Biological Databases

Ignacio Medina
imedina@cipf.es
Index

- Introduction
- Biological Databases
- Ensembl Biomart
- CellBase
- Hands on

Ignacio Medina
imedina@cipf.es

Biological Databases
Introduction
NAR Biological Database Collection

- *Nucleic Acid Research* compilation currently lists 1512 online databases!
- Features:
  - Many different databases for each category, which should I use?
  - No standards: different IDs, methods, servers, formats, ...
  - Lack of international initiatives, many local and small databases
  - Different gene IDs, more than 50
  - *In vivo vs in silico* databases
Data in biology is open and available for all scientific community.

Microarrays repositories:
- ArrayExpress http://www.ebi.ac.uk/arrayexpress

NGS repositories:
- ENA http://www.ebi.ac.uk/ena/about/about
- 1000Genomes http://www.1000genomes.org
- ENCODE http://encodeproject.org
Introduction
Sequence databases

Genome Reference Consortium (GRC)


The GRC is a collaborative effort and only works with input from the larger scientific community.

We strive to work closely with external groups to gather all relevant data.

The GRC is now working to create assemblies that better represent this diversity and provide more robust substrates for genome analysis.

Ignacio Medina
imedina@cipf.es

Biological Databases
Mission:

- To provide freely available data and bioinformatics services to all facets of the scientific community in ways that promote scientific progress.

- To contribute to the advancement of biology through basic investigator-driven research in bioinformatics.

- To provide advanced bioinformatics training to scientists at all levels, from PhD students to independent investigators.

- To help disseminate cutting-edge technologies to industry.

Funded by EMBL

http://www.ebi.ac.uk/
Introduction
Sequence databases


National Center for Biotechnology Information (NCBI)

Set of tools and databases for genomic and biomedical studies and analysis.

Financed by the USA.
Is the American competence of the EBI in both objectives and resources.

Welcome to NCBI
The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

Get Started
- Tools: Analyze data using NCBI software
- Downloads: Get NCBI data or software
- How-To's: Learn how to accomplish specific tasks at NCBI
- Submissions: Submit data to GenBank or other NCBI databases

Human Microbiome Project
NIH Roadmap Initiative designed to characterize the community of microorganisms living on and in the human body.

Ignacio Medina
imedina@cipf.es

Biological Databases
Introduction

Functional Annotation

Some of the biological databases contain *Functional Information* of genes and sequences:

- **Functional databases**
  - KEGG pathways
  - Reactome
  - InterPro Motifs
  - Biocarta pathways
  - Gene Expression in tissues
  - Gene Ontology
    - Biological Process
    - Molecular Function
    - Cellular Component
  - Regulatory elements
    - miRNA
    - CisRed
    - Transcription Factor
    - Binding Sites
  - Bioentities from literature:
    - Diseases terms
    - Chemical terms
  - Keywords
    - Swissprot

**Overview**

- UniProt/Swiss-Prot
- UniProtKB/TrEMBL
- Ensembl IDs
- EntrezGene
- Affymetrix
- Agilent
- HGNC symbol
- EMBL acc
- RefSeq
- PDB
- Protein Id
- IPI ...

**Biological Databases**

Ignacio Medina
imedina@cipf.es
The *Gene Ontology* project provides a **controlled vocabulary** to describe gene and gene product attributes in any organism.

- Latest version has **38137** terms (March, 2013)
- The controlled vocabularies of terms are structured.

http://www.geneontology.org/
Introduction
Functional Annotation

The three categories of GO:

**Molecular Function**
The tasks performed by individual gene products; examples are transcription factor and DNA helicase

**Biological Process**
Broad biological goals, such as mitosis or purine metabolism, that are accomplished by ordered assemblies of molecular functions

**Cellular Component**
Subcellular structures, locations, and macromolecular complexes; examples include nucleus, telomere, and origin recognition complex

Gene Ontology (GO terms)

