Bioinformatics data & parsing

Bioinformatics data analysis and tools -- data & parsing

Bioinformatics analysis

Data

Answer to some biological question

Computational tools

Every bioinformatics analysis starts with data!

==> Welcome to your first problem...

Bioinformatics data analysis and tools -- data & parsing

Aspects of bioinformatics data

Generation & Storage

Exchange & Integration

Representation & Formats

Essential bioinformatics skill: parsing

Bioinformatics data analysis and tools -- data & parsing
Aspects of bioinformatics data

- Generation & Storage
- Exchange & Integration
- Representation & Formats

Essential bioinformatics skill: parsing

Bioinformatics data analysis and tools -- data & parsing

Data sources

- Direct experimental data from the lab
- Databases
- Applications
- Queries

For example: MUSCLE, ClustalW, PRALINE

---

Data sources

- Direct experimental data from the lab
- Databases
- Applications
- Queries

X-ray diffraction pattern of protein crystal

Bioinformatics data analysis and tools -- data & parsing

---

Data sources

- Direct experimental data from the lab
- Databases
- Applications
- Queries

Multiple sequence alignment
For example: MUSCLE, ClustalW, PRALINE

Bioinformatics data analysis and tools -- data & parsing

---

Data sources

- Direct experimental data from the lab
- Databases
- Applications
- Queries

Atomic model of protein structure
For example: PDB

Bioinformatics data analysis and tools -- data & parsing
### Data sources

- Direct experimental data from the lab
- Databases
- Applications
- Queries

### Databases

<table>
<thead>
<tr>
<th>Types of databases</th>
<th>Data(base) issues</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flat file</td>
<td>Consistency</td>
</tr>
<tr>
<td>Relational</td>
<td>Redundancy</td>
</tr>
<tr>
<td>Object-oriented</td>
<td>Normalization</td>
</tr>
<tr>
<td>Hypertext</td>
<td>Indexing</td>
</tr>
<tr>
<td></td>
<td>Machine readable</td>
</tr>
<tr>
<td></td>
<td>Data structure</td>
</tr>
</tbody>
</table>

### PDB record: list of residues

<table>
<thead>
<tr>
<th>SEQRES</th>
<th>1 0 2922</th>
<th>U</th>
<th>U</th>
<th>G</th>
<th>G</th>
<th>C</th>
<th>U</th>
<th>A</th>
<th>C</th>
<th>U</th>
<th>A</th>
<th>U</th>
<th>G</th>
<th>C</th>
<th>PDB record: list of residues</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATOM</td>
<td>1 O5+</td>
<td>U</td>
<td>0</td>
<td>10</td>
<td>16.030</td>
<td>148.355</td>
<td>104.290</td>
<td>1.00</td>
<td>83.79</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2 C5+</td>
<td>U</td>
<td>0</td>
<td>10</td>
<td>16.942</td>
<td>149.450</td>
<td>104.201</td>
<td>1.00</td>
<td>82.36</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3 C4+</td>
<td>U</td>
<td>0</td>
<td>10</td>
<td>18.405</td>
<td>149.086</td>
<td>104.371</td>
<td>1.00</td>
<td>81.58</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### PDB record: list of atoms

<table>
<thead>
<tr>
<th>SEQRES</th>
<th>1 0 2922</th>
<th>U</th>
<th>U</th>
<th>G</th>
<th>G</th>
<th>C</th>
<th>U</th>
<th>A</th>
<th>C</th>
<th>U</th>
<th>A</th>
<th>U</th>
<th>G</th>
<th>C</th>
<th>PDB record: list of atoms</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATOM</td>
<td>1 O5+</td>
<td>U</td>
<td>0</td>
<td>10</td>
<td>20.413</td>
<td>148.134</td>
<td>101.205</td>
<td>1.00</td>
<td>65.72</td>
<td>P</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2 C5+</td>
<td>U</td>
<td>0</td>
<td>10</td>
<td>20.046</td>
<td>148.458</td>
<td>99.805</td>
<td>1.00</td>
<td>67.61</td>
<td>O</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3 C4+</td>
<td>U</td>
<td>0</td>
<td>10</td>
<td>20.863</td>
<td>146.760</td>
<td>101.535</td>
<td>1.00</td>
<td>65.62</td>
<td>O</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Internal consistency example

#### PDB record: missing residues

<table>
<thead>
<tr>
<th>SEQRES</th>
<th>1 0 2922</th>
<th>U</th>
<th>U</th>
<th>G</th>
<th>G</th>
<th>C</th>
<th>U</th>
<th>A</th>
<th>C</th>
<th>U</th>
<th>A</th>
<th>U</th>
<th>G</th>
<th>C</th>
<th>PDB record: missing residues</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATOM</td>
<td>1 O5+</td>
<td>U</td>
<td>0</td>
<td>10</td>
<td>26.018</td>
<td>149.543</td>
<td>99.231</td>
<td>1.00</td>
<td>47.64</td>
<td>P</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2 C5+</td>
<td>U</td>
<td>0</td>
<td>10</td>
<td>26.451</td>
<td>150.604</td>
<td>98.281</td>
<td>1.00</td>
<td>47.81</td>
<td>O</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3 C4+</td>
<td>U</td>
<td>0</td>
<td>10</td>
<td>24.872</td>
<td>148.661</td>
<td>98.281</td>
<td>1.00</td>
<td>46.23</td>
<td>O</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Bioinformatics databases

