Bioinformatics resources for data management

Etienne de Villiers
KEMRI-Wellcome Trust, Kilifi
Typical Bioinformatic Project

This is an iterative process. You may loop back at almost any step.
Knowledge Discovery in Databases (KDD)

Data

Cleaning Integration

Data Warehouse

Selection Transformation

Prepared data

Data Mining

Evaluation Visualization

Patterns

Knowledge Base

Knowledge
Data Acquisition

- Pose Hypothesis
  - Usually suggested by researcher
  - Important to state clearly and precisely
  - Motivates choice of data selection
  - E.g. Selective pressure is different in disease related genes.
Data Acquisition

- Identify relevant data
  - In collaboration with researchers
- What is needed?
  - Gene sequences, amino acid sequences, structure information, literature, etc.
- What quality is required?
  - No errors, some errors
Data Acquisition

- Identify Relevant Databases
  - Which databases contain the data you are interested in?
  - Which ones have the required quality?
  - Do you need general or specific data (or both)?
  - Look in January NAR, Google search, ask other researchers in the field
Data Acquisition

- Retrieve Data
  - What format is required?
  - Track where and when data is retrieved
  - Do you need to update the data regularly?
  - How do updates impact your analysis?
Data Acquisition

- Store data in local database
  - Data management is a fundamental piece of every project
  - Use a DBMS over flat files for projects
    - Flexible queries, remote access, etc.
  - Consider how data from multiple sources will be integrated
  - Examples of successes and failures surround you. Pay attention to the databases you use!
Other projects

- You may work where lots of data is generated:
  - Microarray experiments, sequencing, etc.
  - Don’t collect data without a goal in mind.

- Many steps of the process are the same.
  - Your data is just another database in the mix.

- Carefully consider how you want to store the information.
Databases

- A *database* is a collection of data, typically describing the activities of one or more related organizations. (Ramakrishnan and Gehrke).

- A *database* is a repository for a collection of computerized data files. (Date)
Databases

- Databases typically support the following operations
  - Retrieval
  - Insertion
  - Updating
  - Deletion
Database Models

- Defines data organization

- Relational
  - Entities and relationships stored in tables
  - Oracle, DB2, MySQL, PostgreSQL
  - Predefined schema

- Object Oriented/Object Relational
  - Abstract data types, data and operations
  - Structured types (arrays, lists, sequences, etc.)
  - Flexible
Bioinformatic Databases

- Different kinds of bioinformatic databases
  - General Purpose
  - Data type specific
  - Organism specific
  - Pathway information
  - Specialized data
  - ...
Where to look for databases?

- Search Engines
- Journals related to Bioinformatics
  - Nucleic Acids Research (NAR) online Database issue
- Websites like:
  - www.expasy.ch
There is a lot of databases

- Nucleic Acids Research (NAR) Database Collection, which currently lists 1512 online databases.

- Versioning is an issue.
  - Primary data is always being updated.
  - Data versioning is not always feasible.

- Maintain read-only "master copies" of primary data, version your code, then you can say "this input + this code version will reproducibly generate this output."
Major Databases

- There are many public resources but only a few key resources.
- Many databases are part of projects that has limited lifespan.
- Use major public resources: NCBI, EBI, Ensembl, PDB, KEGG.
- If your field is more specific, identify the major resources in that area (e.g. IMG for microbial genomics)
Retrieving Data

1. Formulate A Query
2. Select A Data Format
3. Download Result
4. Review Result
   - Is All Required Data Collected?
     - No: Return to Formulate A Query
     - Yes: Store Result
Formulate Query

- Databases offer different query options
  - Dependent on the underlying structure
    - E.g. RDBMS can offer SQL queries
  - Dependent on the user interface

- Most databases offer the following features:
  - Keyword
  - Boolean searches (AND/OR/NOT)
  - Restricted searches
    - Date, organism, annotation (e.g. MeSH), etc.
Formulate Query

Example:

- Retrieve all **reviewed nucleotide sequences** from RefSeq mentioning **BRCA1** (a breast cancer gene).
- Restrict the search to mice.

BRCA1 AND srcdb_refseq_reviewed[prop] AND mouse[organism]
Identifiers

- Most entries in bioinformatic databases have one or more *identifiers*
  - Intended to uniquely identify entries
  - Should be used in (your) publications

- Common identifiers for bioinformatic data
  - Locus Name
  - Accession Numbers
  - GenInfo ID
  - Pubmed ID
  - …
Locus IDs

- Original identifiers of GenBank records
  - **LOCUS** line in GenBank entries
- Used to be first 3 letters of organism followed by code for gene
  - **HUMBB** for human β-globin region
- Unmaintainable due to growth of data
  - Homologous genes not named the same
Accession Numbers

- No biological meaning
- Originally an uppercase letter followed by 5 digits: U00002
- Now, two uppercase letters followed by six digits: BC037153
  - ACCESION in GenBank entry
- Version of entry added later: BC037153.1
- Stable way of identifying GenBank entries
- Now being used for both DNA and proteins
Accession Numbers