- **GO:0003673**: Gene Ontology (65883)
  - **GO:0008150**: biological process (44405)
    - **GO:0007610**: behavior (357)
      - **GO:0000004**: biological process unknown (7877)
    - **GO:0009987**: cellular process (32672)
      - **GO:0007154**: cell communication (5384)
      - **GO:0008219**: cell death (744)
      - **GO:0030154**: cell differentiation (464)
      - **GO:0008151**: cell growth and/or maintenance (28802)
      - **GO:0006928**: cell motility (911)
      - **GO:0006944**: membrane fusion (257)
    - **GO:0016265**: death (793)
    - **GO:0007275**: development (4615)
    - **GO:0008371**: obsolete (1581)
    - **GO:0007582**: physiological processes (31124)
    - **GO:0016032**: viral life cycle (115)
    - **GO:0005575**: cellular component (32869)
    - **GO:003674**: molecular function (53910)
Introduction
Functional Annotation

Gene Ontology (GO terms)

- **AmiGO** provides a web interface to search and browse the ontology and annotation data
  
  http://amigo.geneontology.org/cgi-bin/amigo/go.cgi

- **QuickGO** (EBI) provides also a web interface
  
  http://www.ebi.ac.uk/ego
It is a free, online, open-source, curated pathway database encompassing many areas of human biology. Information is authored by expert biological researchers.

http://www.reactome.org/
Introduction
Functional Annotation

- Involved in gene regulation.
- Last version has 16,772 entries (Release 17, April 2011).
- The **target database** contains computationally predicted targets for microRNAs across many species.

http://www.mirbase.org/

http://www.ebi.ac.uk/enright-srv/microcosm/htdocs/targets/v5/

Ignacio Medina
imedina@cipf.es

Biological Databases
The JASPAR database contains a curated, non-redundant set of profiles, derived from published collections of experimentally defined transcription factor binding sites for eukaryotes. The prime difference to similar resources (TRANSFAC, etc) consist of the open data access, non-redundancy and quality.

http://jaspar.genereg.net/
Introduction
Protein databases

UniProtKB/Swiss-Prot
contains 531473 sequence entries

UniProt, protein sequence and information
http://www.uniprot.org/

Welcome

The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.

What we provide

<table>
<thead>
<tr>
<th>UniProtKB</th>
<th>Protein knowledgebase, consists of two sections:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>⭐ Swiss-Prot, which is manually annotated and reviewed.</td>
</tr>
<tr>
<td></td>
<td>⭐ TrEMBL, which is automatically annotated and is not reviewed.</td>
</tr>
<tr>
<td></td>
<td>Includes Complete Proteome Sets.</td>
</tr>
</tbody>
</table>

| UniRef   | Sequence clusters, used to speed up sequence similarity searches. |

| UniParc  | Sequence archive, used to keep track of sequences and their identifiers. |

Supporting data: Literature citations, taxonomy, keywords and more.

Ignacio Medina
imedina@cipf.es

Biological Databases
Introduction
Protein databases

InterPro, protein annotation database

A centralized database of protein families, domains, repeats and sites in which identifiable features found in known proteins can be applied to new protein sequences.

http://www.ebi.ac.uk/interpro/

Member database information

<table>
<thead>
<tr>
<th>Signature Database</th>
<th>Version</th>
<th>Signatures</th>
<th>Integrated Signatures</th>
</tr>
</thead>
<tbody>
<tr>
<td>GENE3D</td>
<td>3.3.0</td>
<td>2386</td>
<td>1377</td>
</tr>
<tr>
<td>HAMAP</td>
<td>021210</td>
<td>1675</td>
<td>1429</td>
</tr>
<tr>
<td>PANTHER</td>
<td>7.0</td>
<td>80933</td>
<td>1777</td>
</tr>
<tr>
<td>PIRSF</td>
<td>2.74</td>
<td>3248</td>
<td>2791</td>
</tr>
<tr>
<td>PRINTS</td>
<td>41.1</td>
<td>2050</td>
<td>2009</td>
</tr>
<tr>
<td>PROSITE patterns</td>
<td>20.66</td>
<td>1308</td>
<td>1292</td>
</tr>
<tr>
<td>PROSITE profiles</td>
<td>20.66</td>
<td>901</td>
<td>877</td>
</tr>
<tr>
<td>Pfam</td>
<td>24.0</td>
<td>11912</td>
<td>11465</td>
</tr>
<tr>
<td>PfamB</td>
<td>24.0</td>
<td>142303</td>
<td>0</td>
</tr>
<tr>
<td>ProDom</td>
<td>2006.1</td>
<td>1894</td>
<td>1008</td>
</tr>
<tr>
<td>SMART</td>
<td>6.1</td>
<td>895</td>
<td>882</td>
</tr>
<tr>
<td>SUPERFAMILY</td>
<td>1.73</td>
<td>1774</td>
<td>1154</td>
</tr>
<tr>
<td>TIGRFAMs</td>
<td>9.0</td>
<td>3808</td>
<td>3796</td>
</tr>
</tbody>
</table>

