Databases for Bioinformatics

Brief overview of databases
relational databases
MySQL
Perl and PHP connections to MySQL

References
Bioinformatics Computer Skills, Gibas [ch 13]
MySQL, Dubois (New Riders)
Programming PHP, Lerdof (O'Reilly)

Mathematically, a relation is a set

Set operations used to define relations include
- union
  \[ x \cup A \cup B \text{ if } x \subseteq A \text{ or } x \subseteq B \]
- intersection
  \[ x \cap A \cap B \text{ if } x \subseteq A \text{ and } x \subseteq B \]
- product
  \[ (x,y) \subseteq A \times B \text{ if } x \subseteq A \text{ and } y \subseteq B \]
- subset
  \[ A \subseteq B \text{ if every element in } A \text{ is also in } B \]

Example: “less than”

Define \( N \) to be a set of integers:
- 0 is an integer
- if \( x \) is an integer, \( s(x) \) is an integer
  \[ 0 = 0, 1 = s(0), 2 = s(s(0)), \text{ etc} \]

The relation \( L \) (“less than”) is a subset of \( N \times N \):
- \( (0,s(0)) \subseteq L \)
- if \((x,y) \subseteq L\) then \((x,s(y)) \subseteq L\)
- if \((x,y) \subseteq L\) then \((s(x),s(y)) \subseteq L\)

A function is a special kind of relation

- \( F = D \times R \)
- \( D \) = domain, \( R \) = range
- There is at most one element \((x,y)\) in \( F \) for any given value of \( x \)

Example: “square”

\( (0,0), (1,1), (2,4), (3,9), \text{ etc} \)

A relational algebra extends the basic notion of relations as sets and defines several new operators
- elements of relations are named
- the order of the elements is not important
- \( A_{\text{name,addr}} \) is a set of name, address pairs
- \( A_{\text{name,addr}} \subseteq A_{\text{addr,name}} \)

Example:
- \( \text{tRNA} = \{ (1,TGC), (2,ACG), (3,GGC), \ldots \} \)
- \( \text{GC,codon,aa} = \{ (TTF,Phe), (TTC,Phe), (TTA,Leu), \ldots \} \)

Operations on relations:
- select
  \[ \text{tRNA} = \{ (1,TGC), (2,ACG) \} \]
- project
  \[ \text{GC,codon,aa} = \{ \text{Phe, Leu,} \ldots \} \]
  \[ \text{no duplicate entries} \]
- rename
  \[ \text{tRNA} = \text{tRNA}_L \]

Relations (cont’d)
Relational Algebra (cont’d)

- Relational Algebra

  - product (must have no element names in common)
    \[ \text{tRNA}_{\text{GC codon}} \times \text{tRNA}_{\text{GBT codon}} = \]
    \[ (1, \text{TGC}, \text{TTT}, \text{Phe}), \]
    \[ (1, \text{TGC}, \text{TTA}, \text{Leu}), \]
    \[ \ldots \]

    Note: with 20 tuples in GC and \( n \) in tRNA, the product has \( 20n \) tuples.

Relational Algebra (cont’d)

- The join operation is a combination of rename, product, select, and project

\[ \text{tRNA}_{\text{GC codon}} \bowtie \text{tRNA}_{\text{GBT codon}} = \]
\[ (1, \text{TGC}, \text{Cys}), \]
\[ (2, \text{ACG}, \text{Thr}), \]
\[ (3, \text{GGC}, \text{Gly}), \]
\[ \ldots \]

Note the two relations have an element name in common (otherwise the join is just a regular product).

