

}
```

You can get the original alignment used to define an homology:

```
use Bio::AlignIO;

my $simple_align = $homology->get_SimpleAlign();
my $alignIO = Bio::AlignIO->newFh(
    -interleaved => 0,
    -fh => \*STDOUT,
    -format => "clustalw",
    -idlength => 20);

print $alignIO $simple_align;
```

Family Objects

Families are clusters of proteins including all the EnsEMBL proteins plus all the metazoan SwissProt and SP-Trembl entries. The object and the adaptor are really similar to the previous ones.

```
my $member_adaptor = Bio::EnsEMBL::Registry->get_adaptor('Multi', 'compara', 'Member');
my $member = $member_adaptor->fetch_by_source_stable_id('ENSEMBLGENE','ENSG00000004059');

my $family_adaptor = Bio::EnsEMBL::Registry->get_adaptor('Multi','compara','Family');
my $families = $family_adaptor->fetch_all_by_Member($member);

foreach my $family (@{$families}) {
    print join(" ", map { $family->$_ }  qw(description description_score))."\n";

    foreach my $member (@{$family->get_all_Members}) {
        print $member->stable_id," ",$member->taxon_id,"\n";
    }

    my $simple_align = $family->get_SimpleAlign();
    my $alignIO = Bio::AlignIO->newFh(
        -interleaved => 0,
        -fh          => \*STDOUT,
        -format      => "phylip",
        -idlength    => 20);

    print $alignIO $simple_align;

    $simple_align = $family->get_SimpleAlign('cdna');
    $alignIO = Bio::AlignIO->newFh(
        -interleaved => 0,
        -fh          => \*STDOUT,
        -format      => "phylip",
        -idlength    => 20);

    print $alignIO $simple_align;
}
```

Further help

For additional information or help mail the ensembl dev mailing list. You will need to subscribe to this mailing list to use it.