Contents of InterPro 31.0 (Feb 2011)

- Active site: 97
- Binding site: 65
- Conserved site: 615
- Domain: 5936
- Family: 14194
- PTM: 16
- Repeat: 262

Ignacio Medina
imedina@ciapf.es
**IntAct**, protein-protein interaction database

http://www.ebi.ac.uk/intact/main.xhtml

*IntAct* provides a freely available, open source database system and analysis tools for protein interaction data. All interactions are derived from literature curation or direct user submissions and are freely available.
Introduction

Variation databases

dbSNP, the repository of all the SNPs


August 2011

Ignacio Medina
imedina@cipf.es

Biological Databases
Introduction
Variation databases

HapMap, Human Haplotype Map

The goal is to develop a public resource that will help researchers find genes associated with human disease and response to pharmaceuticals.
Introduction

Variation databases

Mutations: OMIM, COSMIC, Mitelman, etc.


http://www.sanger.ac.uk/genetics/CGP/cosmic/

Ignacio Medina
imedina@cipf.es

Biological Databases
Introduction
Variation databases

https://www.ebi.ac.uk/ega

The European Genome-phenome Archive

Database of Genomic Variants Archive

Ignacio Medina
imedina@cipf.es

Biological Databases
Introduction
Genome annotation databases and browsers

Ensembl, the most used

http://www.ensembl.org/index.html
Introduction
Genome annotation databases and browsers

http://genome.ucsc.edu/index.html

UCSC Genome Browser

Ignacio Medina
imedina@cipf.es
Ensembl Biomart
Data mining with Biomart

http://www.ensembl.org/biomart/martview

- Oriented to small queries
- Easy interface for users
Motivation

- **Exponential growth** in the number and size of biological databases and repositories. Data size can reach **hundreds of gigabytes** and involves serious problems of data access through Internet and local disks.
- **Biological information is spread out** in different databases and repositories (~1512), using different identifiers → Problem when analyzing genome-wide experiments.
- Researchers have to deal with **complex scripts** (in Perl, generally) or parse horribles XML files.

Goals

- **Integrate and Join** the most relevant and high quality biological information in a single database.
- Facilitate **accessibility** to users.
CellBase Overview

- A comprehensive integrative database and RESTful Web Services API.
- More than 220GB of data and ~100 SQL tables containing the most relevant biological information.
- Available for 11 species: human, mouse, rat, zebrafish, fruitfly, worm, yeast, dog, pig, mosquito and plasmodium.
- Accessible via RESTful web services and by a Perl client.

Java RESTful Web Services API design

Structure
- ws.bioinfo.cipf.es/cellbase/rest/(version)/(species)/(category)/(subcategory)/id/(resource)?(filters)