- NCBI
- UniProt
- Pfam
- swissprot
- PDB
- SCOP
- WormBase

Bioinformatics data analysis and tools -- data & parsing

When you are designing a database or application:
- What data format to use?
- Is the data structured in a logical way?
- Is the data machine readable?
- What type of database best suits my needs?
- Design a good database: worry about data issues!

When you are on the receiving end:
- Where does my data come from?
- What format is used? Is there a formal description?
- Is the data machine readable? What's the structure?
- What obstacles does my parser have to deal with?

Further reading

Bioinformatics databases:

Relational versus Object-oriented:

Tree of life project:
http://tolweb.org/tree/phylogeny.html

Essential bioinformatics skill: parsing

Aspects of bioinformatics data
- Generation & Storage
- Exchange & Integration
- Representation & Formats
Integration of data from multiple sources (databases and/or applications) if often necessary

Sequence
3D structure of gene product

Gene expression
Protein interactions

Two approaches to large-scale data integration
- Data warehousing
- Data federation

Data warehousing

Examples:
- ATLAS
- BioWarehouse
- LCB-DWH
- MolPAGE
- SABIO-RK

Data federation
Data federation

Examples:
- K2
- BioKris
- OPM
- TAMBI
- BioMediator
- SPINE2
- caBIG
- Rodeo

Bioinformatics data analysis and tools -- data & parsing

Ontology: shared ‘vocabulary’ within expert domain
- Open Biomedical Ontologies
- the Gene Ontology
- Sequence Ontology
- RNA Ontology

Exchange/markup language: standardized description
- XML -- general purpose
- BSML -- genomic sequences and biological function
- MAGE-ML -- microarray data
- RNAML -- RNA structure and motifs

Bioinformatics data analysis and tools -- data & parsing

Other standardization issues

• Naming
  – Example: same protein with different name in different databases
• Accession numbers and identifiers
  – Basically version control of data. How to keep things synchronized?
• What data should be recorded? Where?
  – Example: microarray data -- MIAME
  – Example: environmental samples? What should be measured? In what units?
  – Example: R-factor in 3D structures (PDB)

Bioinformatics data analysis and tools -- data & parsing

Analytical part to data integration/exchange:
- Where do I find all the data that I need?
- What are the relationships between these sources?
- What’s the meaning of concepts in these sources?
- How do I compare my results to previous work?

Technical part to data integration/exchange:
- How do I get all the data into an integrated data structure? Parsers!
- Are there common structures in the different sources?
- Generalize parsers?
- What data structure and exchange language should I use? What are the standards in this field?
Further reading

ATLAS- a data warehouse for integrative bioinformatics
Shah et al. BMC Bioinformatics. 2005 Feb 21;6:34

BioWarehouse: a bioinformatics database warehouse toolkit

Federated databases in bioinformatics (2 articles):
http://www.b-eye-network.com/view/2164

MIAME
http://www.mged.org/Workgroups/MIAME/miame.html

Aspects of bioinformatics data

Generation & Storage
Exchange & Integration
Representation & Formats

Essential bioinformatics skill: parsing

Bioinformatics data analysis and tools -- data & parsing

RNA secondary structure can be denoted by different formats

Connect format

```
73 ENERGY = -17.50  S.cerevisiae_tRNA-PHE
1 G 0 2 72 1
2 C 1 3 71 2
3 G 2 4 70 3
4 G 3 5 69 4
5 A 4 6 68 5
6 U 5 7 67 6
7 U 6 8 66 7
8 U 7 9 65 8
```

Bioinformatics data analysis and tools -- data & parsing
Dot-bracket or Vienna format

> S.cerevisiae tRNA-PHE M10740/1-73
GGCGAUUAGCUCAGUUGGAGGACGCACUGAAGAUUUUGAGGUCUGGUUCCGUCCAGAAGAAUUCGCA
(((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((\n
Stockholm or WUSS format

# STOCKHOLM 1.0
# GF AU Infernal 0.55
DF6280 GCGAUAUUGCUCAGUUGGAGGACGCACUGAAGAUUUUGAGGUCUGGUUCCGUCCAGAAGAAUUCGCA
# GC SS_cons ((((((,"<<._.>>>,"<<._.>>>,",",<<.
# GC RF gccgauuaUGcgcAcgU.GGUGGcgcgcacccUgaacgggAggUccg

DF6280 GUUCUCCAGUCCAGAAGAAUUCGCA
# GC SS_cons <<________>>))))):
# GC RF gggUUGCAGUcccccguacggcg

