Whole Genome databases (review)

Hélène Chiapello
Mathématique, Informatique & Génome Unit,
INRA - Jouy-en-Josas – France.
Helene.Chiapello@jouy.inra.fr
The MIG laboratory

mig.jouy.inra.fr

• A multidisciplinary research unit created in 2000 and grouping together mathematicians, computer scientists, physical chemists and biologists around the analysis of the genomes.

• The laboratory has three mains activities:
  (1) it is developing methods and tools for analyzing genomes in silico;
  (2) it is setting up information systems for genomics;
  (3) it is participating in genomic projects.

• The research activity of the lab is structured in three main subjects: genome and evolution, text mining and systems biology.

• The lab has also a bioinformatics platform: MIGALE migale.jouy.inra.fr which provide to the community of the biologists both logistic, software and training support.
Outline

I. General introduction

II. Genome databases

III. Related resources

IV. Conclusion, Challenges

V. Practical exercises
Complete Genomes

7464 genome projects
• 1305 published genomes: 1088 B., 85 A., 132 E.
• In progress: ~440 Bacteria, ~200 Archea, ~1300 Eukaryotes and ~210 Metagenomes.

Source: GOLD (06/07/2010)
www.genomesonline.org

• The 3 phylogenetic domains are covered
• Diverse lifestyles: mesophiles, thermophiles, extreme conditions,…

Bioinformatics and Comparative Genome Analysis
Institut Pasteur Paris, July 5th - July 17th, 2010
I. General Introduction

Evolution of genome projects

• 7464 genome projects (June 2010)
• Number and Size of projects grows at a rapid rate
• We have crossed the Terabyte threshold of genomic data
• These data need to be stored, curated and made available for analysis and knowledge discovery

Number of sequencing projects, 1995-2010
Source: GOLD www.genomesonline.org

Bioinformatics and Comparative Genome Analysis
Institut Pasteur Paris, July 5th - July 17th, 2010
I. General Introduction

Database definition

• What is a database?
  – An organized body of related data
  – It is presumed that (i) the volume of the data is large (ii) data have to be accessed, retrieved, updated “frequently”
  – **Database Management Systems (DBMS)** are software that facilitates the creation, administration and use of the database

• Some common database models:

<table>
<thead>
<tr>
<th>Flat File Model</th>
<th>Relational Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Route No.</td>
<td>EMPLOYEE</td>
</tr>
<tr>
<td>Miles</td>
<td>ID: NUMBER</td>
</tr>
<tr>
<td>Activity</td>
<td>F_NAME: VARCHAR</td>
</tr>
<tr>
<td>Record 1</td>
<td>L_NAME: VARCHAR</td>
</tr>
<tr>
<td>1-95</td>
<td>ADDRESS_ID: NUMBER</td>
</tr>
<tr>
<td>12</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Structured Files</th>
<th>Tables</th>
<th>Objects</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Bioinformatics and Comparative Genome Analysis
Institut Pasteur Paris, July 5th - July 17th, 2010
I. General Introduction

How big are genomes?

