Information Systems in the Life Sciences (ISiLS)

Fons J. Verbeek
Imaging & BioInformatics, LIACS

Lecture 3a

Contents

- Information systems
- Basic Formats
- Annotation
- Integration

Information Systems

- In the life sciences the information is dispersed over a number of databases.
- Information retrieved from these databases and combined with other data
- The life-sciences information system is a set of databases.

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)

5

Operations on Databases

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

6

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.)
  - Inheritance of attributes
- Hierarchical/Semistructured
  - Implicit schema
  - Flexible

7

Integration of Information

- Information thus starts from
  - Doing experiments
  - Submit data to a repository
  - Extract related data from other repository
  - Combine the outcome
- How are the data integrated/exchanged
  - Format of our data
  - Format of our database

8

Formats

- Data is stored/presented in several formats
  - ASN.1
  - FASTA
  - GenBank
  - SwissProt
  - XML
- Native formats of the database with
  - PHP interface
  - DHTML interface

9

What do these formats look like …

10

What is ASN.1

- ASN.1 = Abstract Syntax Notation 1
- International standard language for data specification
  - Used to build complex data types in a hierarchical manner
  - Originated with Xerox
- Used in telephone systems, air traffic, building and machine control, toll highways, smart cards, security and more
- Used by NCBI to store GenBank, PubMed, MMDB and more

11

ASN.1

- International standard
  - Semistructured format
  - Base format NCBI data: heterogeneous data!
- Example:

```
Seq-entry ::= set {
  level 1 ,
  class nuc-prot ,
  descr {
    title "Mus musculus Brca1 mRNA, and translated products" ,
    source {
      org {
        taxname "Mus musculus" ,
        db {
          {
            db "taxon" ,
            tag
              id 10090 } } ,
        orgname {
          name
            binomial {
              genus "Mus" ,
              species "musculus" } , …
```

12

FASTA (or Pearson)

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

```
>gi|1040960|gb|U35641.1|MMU35641 Mus musculus Brca1 mRNA, complete cds
GGCACGAGGATCCAGCACCTCTCTTGGGGCTTCTCCGTCCTCGGCGCTTGGAAGTACGGATCTTTTTTCT
CGGAGAAAAGTTCACTGGAACTGGAAGAAATGGATTTATCTGCCGTCCAAATTCAAGAAGTACAAAATGT
CCTTCATGCTATGCAGAAAATCTTAGAGTGTCCGATCTGTTTGGAACTGATCAAAGAACCTGTTTCCACA
AAGTGTGACCACATATTTTGCAAATTTTGTATGCTGAAACTTCTTAACCAGAAGAAAGGGCCTTCACAAT
GTCCTTTGTGTAAGAATGAGATAACCAAAAGGAGCCTACAGGGAAGCACAAGGTTTAGTCAGCTTGCTGA
AGAGCTGCTGAGAATAATGGCTGCTTTTGAGCTTGACACGGGAATGCAGCTTACAAATGGTTTTAGTTTT
TCAAAAAAGAGAAATAATTCTTGTGAGCGTTTGAATGAGGAGGCGTCGATCATCCAGAGCGTGGGCTACC
GGAACCGTGTCAGAAGGCTTCCCCAGGTCGAACCTGGAAATGCCACCTTGAAGGACAGCCTAGGTGTCCA
GCTGTCTAACCTTGGAATCGTGAGATCAGTGAAGAAAAACAGGCAGACCCAACCTCGAAAGAAATCTGTC
TACATTGAACTAGACTCTGATTCTTCTGAAGAGACAGTAACTAAGCCAGGTGATTGCAGTGTGAGAGACC
…
```


GenBank

- Flat file format used by GenBank
  - Annotation, author, version, etc.
- Example (*head -14*)

```
LOCUS       MMU35641                5538 bp    mRNA    linear   ROD 18-OCT-1996
DEFINITION  Mus musculus Brca1 mRNA, complete cds.
ACCESSION   U35641
VERSION     U35641.1  GI:1040960
KEYWORDS    .
SOURCE      house mouse strain=C57Bl/6.
  ORGANISM  Mus musculus
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Mus.
REFERENCE   1  (bases 1 to 5538)
  AUTHORS   Sharan,S.K., Wims,M. and Bradley,A.
  TITLE     Murine Brca1: sequence and significance for human missense
            mutations
  JOURNAL   Hum. Mol. Genet. 4 (12), 2275-2278 (1995)
  MEDLINE   96177660
   PUBMED   8634698
```