Categories
- genomic
  - ws.bioinfo.cipf.es/cellbase/rest/(version)/(species)/genomic/(subcategory)/id/(resource)
  - Example: ws.bioinfo.cipf.es/cellbase/rest/latest/human/genomic/region/15:32973105-32973105
- feature
  - ws.bioinfo.cipf.es/cellbase/rest/latest/human/genomic/feature/(subcategory)/id/(resource)
  - Example: ws.bioinfo.cipf.es/cellbase/rest/latest/human/genomic/feature/ENCSR000B3J/ENCSR000B3J
- regulatory
  - ws.bioinfo.cipf.es/cellbase/rest/(version)/(species)/regulatory/(subcategory)/id/(resource)
  - Example: ws.bioinfo.cipf.es/cellbase/rest/latest/human/regulatory/tf/TF binding
- network
  - ws.bioinfo.cipf.es/cellbase/rest/(version)/(species)/network/(subcategory)/id/(resource)
  - Example: ws.bioinfo.cipf.es/cellbase/rest/latest/human/network/pathway/123

Server
- CLI client has been implemented

Client
- TXT or JSON

Programmatic access

Web browser access

Usage in web applications

Data sources
- Core features: genes, transcripts, exons, proteins (UniProt), etc.
- Regulatory: TFBSs, miRNAs, regulatory regions, PWMs, conserved regions, etc.
- Functional annotation: GO ontologies (Gene ontology, disease ontology, etc.), InterPro, etc.
- Variation: dbSNP, HapMap, 1000 Genomes project, COSMIC, protein variants, etc.
- Systems biology: IntAct, Reactome, gene co-expression, etc.

Biological Databases

Ignacio Medina
imedina@cipf.es
CellBase
RESTful web services

General Structure
ws.bioinfo.cipf.es/cellbase/rest/{version}/{species}/{category}/{subcategory}/id/{resource}?{filters}


Categories

- Genomic
  Subcategories: region, variant and position

- Feature
  Subcategories: gene, transcript, exon, protein, snp and karyotype
  http://ws.bioinfo.cipf.es/cellbase/rest/latest/hsa/feature/snp/rs3934834/phenotype

- Regulatory
  Subcategories: mirna_gene, mirna_mature and tf

- Network
  Subcategories: pathway
CellBase, a comprehensive collection of RESTful web services for retrieving relevant biological information from heterogeneous sources

Marta Bleda¹,², Joaquin Tarraga¹,³, Alejandro de Maria¹, Francisco Salavert¹,², Luz Garcia-Alonso¹, Matilde Celma⁴, Ainoa Martin⁴, Joaquin Dopazo¹,²,³,* and Ignacio Medina¹,³,*

Published online 12 June 2012

doi:10.1093/nar/gks575
## Accessing CellBase information

<table>
<thead>
<tr>
<th>Category</th>
<th>Subcategory</th>
<th>ID</th>
<th>Resources</th>
</tr>
</thead>
<tbody>
<tr>
<td>genomic</td>
<td>position</td>
<td>chr:position</td>
<td>gene, snp</td>
</tr>
<tr>
<td></td>
<td>variant</td>
<td>chr:position:reference allele:other allele chr:position:other allele</td>
<td>snp_phenotype, mutation_phenotype, consequence_type</td>
</tr>
<tr>
<td></td>
<td>region</td>
<td>chr:start-end</td>
<td>Gene, transcript, exon, snp, mutation, structural_variation, sequence, tfbs, mirna_target, cpg_island, conserved_region, regulatory</td>
</tr>
<tr>
<td>feature</td>
<td>gene</td>
<td>Any gene ID</td>
<td>list, info, transcript, snp, mutation, tfbs, mirna_target, protein_feature</td>
</tr>
<tr>
<td></td>
<td>transcript</td>
<td>Ensembl transcript</td>
<td>info, all, gene, exon, sequence, mutation, protein_feature</td>
</tr>
<tr>
<td></td>
<td>snp</td>
<td>rsID</td>
<td>Info, consequence_type, regulatory, phenotype, population_frequency, xref</td>
</tr>
<tr>
<td></td>
<td>exon</td>
<td>Ensembl exon</td>
<td>Info, info, sequence, transcript</td>
</tr>
<tr>
<td></td>
<td>protein</td>
<td>Uniprot ID and Accession</td>
<td>Info, feature, xref, interaction</td>
</tr>
<tr>
<td></td>
<td>id</td>
<td>any</td>
<td>xref</td>
</tr>
<tr>
<td>regulatory</td>
<td>tf</td>
<td>Gene name</td>
<td>Info, tfbs, gene, target_gene, pwm, annotation</td>
</tr>
<tr>
<td></td>
<td>mirna_gene</td>
<td>miRBase gene ID</td>
<td>Info, target, disease</td>
</tr>
<tr>
<td></td>
<td>mirna_mature</td>
<td>miRBase gene Accession</td>
<td>Info, gene, mirna_gene, target_gene, target, disease, annotation</td>
</tr>
</tbody>
</table>
Accessing CellBase information