Relational Database

- A relational database uses relational algebra as the basis for modeling data and operations on data
  - A database is a set of tables
  - A table is defined by a set of named columns
  - Each row in a table is a tuple

- Database schema: definition of the tables and their structures

SQL

- SQL (structured query language) is a standard language for defining database schemas and writing queries

Example:

```sql
CREATE TABLE tRNA (
    id int(4) NOT NULL,  # tRNA number
    aa char(3),  # amino acid
    codon char(3),
    gene char(12),  # name of tRNA gene
    PRIMARY KEY id
);```

SQL Query

- The SELECT statement is used to retrieve data

Examples:

```sql
SELECT * FROM tRNA;
# list all tRNA genes

SELECT name, start, end FROM Gene WHERE chromosome = "X";
# what are the names and locations of the genes on the X chromosome?
```

SQL Joins

- Joins are performed automatically when selecting items from two or more tables

```sql
SELECT name, start, end FROM Gene, tRNA
WHERE gene.name = tRNA.gene
AND chromosome = "X";
```

- Note the highlighted formula in the WHERE clause; without it the query would return all tuples in Gene \( \times \) tRNA
MySQL

- MySQL is an open source relational database system
- mysql
  - server (daemon), maintains tables, responds to queries
- mysql
  - command line interface; client that connects to a MySQL daemon (local or remote)
- mysqladmin
- mysqldump
  - many other administrator and user utilities

MySQL Interfaces

- There are many other ways to interact with a MySQL server
- phpMyAdmin
  - PHP scripts; connect via browser
- YourSQL
  - Mac OS/X application with GUI
- Perl API
  - library of Perl functions to establish a connection, send a query, process results

Perl Example

- Here is some code from a Perl program that uses MySQL for a gene database

```perl
my $dbh = DBI->connect("DBI:mysql:$db:$host", $user)
  or die "Can't connect: $DBI::errstr\n";

my $query = "LOAD DATA LOCAL INFILE '$f' INTO TABLE Genes";
$dbh->do($query)
  or die "Load failed: $DBI::errstr\n";
```

Perl Example (cont'd)

```perl
$query = "SELECT id, codon, aa FROM tRNA, GC";
my $sth = $dbh->prepare($query);
$sth->execute();
while (($id,$x,$aa) = $sth->fetchrow_array()) {
  ...
}
```

PHP Example

- Here is a similar example in PHP
- The PHP script connects to the database, sends it a query, extracts the information, formats it as an HTML table, returns it to a web browser

```php
if (!$mysql_pconnect($host,$user,$password)) {
  $error = "couldn't connect to $host as $user";
}

$result = mysql_query($query);
$n = mysql_num_fields($result);
while ($row = mysql_fetch_row($result)) {
  printf('<TR bgcolor="$row_color">
  for ($i = 0; $i < $n; $i++) {
    printf(' "%s" <TD>$row[$i]
  }
  printf('</TR>

```
MySQL and Bioinformatics

- Several bioinformatics projects use MySQL
- Human Genome Browser at UC Santa Cruz
  - Home page: genome.ucsc.edu
  - For a description of their database schema: http://genome.ucsc.edu/goldenPath/gbdDescriptions.html
- Ensembl
  - Home page: www.ensembl.org
  - Joint project of EMBL, Sanger Center
  - Raw sequence data, assembled and annotated sequences, ...
  - Extensive Perl API for using the data

Downloading a MySQL Database

- One of the advantages of MySQL: databases can be easily copied to a new system
- Database administrators at the original site run mysqldump
  - writes a file named x.sql containing all the "create table" commands for the database named x
  - writes a series of .table files containing the rows of each table
- Tar and compress the directory

Downloading (cont'd)

- To install the database at your own site:
  - create a new database on your server
    mysql> CREATE DATABASE fugu;
  - define the tables
    $ mysql -h ix -p fugu < fugu_rubripes_10_2.sql
  - load the data
    $ mysqlimport -h ix -p fugu -L *.table

Object Databases

- A topic for another day: object databases
- Different semantic foundations
  - RDB: relations, relational algebra
  - ODB: abstract data types, object-oriented programming
- Many adherents in bioinformatics
- ACeDB (A. C. elegans DataBase) now widely used as a model for several genomics databases