Bioinformatics data analysis and tools -- data & parsing

StructureString

(...(.-)\(.-)\(.-)\)

Pairs

[(0.22),(4.9),(5.8),(12.19),(13.17),(14.16)]

Partners

[22,None,None,None,9.8,None,None,5.4, None,None,19,17,16,None,14,13,None,12, None,None,0]

Data formats need to be machine readable

You need parsers to extract data from various formats and get it into a useful data structure (objects)

You need parsers to switch between different data structures

You need to be able to produce different output formats/visualizations from various data structures

Don’t try to find/design a single data format or structure that can fulfill all goals. It doesn’t exist!
Rather invest time in flexible data management: parsing.
Aspects of bioinformatics data

- Generation & Storage
- Exchange & Integration
- Representation & Formats

Essential bioinformatics skill: parsing

Bioinformatics data analysis and tools -- data & parsing

Parsing

Digesting the data format into logical components or structured units

- Input file Datastream Lines
- Data structures Objects

Parser

Case study: RNA families

Suppose we found some feature in ribosomal RNA

We want to know:
Is this feature universal for all families of RNA?

Approach:
Examine the same feature in all families in the RFAM (RNA families) database

First step: get the data out of the database

Rfam data format

- Annotation
  - Per-file (GF), per-column (GC), per-sequence (GS), per-sequence and per-column (GR)
  - Both compulsory and non-compulsory fields
  - Multiple entries for a field possible
- Sequences (in alignment)
- Structure
  - Actually part of annotation (GC SS_cons)
**Annotation example**

Bioinformatics data analysis and tools -- data & parsing

---

**Sequences example**

Bioinformatics data analysis and tools -- data & parsing

---

**Structure example**

Bioinformatics data analysis and tools -- data & parsing

---

**Parsing in general**

- Split data into records
  - Might use general record-finder functions
- Split record into separate elements
  - E.g. header, sequences
- Process every element
  - E.g. create Info object from label and remove bad characters from sequences
- Join all processed elements into an object
  - E.g. Sequence object

---

PLATFORMS

- Bioinformatics tools
  - CLC Genomics Workbench
  - Geneious
- Tools for specific tasks
  - FASTA
  - Sequence alignment
- Bioinformatics visualization
  - Circos
  - D3.js
- Bioinformatics databases
  - GenBank
  - NCBI
- Bioinformatics software
  - BLAST
  - Sequence analysis
- Bioinformatics web services
  - Ensembl
  - Genes
- Bioinformatics applications
  - DNA sequencing
  - RNA editing
- Bioinformatics algorithms
  - DNA alignment
  - RNA structure prediction
- Bioinformatics databases
  - PDB
  - Protein sequences
- Bioinformatics software
  - Cytoscape
  - IPA
- Bioinformatics applications
  - Gene expression
  - Protein function
- Bioinformatics algorithms
  - Clustering
  - Classification
- Bioinformatics databases
  - HMDB
  - Metabolites
### Rfam parser

- **Data**
  - RfamFinder
  - MinimalRfamParser
  - RfamSequence

- **Record**
  - **Family**
    - headerToInfo
  - **Info**
  - **Sequence, Info**
  - **Structure**

- **Record**

### Flexibility

- **Clustal W output format produced by the Clustal W program**

```
CLUSTAL W (1.83) multiple sequence alignment

#seq1 --MGAAI0QGQRGAP
#seq2 BCMAAI0QG--GAP
#seq3 --CVAEQQQRGG--GAPT
```

- **Clustal W output format produced by MUSCLE**

```
MUSCLE (3.6) multiple sequence alignment

#seq3 --CVAEQQQRGG--GAPT
#seq1 --MGAAI0QG--RGAP
#seq2 BCMAAI0QG----GAP
```

### Error handling

- **What errors/exceptions should be handled?**
- **Possible mechanism: two modes**
  - **Strict=True**: raise exception when error occurs. It’s up to downstream code to make a decision
  - **Strict=False**: return all ‘parsable’ records, just skip records that raise exceptions
- **It’s not always the fault of the parser, there are errors in databases (more than expected!)**
What about regular expressions?

- Pattern matching standard for string parsing and replacement
- Regular expressions are good for
  - Pattern matching (e.g. finding sequence motifs)
  - Local parsing of simple units (dates, accession numbers, etc)
- Regular expressions have disadvantages
  - Limited (e.g. find the same number of As and Bs)
  - Not flexible
  - Not readable (other people might want to use your code too!)

When designing a parser:

- What is the structure of the data that I need to parse?
- What is the best design for my parser?
- Hierarchical? Regular expressions?
- What exceptions are acceptable?
- Will other people use this code? Documentation?
- If something changes in the data format, can the parser easily be adjusted? Is my code modular?

Bioinformaticians need to be good at handling data. Writing parsers is an essential skill.

Assignment

- Write a small parser, perform some calculation, report some output
- Three or four cases. Pick one.
- Source code
  - Modular
  - Documented
  - Instructions to run the program
- Time estimate: ~ few hours
- Assignment will be online soon...