- Web services
  
  `ws.bioinfo.cipf.es/cellbase/rest/latest/hsa/feature/gene/BRCA2/transcript`

- Web client
  
  `http://bioinfo.cipf.es/apps-beta/cellbase.html`

- Perl RESTful WS client
  
  
  - Download the Perl client
  - Execute it!
    
    `./cellbase_client.pl --input-type gene --id BRCA2 --get transcript`
QUESTION 1:

We are interested in a particular region of the genome, 1:2201105-2319315, and we want to know if this region contains mutations already catalogued.

HELP:

- Version: latest
- Species: hsa
- Category: genomic
- Subcategory: region
- ID: 1:2201105-2319315
- Resource: mutation
Accessing CellBase information

**QUESTION 2:**
We want to know which **microRNAs** regulate the **gene BRCA2**.

**HELP:**
- **Version:** latest
- **Species:** hsa
- **Category:** feature
- **Subcategory:** gene
- **ID:** BRCA2
- **Resource:** mirna_target
QUESTION 3:

We want to know the allelic and **genotypic frequencies** for a SNP, rs158691, across populations.

HELP:
- Version: latest
- Species: hsa
- Category: feature
- Subcategory: snp
- ID: rs158691
- Resource: population_frequency
QUESTION 4:

We need to **convert** some gene names (BRCA2, PAEP, GATA2) into Ensembl Gene **identifiers**.

HELP:

- Version: latest
- Species: hsa
- Category: feature
- Subcategory: id
- ID: BRCA2, PAEP, GATA2
- Resource: xref
- Filter: dbname=ensembl_gene
QUESTION 5:
We have obtained a microRNA of interest (hsa-miR-149-3p) in our analysis and we want to know if it has been related with any disease.

HELP:
- Version: latest
- Species: hsa
- Category: regulatory
- Subcategory: mirna_mature
- ID: hsa-miR-149-3p
- Resource: disease
Accessing Ensembl BioMart information

http://www.ensembl.org/biomart/martview
**QUESTION 1:**

We need the complete list of **correspondences** between **genes** and their **transcripts**.

**HELP:**

- **Database:** Ensembl Genes 70
- **Dataset:** Homo sapiens genes
- **No filters.** We want all the information!
- **Attributes:** Gene > Ensembl Gene ID
  
  Ensembl Transcript ID
QUESTION 2:

We want to retrieve all known somatic variants that have been associated with a cancer and we just want those in COSMIC.

HELP:

- Database: Ensembl Variation 70
- Dataset: Homo sapiens Somatic Short Variation (SNPs and indels) Filters: GENERAL VARIATION FILTERS > COSMIC
- Attributes: Sequence variation > Variation Name, Variation source, Chromosome name, Position on Chromosome (bp), Phenotype description
Solutions to CellBase questions

**Question 1:**
CLI: ./cellbase_client.pl --input-type region --id 1:2201105-2319315 --get mutation

**Question 2:**
CLI: ./cellbase_client.pl --input-type gene --id BRCA2 --get mirna_target

**Question 3:**
WS: http://ws.bioinfo.cipf.es/cellbase/rest/latest/hsa/feature/snp/rs158691/population_frequency
CLI: ./cellbase_client.pl --input-type snp --id rs158691 --get population_frequency

**Question 4:**
CLI: ./cellbase_client.pl --input-type id --id BRCA2,PAEP,GATA2 --get xref?dbname=ensembl_gene

**Question 5:**
CLI: ./cellbase_client.pl --input-type regulatory --id hsa-miR-149-3p --get disease