SWISS-PROT

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

```
ID   BRC1_MOUSE     STANDARD;      PRT;  1812 AA.
AC   P48754; Q60957; Q60983;
DT   01-FEB-1996 (Rel. 33, Created)
DT   01-NOV-1997 (Rel. 35, Last sequence update)
DT   16-OCT-2001 (Rel. 40, Last annotation update)
DE   Breast cancer type 1 susceptibility protein homolog.
GN   BRCA1.
OS   Mus musculus (Mouse).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Mus.
OX   NCBI_TaxID=10090;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   STRAIN=C57BL/6; TISSUE=Embryo;
RX   MEDLINE=96177659; PubMed=8634697;
RA   Abel K.J., Xy J., Yin G.Y., Lyons R.H., Meisler M.H., Weber B.L.;
RT   "Mouse Brca1: localization sequence analysis and identification of
RT   evolutionarily conserved domains.";
RL   Hum. Mol. Genet. 4:2265-2273(1995).
…
```


basics of XML

- eXtensible Markup Language
  - Tags like HTML
  - International Standard
  - Semi Structured
  - Subset of SGML
- Many dedicated XML schema's
  - MAGE-ML (micro array)
  - CML, chemical markup language
  - Bio XML
  - …

Example XML

```
<?xml version="1.0"?>
<!DOCTYPE GBSeq PUBLIC "-//NCBI//NCBI GBSeq/EN"

http://www.ncbi.nlm.nih.gov/dtd/NCBI_GBSeq.dtd">

<GBSet>
<GBSeq>
  <GBSeq_locus>MMU35641</GBSeq_locus>
  <GBSeq_length>5538</GBSeq_length>
  <GBSeq_strandedness value="not-set">0</GBSeq_strandedness>
  <GBSeq_moltype value="mrna">5</GBSeq_moltype>
  <GBSeq_topology value="linear">1</GBSeq_topology>
  <GBSeq_division>ROD</GBSeq_division>
  <GBSeq_update-date>18-OCT-1996</GBSeq_update-date>
  <GBSeq_create-date>25-OCT-1995</GBSeq_create-date>
  <GBSeq_definition>Mus musculus Brca1 mRNA, complete cds</GBSeq_definition>
  <GBSeq_primary-accession>U35641</GBSeq_primary-accession>
  <GBSeq_accession-version>U35641.1</GBSeq_accession-version>
```

17

XML

- Usually defined in document type definition (DTD) or in an XML schema
  - Defines valid tags, valid value types
  - Used for format validation
  - Used for data validation
- XSLT
  - Defines how to translate documents to other formats
- XML used heavily in business, becoming more popular in science

18

Transforming formats

- Tools handle only one of the data formats
  - Database specific
- Software transform between formats
  - Database filters
  - Software toolkits
  - Examples:
    - ReadSeq
    - http://searchlauncher.bcm.tmc.edu/seq-util/Options/readseq.html
    - SEQIO
    - http://www.cs.ucdavis.edu/~gusfield/seqio.html

19

Integrating Data from Databases

- CORBA: common object request brooker
- BioMoby – Interoperability of biological data
  - Provides WSDL service descriptions to client through a MOBY central repository, which in turn tracks available data hosts and services
- BioDAS – Distributed Annotation Server
  - uses DAS/1 protocol to talk to various annotation servers and provide a single view to client
  - e.g. WormBase, FlyBase, Ensembl, TIGR
- XML
- Ontologies: OBO, GO

20

The data in the databases were annotated …

21

Different Levels of Annotation

- Sparse – typical in
  - usually just includes name and accession number
  - gel annotations, sequence annotations
  - microarray annotations,
- Moderate – typical in
  - many sequence databases or
  - of experiments aimed at identifying protein complexes or ligands
- Detailed – not typical?, but effort is directed
  - found in organism-specific databases