Top 20 organisms in public banks

<table>
<thead>
<tr>
<th>Species</th>
<th>Genome size</th>
<th>Bases</th>
<th>Entries</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homo sapiens</td>
<td>3,400,000,000</td>
<td>13,669,851,495</td>
<td>12,838,795</td>
</tr>
<tr>
<td>Mus musculus</td>
<td>3,454,200,000</td>
<td>8,445,993,792</td>
<td>7,347,636</td>
</tr>
<tr>
<td>Rattus norvegicus</td>
<td>2,900,000,000</td>
<td>6,284,206,670</td>
<td>1,997,976</td>
</tr>
<tr>
<td>Bos taurus</td>
<td>3,651,500,000</td>
<td>5,319,815,212</td>
<td>2,135,747</td>
</tr>
<tr>
<td>Zea mays</td>
<td>5,000,000,000</td>
<td>5,007,807,286</td>
<td>3,870,406</td>
</tr>
<tr>
<td>Sus scrofa</td>
<td>3,108,700,000</td>
<td>4,229,790,475</td>
<td>2,556,492</td>
</tr>
<tr>
<td>Danio rerio</td>
<td>1,900,000,000</td>
<td>3,074,615,557</td>
<td>1,695,362</td>
</tr>
<tr>
<td>Stronglyocentrotus purpur</td>
<td>900,000,000</td>
<td>1,352,840,985</td>
<td>228,153</td>
</tr>
<tr>
<td>Nicotiana tabacum</td>
<td>900,000,000</td>
<td>1,184,330,809</td>
<td>1,752,654</td>
</tr>
<tr>
<td>Oryza sativa Japonica Gro</td>
<td>900,000,000</td>
<td>1,176,024,629</td>
<td>1,217,983</td>
</tr>
<tr>
<td>Xenopus (Silurana) tropic</td>
<td>900,000,000</td>
<td>1,146,732,476</td>
<td>1,423,046</td>
</tr>
<tr>
<td>Drosophila melanogaster</td>
<td>180,000,000</td>
<td>1,038,512,618</td>
<td>1,202,127</td>
</tr>
<tr>
<td>Pan troglodytes</td>
<td>3,577,500,000</td>
<td>997,816,950</td>
<td>213,217</td>
</tr>
<tr>
<td>Arabidopsis italiana</td>
<td>100,000,000</td>
<td>950,139,115</td>
<td>2,240,601</td>
</tr>
<tr>
<td>Canis lupus familiaris</td>
<td>100,000,000</td>
<td>931,176,470</td>
<td>1,434,100</td>
</tr>
<tr>
<td>Vitis vinifera</td>
<td>100,000,000</td>
<td>910,760,908</td>
<td>655,685</td>
</tr>
<tr>
<td>Gallus gallus</td>
<td>1,200,000,000</td>
<td>884,489,747</td>
<td>806,871</td>
</tr>
<tr>
<td>Glycine max</td>
<td>1,115,000,000</td>
<td>846,429,180</td>
<td>1,828,912</td>
</tr>
<tr>
<td>Macaca mulatta</td>
<td>3,543,000,000</td>
<td>808,403,289</td>
<td>78,410</td>
</tr>
<tr>
<td>Ciona intestinalis</td>
<td>200,000,000</td>
<td>748,153,905</td>
<td>1,216,132</td>
</tr>
<tr>
<td>Total</td>
<td>106,533,156,756</td>
<td>108,431,692</td>
<td></td>
</tr>
</tbody>
</table>

Genome typical sizes:

- Virus: 1 to 360 kb (mimivirus: 1.2 Mb !)
- Bacteria: 0.5 to 13 Mb
- Eukaryotes: 8 Mb to 670 Gb
I. General Introduction

What are the genomic data?

- Genomes
- Chromosomes
- Genes (cDNA, ESTs, RNAseq, ...)
- Nucleotide sequences
- Protein sequences
- Annotations
- Structures
- Etc...

Bioinformatics and Comparative Genome Analysis
Institut Pasteur Paris, July 5th - July 17th, 2010
I. General Introduction

Basic definitions

Genome
The genome of a cell is formed by the collection of the DNA it comprises. The genome size is the total of its DNA bases.

Gene
Is a particular DNA sequence situated in a specific position on a chromosome and that codes for a specific function.

Protein
Is a sequence composed of amino-acids ordered according to the DNA sequences of the gene it codes for.

Proteome
Is the set of proteins in an organism.

Genomics
Is the exhaustive study of a genome of an organism: genetic material, genes; their functions, their organization....