- Some Accession numbers have special meaning
  - NM_, NP_, XM_, … are RefSeq

GenInfo (gi) IDs

- Identifier for a particular sequence only
  - Each entry gets a unique gi number
  - GI: 22477487

- Not subject to versioning
  - The entry remains the same over all time
  - Different versions of the same sequence are managed by accession numbers
Pubmed Ids (PMID)

- Identifies articles managed by NCBI
- Reliable, stable link to citation

**Example:** PMID: 12205585

- A related ID is the MedLine ID
  - PMIDs are preferred if both available
Other IDs

- Other databases have their own ID definitions
- Important features to look for
  - Uniquely identify a database entry
  - Version information
  - Associations with other databases
    - Is there a mapping of identifiers available?
- Read documentation on database for details
Formats

- Data is stored/presented in a variety of formats
  - FASTA, GenBank, SwissProt, ASN.1, XML

- When considering a format for retrieval
  - What is easy to parse?
  - What format do the tools need?
  - What information is needed?
FASTA (or Pearson)

- Used by FASTA tools
- Comment line followed by sequence data
  - No Annotation, just sequence
- Example:

```plaintext
>gi|1040960|gb|U35641.1|MMU35641 Mus musculus Brca1 mRNA, complete cds
GGCAGGGATCCAGCAGCTCTTGGGGCTCTCTGCTGACCTCGCCGCTTTGGAAGTACGGATCTTTTTCTCTC
GGAGAAAAGTTTCACTGGAACCTGGAAGAAATGGATTATCTGCGTCCAAATTCAAGAAGTACAAATGTGC
TTCTGCTATGCAGAAAATCTTGGACTGCTGCTTTTGGAACCTGCTGAAATGGTTTTAGTTTTTCG
AGGAGACACTATATTTGCAATTTTTGATGTGCAACCTTACCAAGAAGAAAGGCCCTTCAATATGTC
TCTTTTGTGTAAGAATGAGATAACCAAAAGGAGCCTACAGGGGAAGCAAGGATTATGCACTCTGCTG
AGAGCTGCTGAGAAATAATGGCTGCTTTTGAGCTTACCGGGAATGGCTGCTTTTGAGCTGTGAGACCT
GCTGTGACCTTCTGGAATCGTGAGATCAGTGAAGAAAAACAGGCAGACCCAACCTCGAAAGAAATCTG
ATCATTGAAACTAGACTCTTCTGTGAGAGAGACC
```
GenBank

- Flat file format used by GenBank
- Annotation, author, version, etc.
- Example (just the top)

<table>
<thead>
<tr>
<th>LOCUS</th>
<th>MMU35641</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEFINITION</td>
<td>Mus musculus Brca1 mRNA, complete cds.</td>
</tr>
<tr>
<td>ACCESSION</td>
<td>U35641</td>
</tr>
<tr>
<td>VERSION</td>
<td>U35641.1 GI:1040960</td>
</tr>
<tr>
<td>KEYWORDS</td>
<td>.</td>
</tr>
<tr>
<td>SOURCE</td>
<td>house mouse strain=C57B1/6.</td>
</tr>
<tr>
<td>ORGANISM</td>
<td>Mus musculus</td>
</tr>
<tr>
<td></td>
<td>Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Mus.</td>
</tr>
<tr>
<td>REFERENCE</td>
<td>1 (bases 1 to 5538)</td>
</tr>
<tr>
<td>AUTHORS</td>
<td>Sharan,S.K., Wims,M. and Bradley,A.</td>
</tr>
<tr>
<td>TITLE</td>
<td>Murine Brca1: sequence and significance for human missense mutations</td>
</tr>
<tr>
<td>MEDLINE</td>
<td>96177660</td>
</tr>
<tr>
<td>PUBMED</td>
<td>8634698</td>
</tr>
</tbody>
</table>
### SWISS-PROT

