Integrating and Accessing Molecular Biology Resources

Alexandre Karlov

Introduction

• Applications operating on the collected data produce more data
• Data should be integrated so that relationships between data can be found

Outline

• Introduction
• Molecular Biology Resources
  – Databases
  – Applications
• SRS
• Data Integration & Access
• Conclusion

Problems

• Obtaining a consistent, cross-referenced view of the data
• Heterogeneity of sources
• Quality of data
• Maintaining and updating data
Databases

- Sequence databases
- Sequence-related databases
- Genome databases
- Mutations and SNPs
- Pathway and chemical compounds
- Transcription factors
- Literature databases

Sequence-related databases

- Contain further information about the sequences in the sequence database
- Provide an explicit link to a sequence database

Genome Databases

- GDB (Genome Databank)
- HGMP (Human Genome Mapping Project)

Sequence databases

- EMBL (Europe), GENBANK (USA), DDBJ (Japan)
  - DNA databases of sequences and their annotations
  - Sharing their data
  - Annotated feature tables
- Protein databases
  - SWISS-PROT

A.Karlov, “Integrating and Accessing Molecular Biology Resources” 12/01/2005
Mutations and SNPs

- At DNA level
- 3 types (exchange, insertion, deletion)
- Give a different protein product
- OMIM (Online Mendelian Inheritance in Man) - catalog of human genes and genetic disorders
- SNP (Single Nucleotide Polymorphisms)
- MITSNP - SNP database
- SWISS-PROT and EMBL include variation information

A. Karlov, “Integrating and Accessing Molecular Biology Resources” 12/01/2005

Transcription factors

- Protein that binds DNA at a specific promoter or enhancer region where it regulates transcription
- Transfac - database on eukaryotic transcription factors, their genomic binding sites and DNA-binding profiles
- TFD

A. Karlov, “Integrating and Accessing Molecular Biology Resources” 12/01/2005

Pathway and Chemical compounds

- Enzyme databases (Enzyme, Brenda, Ligand, EMP)
- Contain enzymes involved in a large number of reactions
- Reaction databases (MPW, KEGG, EcoCyc)
- Pathways of a reaction, input, products and enzyme information

A. Karlov, “Integrating and Accessing Molecular Biology Resources” 12/01/2005

Literature

- Some important data is in literature
- Medline (by National Library of Medicine)
- Excerpta Medica (by Elsevier)
- Contain abstracts of published papers

A. Karlov, “Integrating and Accessing Molecular Biology Resources” 12/01/2005
Applications

- Similarity search engines
  - BLAST and FASTA
- Sequence matching
  - CLUSTALW - sequence alignment tool
- Results form themselves a database

SRS - an overview

- Leaves data in their original content
- The use of meta-description to describe the data
- Utilizes the special purpose programming language
- ICARUS (Interpreter of Commands and RecUrsive Syntax)
- Used for the meta definition of resources and the syntactic description of the data

<table>
<thead>
<tr>
<th>Clients</th>
<th>WWW Server</th>
<th>Automated Analysis</th>
<th>Data Viewers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wrappers</td>
<td>C++</td>
<td>CORBA</td>
<td>Perl</td>
</tr>
<tr>
<td>Icarus Meta-Level</td>
<td>Databanks</td>
<td>Applications</td>
<td>Object Loaders</td>
</tr>
<tr>
<td>Resources</td>
<td>DBMS</td>
<td>External</td>
<td>Flat File</td>
</tr>
</tbody>
</table>

SRS - an overview

- Sequence Retrieval System
- Creates indices of the data using the meta definitions of the format and syntax of the resources
- Querying is very fast
- The whole entry or just the information required by the user can be extracted
- Links between the data and other databases can be made, which allows navigation
**Meta-Definition layer**

- Used to describe both the resources to be integrated and the format of the data
- Extracted data can be viewed or passed to another program
- Changes to the input/output format requires minor changes to the meta definition and none to any core component of SRS

---

**SRS core**

- Several programs using the meta definition layer to allow the resources
  - to be queried
  - Data from these resources to be returned
  - Other programs to run on selected data
- Key concept: creation of indices from the data resources
- Makes the querying of the database very quick

---

**Wrappers**

- Provide a programming interface to SRS
- Full access to the SRS system
- Link programs to the SRS system
- C++, Perl, Java, Python
- Use CORBA to connect to a remote SRS system

---

**Clients**

- Top layer of the SRS system
- Web clients
- Also data viewers and automated analysis programs
Integrating molecular biology resources

- Importance of data structure for pertinent information retrieval
- Syntax of the field must be known
- SRS token server
  - Set of rules must be provided in order to describe the data