Metagenomics
It describes the analysis of microbial genomes contained in an environmental sample (no clonal culture).
Main resources for genomes

II. Genome databases

- Two main international resources:
  - EBI: European Bioinformatics Institute
    [www.ebi.ac.uk/genomes](http://www.ebi.ac.uk/genomes)
  - NCBI: National Center for Biotechnology Information

- Many other resources from sequencing institutions:
  - Sanger: the Welcome Trust Sanger Institute
    [www.sanger.ac.uk](http://www.sanger.ac.uk)
  - JCVI: the Craig Venter Institute (formerly TIGR)
    [www.jcvi.org](http://www.jcvi.org)
  - JGI: the Joint Genome Institute (DOE)
    [genome.jgi-psf.org](http://genome.jgi-psf.org)
  - Broad: the Broad Institute (Harvard, MIT)
    [www.broadinstitute.org](http://www.broadinstitute.org)
  - Genoscope: the french sequencing center
    [www.genoscope.cns.fr](http://www.genoscope.cns.fr)
  - ....
II. Genome databases

Browsing genomes at EBI

ResearchGate is a social networking platform for scientists and researchers. It allows scientists to connect with one another, share research, and collaborate on projects.

Access to Completed Genomes

The first completed genomes from *viruses*, *phages* and *organelles* were deposited into the EMBL Database in the early 1980s. Since then, molecular biology's shift to obtain the complete sequences of as many genomes as possible combined with major developments in sequencing technology resulted in hundreds of complete genome sequences being added to the database, including *Arabidopsis*, *Bacillus* and *Pulvinus*. These web pages give access to a large number of complete genomes. Help is available to describe the layout.

Whole Genome Shotgun Sequences (WGS)

Methods using whole genome shotgun data are used to gain a large amount of genome coverage for an organism. WGS data for a growing number of organisms are being submitted to DDBJ/EMBL/GenBank and are made available via EBI's Sequence Retrieval System (SRS) at http://srs.ebi.ac.uk and the EBI FTP server at ftp.ebi.ac.uk/pub/databases/embl/wgs. More information about WGS projects...

Human Draft Genome

The completion of the human draft genome sequence was announced and published in February 2001 in *Nature* and *Science*. Since the beginning of the Human Genome Project, the international Human Genome Sequencing Consortium has been submitting human draft sequence data to the International Nucleotide Sequence Databases DDBJ/EMBL/GenBank. High-throughput human sequences have been made available to the public immediately via the EMBL Database high-throughput genome division (HTG), while finished sequences have been included in the Human division (HUM).

Genome Annotation and Proteome Analysis

Genome Reviews are a set of reannotated proteomes, based on manually curated data from UniProt.

The *Ensembl Genome Browser* provides the best possible automatic annotation, graphical views and web searchable datasets for a number of eukaryotic genomes including human, mouse, drosophila, anopheles, zebrafish with others to follow.

*Internet* has proteome analysis information on a large number of organisms.

Last 40 Genome Entries

NB The genomes pages are normally updated ahead of SRS, if links fail to work for recent genomes, please try in 1-2 days.

<table>
<thead>
<tr>
<th>Date</th>
<th>Accession</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>01-JUL-2010</td>
<td>CP002185.1</td>
<td><em>Serratia</em> rothous DSM 44685</td>
</tr>
<tr>
<td>27-MAY-2010</td>
<td>GQ048411.1</td>
<td>Human <em>pseudomarus</em> type 119</td>
</tr>
<tr>
<td>27-MAY-2010</td>
<td>GQ048412.1</td>
<td>Human <em>pseudomarus</em> type 120</td>
</tr>
<tr>
<td>27-MAY-2010</td>
<td>GQ048413.1</td>
<td>Human <em>pseudomarus</em> type 121</td>
</tr>
<tr>
<td>27-MAY-2010</td>
<td>GQ048414.1</td>
<td>Human <em>pseudomarus</em> type 122</td>
</tr>
<tr>
<td>27-MAY-2010</td>
<td>GQ048415.1</td>
<td>Human <em>pseudomarus</em> type 123</td>
</tr>
</tbody>
</table>
II. Genome databases

