Ensembl Compara Perl API

Stephen Fitzgerald

EBI - Wellcome Trust Genome Campus, UK

http://www.ebi.ac.uk/~stephenf/Workshops/EBI_may_2013/
Sequence types and outputs

Nucleotide sequence
- Pairwise alignments
- Multiple alignments
- Syntenic regions
- Conservation

Protein sequence
- Protein trees
- Homologues
- Families
Nucleotide sequence analyses

Pairwise Alignments
- BLASTZ-net
- LASTZ-net
- t-BLAT-net

Syntenic regions
- Only for species with chromosomal mappings

Multiple alignments
- Mercator-Pecan
- Enredo-Pecan-Ortheus

Conservation
- GERP Cons. Scores
- GERP Constr. Elements
Pairwise Alignments in Compara 71

- 43 tBLAT alignments of distantly related species
- 65 BlastZ/LastZ alignments of closely related species
- 26 analyses of species with karyotypes
Multiple Alignments in Compara 71

- Compara 71
  - Mercator
  - Pecan
  - EPO
  - EPO 2x
  - 20-way amniota vertebrates
  - 13-way eutherian mammals
  - 6-way primate
  - 5-way fish
  - 3-way birds
  - 36-way eutherian mammals
  - 8-way fish
Genomic alignments schema
Compara adaptors

# get a compara MethodLinkSpeciesSet adaptor
my $method_link_species_set_adaptor =
    Bio::EnsEMBL::Registry->get_adaptor(
        "Multi", "compara", "MethodLinkSpeciesSet");

# get a Compara DnaFrag Adaptor
my $dnafrag_adaptor = Bio::EnsEMBL::Registry->get_adaptor(
        "Multi", "compara", "DnaFrag");

# get a compara GenomicAlignBlock adaptor
my $genomic_align_block_adaptor =
    Bio::EnsEMBL::Registry->get_adaptor(
        "Multi", "compara", "GenomicAlignBlock");
Exercise GenomeDB

A GenomeDB is used to link the Compara database to each of the Core species databases. Print the name, assembly version and genebuild version for all the GenomeDBs in the compara db.

Hint: First you will need an adaptor of type "GenomeDB". Then use the fetch_all() method from the GenomeDB adaptor to bring back GenomeDB objects (these will be returned as an array-ref). Then get the name, assembly version and genebuild version from these GenomeDB objects.
Sample code output

homo_sapiens/MT/3071-3091	ttcagaccggagtaatccagg
gorilla_gorilla/MT/2491-2511	ttcagaccggagtaatccagg
bos_taurus/MT/2865-2885	ttcagaccggagtaatccagg
pongo_abelii/MT/2494-2514	ttcagaccggagcaatccagg
sus_scrofa/MT/3681-3701	ttcagaccggagcaatccagg
macaca_mulatta/MT/3021-3041	ttcagaccggagcaatccagg
mus_musculus/MT/2513-2533	ttcagaccggagcaatccagg
oryctolagus_cuniculus/MT/2508-2528	ttcagaccggagcaatccagg
rattus_norvegicus/MT/2502-2522	ttcagaccggagcaatccagg
canis_familiaris/MT/2504-2524	ttcagaccggagtaatccagg

************ ********
use strict;
use warnings;
use Data::Dumper;
use Bio::AlignIO;

use Bio::EnsEMBL::Registry;
#Bio::EnsEMBL::Registry->no_version_check(1);

# Auto-configure the registry
Bio::EnsEMBL::Registry->load_registry_from_db(
    -host=>'ensembldb.ensembl.org', -user=>'anonymous',
    -port=>'5306');

# set up an AlignIO to format SimpleAlign output
my $alignIO = Bio::AlignIO->newFh(-interleaved => 0,
    -fh => \\*STDOUT,
    -format => 'clustalw',
    -idlength => 20);
Sample code MTSS adaptors

# get a compara MethodLinkSpeciesSet adaptor
my $method_link_species_set_adaptor =
    Bio::EnsEMBL::Registry->get_adaptor(
        "Multi", "compara", "MethodLinkSpeciesSet");

# get the method_link_species_set (data object) from the adaptor
my $methodLinkSpeciesSet = $method_link_species_set_adaptor->
    fetch_by_method_link_type_species_set_name("EPO", "mammals");
Exercise GenomeDB

A GenomeDB is used to link the Compara database to each of the Core species databases. Print the name, assembly version and genebuild version for all the GenomeDBs in the compara db.

Hint: First you will need an adaptor of type "GenomeDB". Then use the fetch_all() method from the GenomeDB adaptor to bring back GenomeDB objects (these will be returned as an array-ref). Then get the name, assembly version and genebuild version from these GenomeDB objects.
Exercise DnaFrag

A DnaFrag represents a top-level SeqRegion in the Compara database. Print all the DnaFrgags for chimp.

Hint: First you will need an adaptor of type DnaFrag. Then use the fetch_all_by_GenomeDB_region() from the adaptor to bring back all the dnafrags associated with a region (the region you want is "chromosome").
Exercise MethodLinkSpeciesSet

The MethodLinkSpeciesSet is a central component in the Compara database, it stores information connecting the various analyses (method_link_type) with a set of species (species_set).

a) (ii) Print the total number of MethodLinkSpeciesSet entries stored in the database.

*Hint: use the MethodLinkSpeciesSet adaptor fetch_all() method.*

(iii) Print a unique list of method_link_types and a count of their number in the database.

*Hint: With the MethodLinkSpeciesSet adaptor use the "method->type" method to get all the "method_type" entries in the database. In order to get a unique set of method_types, use the returned values from the "method->type" method as keys in a hash (and increment this for the total number of entries of that type).*
Exercise MethodLinkSpeciesSet

b) Print a list of the species and their internal ids (dbIDs) for the 12 eutherian mammal EPO alignments.

Hint: use the MethodLinkSpeciesSet adaptor fetch_by_method_link_type_species_set_name() method, and then the MethodLinkSpeciesSet (object) "species_set_obj()->genome_dbs" method (this brings back a list-ref of genome_db objects). The method_link_type is 'EPO' and the species_set_name is 'mammals'.