Ensembl

Accessing Ensembl from Java and Jython

Craig Melsopp

BOSC July 2004

Ensembl

- Genomic databases
  - genes, transcripts, exons, sequence, Affymetrix probes, snps ...
- Multi-species
  - human, mouse, zebra fish, rat, chicken, mosquito, fugu, fruit fly, chimp, c. elegans, c. briggsae, honey bee, dog
- Availability
  - ensembldb.ensembl.org
  - Local mirrors

Access options

- Web site
  - www.ensembl.org
- SQL
  - mysql client
  - Java JDBC, Python MySQLdb, Perl DBI ...
- APIs
  - Perl
  - Java (Jython)

Java / Jython options

(a)JDBC + SQL
(b)Java API (ensj)

Why ensj?

- Easier (=faster development)
  - API simpler than schema + SQL
- More Stable
  - SQL brittle wrt schema changes
  - Ensj had 1 minor code breaking change in 3 years
- “ inline” SQL if needed
  - query adaptor

The big picture

- [TODO – draw this as pici]
  - Java App <> ensj <> db
  - Jython app <> ensj <> db
  - Jython app <>enembl ><>ensj <>db

Ensj overview

- Drivers and adaptors
  - genes = humanDriver.getGeneAdaptor().fetch(new Location(“ chromosome:22:20m-21m”))
- Flexible
  - 1 driver + multiple databases e.g. core + snps
- Extensible
  - Plug in drivers and adaptors e.g. compara

Example – Pseudo code

1. Create a driver
2. Get the gene adaptor
3. Get the genes for chromosome:22:20m-21m
4. Print the name and number of transcripts for all the genes

database.conf

```
host=ensembldb.ensembl.org
user=anonymous
database=homo_sapiens_core_22_34d

#port=3333
#password=secret
#ensembl_driver=org.ensembl.driver.plugin.compara.ComparaMySQ
LDriver
#connection_pool_size=4
#ensid_prefix=ENS
```

Example – Java

```
import org.ensembl.driver.*;
import org.ensembl.datamodel.*;

...

Driver human = DriverManager.load("database.conf");
GeneAdaptor ga = human.getGeneAdaptor();
Location loc = new Location(“chromosome:22:20m-21m”);
List genes = ga.fetch(loc);
for (int i=0; i<genes.size(); ++i) {
  Gene g = (Gene)genes.get(i);
  System.out.println(g.getAccessionID() + “ ”
                     + g.getTranscripts().size());
}
```

Example – Jython

```
from org.ensembl.driver import *
from org.ensembl.datamodel import *

human = DriverManager.load("database.conf")
ga = human.geneAdaptor
loc = Location(“chromosome:22:20m-21m”)
genes = ga.fetch(loc)
for gene in genes:
  print gene.accessionID + “ “ + gene.transcripts.size()
```

Example – Jython (ensembl.py)

```
from ensembl import *

genes = human.ga.fetch(Location("chromosome:22:20m-21m"))
for gene in genes:
  print gene.accessionID + " " + gene.transcripts.size()
```

human is predefined by the ensembl module and always points to the latest human db on ensembldb.ensembl.org

Future

- Keep ensj synchronised with new schemas
- Optimizations
- Thread safety

Acknowledgments

- Arne Stabenau
- Glenn Proctor
- Vivek Lyer
- Ensembl team
- Users who reported bugs and contributed code

More information and Question time...

www.ensembl.org/java

ensembl-dev@ebi.ac.uk

craig@ebi.ac.uk