EBI: genomes of archaea
II. Genome databases

EBI: genome of A. pernix K1

Aeropyrum pernix (strain K1) - Tax ID: 272557  QAS: 12345

A strictly aerobic hyperthermophilic archaean isolated from a coastal sulfataric thermal vent at Kodakara-Jima Island, Japan in 1993. It is a heterotroph that grows optimally at 90 to 95 degrees Celcius, pH 7.0, and a salinity of 3.5%. It is spherical shaped and covered by a cell envelope (S-layer-like structure). It contains (25-isoprenyl)archael (glycerol diether) as the hydrophobic chain in the core lipid.
II. Genome databases

EBI: genomes of eukaryotes

<table>
<thead>
<tr>
<th>Description</th>
<th>Length (bp)</th>
<th>Sequence</th>
<th>Project</th>
<th>Proteins</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anopheles gambiæ</td>
<td>5,066,924</td>
<td>Y58501</td>
<td>Y96501</td>
<td></td>
</tr>
<tr>
<td>Arachniotes thaliæ</td>
<td>366,924</td>
<td>Y58501</td>
<td>Y96501</td>
<td></td>
</tr>
<tr>
<td>Aspergillus niger</td>
<td>4,142,115</td>
<td>AA000172</td>
<td>AA000172</td>
<td></td>
</tr>
<tr>
<td>Aspergillus niger</td>
<td>4,142,115</td>
<td>AA000172</td>
<td>AA000172</td>
<td></td>
</tr>
<tr>
<td>Aspergillus niger</td>
<td>4,142,115</td>
<td>AA000172</td>
<td>AA000172</td>
<td></td>
</tr>
</tbody>
</table>

The table contains information on various species and their genome lengths, along with sequence information and project details.
II. Genome databases

EBI: download genomic data of *A. thaliana*

---

**Bioinformatics and Comparative Genome Analysis**

**Institut Pasteur Paris, July 5th - July 17th, 2010**
II. Genome databases

Browsing genomes at NCBI
II. Genome databases

Prokaryotes genomes at NCBI

![Genome Project screenshot]
II. Genome databases

**Acetobacter Pasteurianus** proteome at NCBI

![Image of NCBI genome result for Acetobacter pasteurianus IFO 3283-01, complete genome]
II. Genome databases

Other genome databases

• First generation resources often focused on the genome of a model species, for example:
  - Subtilist (\textit{B. subtilis}) \texttt{genolist.pasteur.fr/SubtiList}
  - WormBase (\textit{C. elegans}) \texttt{www.wormbase.org}

• Second generation resources integrate genomic data of a category of organisms with sequenced genomes, for example:
  - Genolist (bacteria) \texttt{genolist.pasteur.fr}
  - Genolevures (yeast) \texttt{www.genolevures.org}
  - Ensembl (vertebrates) \texttt{www.ensembl.org}
  - …
## Genome databases by phylogenetic groups

### Procaryotes
- **GenomeReviews**: [www.ebi.ac.uk/GenomeReviews](http://www.ebi.ac.uk/GenomeReviews)
- **ASAP**: [asap.ahabs.wisc.edu/asap/ASAP1.htm](http://asap.ahabs.wisc.edu/asap/ASAP1.htm)
- **IMG**: [img.jgi.doe.gov](http://img.jgi.doe.gov)

### Animals
- **Ensembl**: [www.ensembl.org](http://www.ensembl.org)
- **FlyBase**: [www.informatics.jax.org](http://www.informatics.jax.org)
- **MGD**: [www.informatics.jax.org](http://www.informatics.jax.org)
- **UCSC**: [www.genome.ucsc.edu](http://www.genome.ucsc.edu)
- **AceDB**: [www.acedb.org](http://www.acedb.org)

### Plants
- **TAIR**: [www.arabidopsis.org](http://www.arabidopsis.org)
- **FlagDB++**: [urgv.evry.inra.fr/FLAGdb](http://urgv.evry.inra.fr/FLAGdb)
- **MIPSPlantsDB**: [mips.gsf.de/proj/plant/jsf/index.jsp](http://mips.gsf.de/proj/plant/jsf/index.jsp)