22

Integrating information: KEGG

- 5904 chemical reactions
- **15,037 pathways**
- 229 reference pathways
- 85 ortholog tables
- 181 organisms
- http://www.genome.ad.jp/kegg/

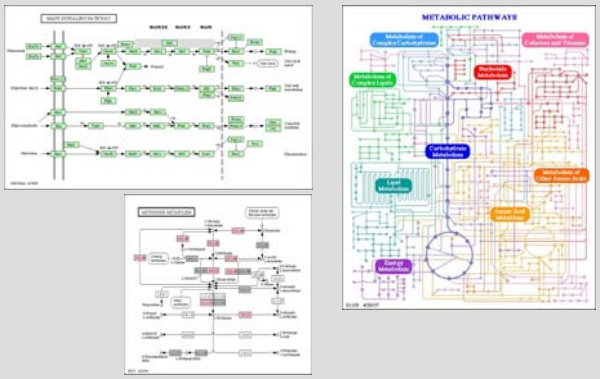
23

KEGG

- GENES Database
  - The universe of genes and proteins in complete genomes
- LIGAND Database
  - The universe of chemical reactions involving metabolites and other biochemical compounds
- Pathway Database
  - Molecular interaction networks, metabolic and regulatory pathways, and molecular complexes

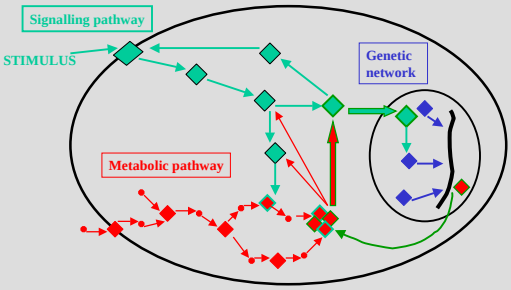
24

Pathways: regulatory - metabolic



25

Pathways are inter-linked



26

Query KEGG

PAX6: **DBGET Search Result**

Database: GENES

KEGG Genes Database Release 31.0+/09-30, Sep 04 Institute for Chemical Research, Kyoto University 772,389 entries, 1,068,482,587 residues **Keyword(s): PAX6** (Total 6 hits. / limit=1000)

1. hsa:5080 PAX6, AN2; paired box gene 6 (aniridia, keratitis)
2. mmu:18508 Pax6; paired box gene 6
3. rno:25509 Pax6; paired box gene 6
4. rno:311270 LOC311270; similar to elongation protein 4 homolog; PAX6 neighbor gene; chromosome 11 open reading frame 19
5. dre:30567 pax6a; paired box gene 6a
6. dre:60639 pax6b; paired box gene 6b

27

Result Query KEGG

28

Result KEGG linked

BLAST Search Result: dre:60639 ->
Database: nr-aa

Non-redundant protein sequence da
Release 04-09-27, Sep 04
1,711,502 entries, 554,776,211 re
Last update: 04/09/30
Protein sequence database entries

Managing the data deluge …

30

Data compression

- Feb 2004 release of GenBank has:
  - 37,893,844,733 base pairs
  - 32,549,400 sequences
  - 143Gb data in all
- Doubles every 1-2 years!
- Mostly A, C, G, T ()
  Compression : 4 bit encoding, provides big saving
- Simple gzip reduces 143Gb to about 21Gb
- Disks are cheap.. 250Gb for €300

31

Data compression

- High-Resolution images take up space too
  - CLSM images
  - Bright Field and Fluorescence Microscopy Images
  - Micro Array plates
  - Electron micrographs
  - Atomic Force Microscopy (AFM) images
- Compression helps greatly –
  for images:
  - Lossless (GIF, RLE, PNG, TIFF, some JPEG)
  - Lossy (JPEG)

32

Data security

- Encryption, digital signatures
- Very useful when transferring data from server to client machine, and vice-versa
- Help maintain data integrity
- Prevent ‘eavesdroppers’ from seeing confidential data: company related
- Especially important in a corporate setting Pharma-companies, Seed-companies etc…

33