- Defined by SWISS-PROT database
- Includes annotation, other info

#### Example:

<table>
<thead>
<tr>
<th>ID</th>
<th>BRC1_MOUSE STANDARD; PRT; 1812 AA.</th>
</tr>
</thead>
<tbody>
<tr>
<td>AC</td>
<td>P48754; Q60957; Q60983;</td>
</tr>
<tr>
<td>DT</td>
<td>01-FEB-1996 (Rel. 33, Created)</td>
</tr>
<tr>
<td>DT</td>
<td>01-NOV-1997 (Rel. 35, Last sequence update)</td>
</tr>
<tr>
<td>DT</td>
<td>16-OCT-2001 (Rel. 40, Last annotation update)</td>
</tr>
<tr>
<td>DE</td>
<td>Breast cancer type 1 susceptibility protein homolog.</td>
</tr>
<tr>
<td>GN</td>
<td>BRCA1.</td>
</tr>
<tr>
<td>OS</td>
<td>Mus musculus (Mouse).</td>
</tr>
<tr>
<td>OC</td>
<td>Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;</td>
</tr>
<tr>
<td>OC</td>
<td>Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Mus.</td>
</tr>
<tr>
<td>OX</td>
<td>NCBI_TaxID=10090;</td>
</tr>
<tr>
<td>RN</td>
<td>[1]</td>
</tr>
<tr>
<td>RP</td>
<td>SEQUENCE FROM N.A.</td>
</tr>
<tr>
<td>RC</td>
<td>STRAIN=C57BL/6; TISSUE=Embryo;</td>
</tr>
<tr>
<td>RX</td>
<td>MEDLINE=96177659; PubMed=8634697;</td>
</tr>
<tr>
<td>RT</td>
<td>&quot;Mouse Brca1: localization sequence analysis and identification of evolutionarily conserved domains.&quot;;</td>
</tr>
</tbody>
</table>

...
Peptide Databases

- Protein identification from tandem mass spectra.
- Peptide *de novo* sequencing.
- PeptideAtlas database
- PRIDE database
- ProteinProspector
Organizing Bioinformatics projects

- Paper based lab notebook.
  - Difficult to maintain.

- Computer based lab notebook.

- Use a Wiki.
Computer based lab notebook

- Use your computer as your notebook.
- Create a sensible file/directory hierarchy.
- Create notes in text files in root directory
- Use “history” function of Unix.
Analysis 28 June 2013

Convert 454 sff file into fasta and qual files:

sff2fastq -o M1104_454Reads.RL16.fastq M1104_454Reads.RL16.sff

Assembly reads using cap3:

cap3 M1176.RL6.fasta M1176.RL6.qual

Identify dbla alpha tags

ruby dbla_finder.rb -i file -o out.file
ruby ../dbla_finder.rb -i T08_454Reads.RL16.fasta.cap.contigs > T08_454Reads.RL16.fasta.cap.contigs.tags

Translate sequence into amino acids

transeq T08_454Reads.RL16.fasta.cap.contigs.tags
T08_454Reads.RL16.fasta.cap.contigs.tags.aa

Run dbl alpha classifier on
ruby ../dbla_classifier.rb T08_454Reads.RL16.fasta.cap.contigs.tags.aa

Output file columns:

entry.definition	#{tag.cyspolv_group}	#{tag.dsid}	#{tag.bs_group}	#{tag.cys_count}	#{tag.length}	#{tag.var1_status}	#{tag.sig2_status}	#{tag.groupA_status}

Create network:

ruby ../net.rb T08_454Reads.RL16.fasta.cap.contigs.tags.aa
awk '{print $1, $3, $9}' T176_454Reads.RL18.fasta.cap.singlelets.tags.aa.tab | more
Wiki

- A wiki is a website that allows collaborative editing of its content and structure by its users.
- Dokuwiki is a popular wiki for personal computers.
EMBOSS

- A suite of >150 bioinformatics applications.
- Open source software – freely available.
- Runs on Unix, MacOSX and MS Windows
- They share a common look and feel.
- Incorporates many small and large programs.
- Easy to run from the command line or graphical interfaces.
EMBOSS programs

• General:
  • wossname - list EMBOSS programs
  • showdb - shows the available databases

• Sequence retrieval
  • seqret - retrieve and/or changes format of sequence
  • transeq - translate DNA sequence to protein
  • backtranseq - translate protein sequence to DNA
  • extractseq - extract region from a sequence
  • cutseq - remove regions from sequence
Working with sequences

- EMBOSS reads sequences from files or databases.
- It automatically recognizes the input sequence format.
- You can easily specify many output formats.
Sequence formats

- Sequences can be read and written in a variety of formats.
- Sequences are stored in databases or in files as simple text (ASCII text).
- Microsoft Word format is not a sequence format (save the files as text *.txt file!).
- The default sequence file format is **fasta**

```
>seq_name!
acgcacgctagcagcagctagctagcagcgcgatcgcctagcatctagctagctcg
atgcgatcgatcgcagcgcagcgcgatcgcctagcatcgcctagcatcgcctagc
gcagcgcgatcgcctagcatcgcctagcatcgcctagcatcgcctagcatc
gcagcgcgatcgcctagcatcgcctagcatcgcctagcatcgcctagcat
```
wEMBOSS

- Web based interface to EMBOSS.
- Each user has a separate and private workspace.
- Organize your work by creating projects and subprojects.
- Results are save for easy recovery and review.
Project Management

Organize your work by creating projects

EMBOSS applications

EMBOSS applications are grouped by type.

An alphabetic list of the programs is also available.

This list can be searched by keywords.

Reminder: wossname program to find a given EMBOSS application
JEMBOSS

- Java based interface to EMBOSS.
Conclusion

- There are many challenges in managing bioinformatics data.
  - Organise your data.
  - Use appropriate databases.
  - Use appropriate tools.
JEMBOSS workshop