### Fungi
- **YPD**: [www.incyte.com/sequence/proteome/databases/YPD.shtml](http://www.incyte.com/sequence/proteome/databases/YPD.shtml)
- **M-CYGD**: [mips.gsf.de/genre/proj/yeast](http://mips.gsf.de/genre/proj/yeast)
## II. Genome databases

### Genome databases by phylogenetic groups

#### Procaryotes
- **GenomeReviews**: [www.ebi.ac.uk/GenomeReviews](http://www.ebi.ac.uk/GenomeReviews)
- **ASAP**: [asap.ahabs.wisc.edu/asap/ASAP1.htm](http://asap.ahabs.wisc.edu/asap/ASAP1.htm)
- **IMG**: [img.jgi.doe.gov](http://img.jgi.doe.gov)

#### Animals
- **Ensembl**: [www.ensembl.org](http://www.ensembl.org)
- **FlyBase**: [www.informatics.jax.org](http://www.informatics.jax.org)
- **MGD**: [www.informatics.jax.org](http://www.informatics.jax.org)
- **UCSC**: [www.genome.ucsc.edu](http://www.genome.ucsc.edu)
- **AceDB**: [www.acedb.org](http://www.acedb.org)

#### Plants
- **TAIR**: [www.arabidopsis.org](http://www.arabidopsis.org)
- **FlagDB++**: [urgv.evry.inra.fr/FLAGdb](http://urgv.evry.inra.fr/FLAGdb)
- **MIPSPlantsDB**: [mips.gsf.de/proj/plant/jsf/index.jsp](http://mips.gsf.de/proj/plant/jsf/index.jsp)

#### Fungi
- **YPD**: [www.incyte.com/sequence/proteome/databases/YPD.shtml](http://www.incyte.com/sequence/proteome/databases/YPD.shtml)
- **M-CYGD**: [mips.gsf.de/genre/proj/yeast](http://mips.gsf.de/genre/proj/yeast)
II. Genome databases

The NAR database issue: a reference

• The 2010 update includes 1230 databases:
  58 new databases, updates of 73 existing databases

• Category list: Nucleotide, RNA, Protein sequence databases,
  Genomics databases, Structure databases,…

• The complete database list and summaries are available online on
  the Nucleic Acids Research web site
  http://www.oxfordjournals.org/nar/database/a/
III. Related resources

Comparative genomics resources

UCSC Genome4 Bioinformatics: genome.ucsc.edu
Ensembl: www.ensembl.org
VISTA Genome Browser: pipeline.lbl.gov
Comparative Regulatory Genomics: corg.molgen.mpg.de
GALA: www.bx.psu.edu
EnsMart: www.ensembl.org/EnsMart
PipMaker and MultiPipMaker: www.bx.psu.edu
VISTA server: www-qsd.lbl.gov/vista
MAVID server: baboon.math.berkeley.edu/mavid
zPicture server: zpicture.dcode.org
rVISTA server: rvista.dcode.org
COGs: www.ncbi.nlm.nih.gov/COG
MOSAIC: genome.jouy.inrafr/mosaic
### III. Related resources

#### "Omics" databases

<table>
<thead>
<tr>
<th><strong>Transcriptomics</strong></th>
<th><strong>Proteomics</strong></th>
<th><strong>Protein-Protein interactions</strong></th>
<th><strong>Metabolomics</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>SMD</td>
<td>SWISS-2DPAGE</td>
<td>DIP</td>
<td>Kegg</td>
</tr>
<tr>
<td>ArrayExpress</td>
<td>OPD</td>
<td>BIND</td>
<td>Metacyc</td>
</tr>
<tr>
<td><a href="http://www.ebi.ac.uk/microarray/ArrayExpress/arrayexpress.html">www.ebi.ac.uk/microarray/ArrayExpress/arrayexpress.html</a></td>
<td>bioinformatics.icmb.utexas.edu/OPD</td>
<td><a href="http://www.bind.ca">www.bind.ca</a></td>
<td>biocyc.org/meta/</td>
</tr>
<tr>
<td>GEO</td>
<td>PARIS</td>
<td>BRITE</td>
<td>WIT</td>
</tr>
</tbody>
</table>