SRS token server

- Tokens are then written to token tables
- The data in those tables are then indexed and used during the query process
- Flat file DB contains a list of entries with each entry containing a list of data-fields
- ID, Accession number, key words, sequence, etc...
- Each of data-fields consists of string or tokens
- Tokens are inserted into index

Example of a typical DB entry (EMBL)

- ID: AB000000; standard; genomic DNA; HUM; 2582 BP.
- AC: AB000000;
- DE: Homo sapiens PGIC gene; complete cds.
- KW: glycosylphosphatidylinositol-synthesis gene; PGIC.
- SO: Sequence 2582 BP: 694 A; 434 C; 581 G; 813 T; 0 other;
- ggtctcgc tcgagaaggq glnacggtf cngctgccc aagcaaatlt tgttgtggc

Subset of rules describing an entry in the EMBL DB

entry: ~ id_line ac_line da_line kw_line sq_line end_line ~
id_line: ~ ID id ~
ac_line: ~ AC ac ~
kw_line: ~ KW (word) ~
sq_line: ~ SQ sequence ~
end_line: ~ / ~

id: ~ word $Wt ~
kw: ~ word $Wt ~

...word: ~ [.a-zA-Z0-9-] ~
Meta definition of molecular biology resources

- Structure description
- Written also using ICARUS
- Example - EMBL:

```plaintext
EMBL_DBLibrary[EMBL_group:@SEQUENCE_LIBS format:@EMBL_FORMAT searchName:*
] EMBL_FORMAT:LibFormat[LibType:@DAT_FILE @SEQ_FILE] syntax:@EMBL_SYNTAX contains:@DNASEQ_DATA
fields:
  $Field[@DF_ID index id indexToken:id]
  $Field[@DF_Accession index str indexToken:acc]
  $Field[@DF_Keywords index str indexToken:des]
  $Field[@DF_Description index str indexToken:des]
  $Field[@DF_DNASequence format:embl]
```

A.Karlov, “Integrating and Accessing Molecular Biology Resources” 12/01/2005

Querying and linking DB

- SRS provides a full query language
- Query strings and numeric fields
- Combining searches across multiple libraries and fields
- Returns a set as the result
- Sets can be combines using AND (‘&’), OR(‘|’) and BUT NOT (‘!’) operators
- Link operators are also provided

A.Karlov, “Integrating and Accessing Molecular Biology Resources” 12/01/2005

Views and object loader

- Views provide a way of viewing part of the information from a databank
- Related information from different databanks can also be include (virtual entry)
- Object loaders provide a flexible way of obtaining data
- Those objects can be passed to other programs (C, C++, Java, Perl or Python)
- Additional operations can be performed on objects

A.Karlov, “Integrating and Accessing Molecular Biology Resources” 12/01/2005

Analyzing data

- Using different tools on experemental data leads to more data
- Some DB are entirely based on those results
- Close relations between DBs

A.Karlov, “Integrating and Accessing Molecular Biology Resources” 12/01/2005
Consistency

- Keeping all relevant DBs up to date is a difficult task
- SRS uses Prisma, which automatically downloads, reformats, re-indexes and reports the progress of each step
- Configuration of SRS Prisma is done through an easy to use graphical interface
- Has proved to be very useful at keeping databanks up to date

Accessing integrated data

- Must be easy and powerful enough
- It must be possible to access the data from anywhere (different programming environments, languages or across a network)
- The Web interface is simple and powerful way to query and access the data

Accessing integrated data

- APIs provide programmers a powerful and flexible way to access SRS system
- Example (Perl loader for LiveSeq from EMBL entries with SRS)

```perl
my $db="EMBL";
my $acc_id="M20132";
my $query="embl-acc:$acc_id";
my $loader=Loader::LiveSeq::IO::SRS->load(db=>"EMBL", -query=>"$query");
my @translationobjects=$loader->entry2liveseq();
my $gene="AR";
my $gene=$loader->gene2liveseq("gene")
```

Accessing integrated data

- Other ways of accessing SRS system
- ICARUS
  - Allows parsing and easier definition of object loaders
- Using unix command line and the ‘getz’ program
Conclusions

• Available Molecular Biology resources
• How the data may be integrated to allow users an easy and powerful access in different databases
• SRS allows browsing and querying existing DBs and integration of the new ones

References

• Chapter “Integrating and Accessing Molecular Biology Resources” from “Bioinformatics - From Genomes to Drugs”, Prof. Dr. Thomas Lengauer
• Google

Thank you!

Questions?