---

**Bioinformatics and Comparative Genome Analysis**  
**Institut Pasteur Paris, July 5th - July 17th, 2010**
III. Related resources

Databases of Motifs and Mobile elements

Regulation motifs
RegulonDB  regulondb.ccg.unam.mx

Protein motifs
Interpro  www.ebi.ac.uk/interpro
Pfam  www.sanger.ac.uk/Software/Pfam

Mobile elements
Isfinder  www-is.biotoul.fr/is.html
ACLAME  aclame.ulb.ac.be

Repeat elements
Repbase  www.girinst.org/repbase/index.html
CRISPRdb  crispr.u-psud.fr/crispr
III. Related resources

Bibliographic databases

- **Pubmed**: [www.pubmed.org](http://www.pubmed.org)
  - Comprises more than 19 million citations for biomedical literature from MEDLINE, life science journals, and online books.
  - Citations may include links to full-text content from PubMed Central and publisher web sites.

- **ISI Web of Science**: [wokinfo.com/products_tools/multidisciplinary/webofscience](http://wokinfo.com/products_tools/multidisciplinary/webofscience)
  - A research platform providing access to the world's leading citation databases
  - Web of Science information is carefully evaluated and selected.

- **Free access journals**: authors pay to get the papers free
  - The Biomed Central BMC initiative: [www.biomedcentral.com](http://www.biomedcentral.com)
  - The Public Library Of Sciences initiative: [www.plos.org](http://www.plos.org)
Conclusion

• **NGS (Next Generation Sequencing) technologies** provide more and more data at ever lower cost.

• **Diversification of projects**: de novo sequencing, re-sequencing, metagenomics, RNAseq,…

  *Example: the 1000 genomes project*  [www.1000genomes.org](http://www.1000genomes.org)

• We will have to deal with increasing amounts of sequencing data and **it is still a challenging task to provide adequate structures to produce, store and analyze data**
IV. Conclusion/Challenges

Genome resources challenges

Some biological challenges:

- **Genome annotation**: Merging automated, experimental and curated information: “reference genomes” vs “draft genomes”

- **Dealing with multiple genomes concerning**: individuals, strains and related species

- **Linking polymorphisms with phenotypes and functional studies**

Some bioinformatics challenges:

- **Data archiving**: format, volume and standardization problems

- **Data integration**: physical, virtual, semantical

- **Data visualization** of large volume of data in a visually intuitive format
IV. Conclusion/Challenges

Data integration

• **Biomart** [www.biomart.org](http://www.biomart.org)
BioMart is a query-oriented data management system developed jointly by the Ontario Institute for Cancer Research (OICR) and the European Bioinformatics Institute (EBI).
The system can be used with any type of data and is particularly suited for providing 'data mining' like searches of complex descriptive data.

• **Web services** [www.biocatalogue.org](http://www.biocatalogue.org)
Web services are application programming interfaces (API) or web APIs that are accessed via Hypertext Transfer Protocol and executed on a remote system hosting the requested services
Data visualization

Genome Browsers (GBs) hold a central place in genomic projects

- Some popular GBs: Gbrowse, Artemis, Ensembl, UCSC,…

- Many GBs are now available but the choice of a well adapted GB can be a challenging task

- Web site: genome.jouy.inra.fr/CompaGB
V. Practical exercises

Practice (1h30)

**Exercise 1:** Get the genome of two *Mycobacterium* species via the NCBI web site and manipulate gene identifiers (60’)

**Exercise 2:** Ensembl/Biomart usage via the EBI interface (30’)