Relational Databases for Biologists: Efficiently Managing and Manipulating Your Data

Session 1:
Data Conceptualization and Database Design

George Bell, Ph.D.
WIBR Bioinformatics and Research Computing
What is a Database?

- A collection of data
- A set of rules to manipulate data
- A method to mold information into knowledge
- Is a phonebook a database?
  - Is a phonebook with a human user a database?

<table>
<thead>
<tr>
<th>Name</th>
<th>Address</th>
<th>Phone</th>
</tr>
</thead>
<tbody>
<tr>
<td>Babbitt, S.</td>
<td>38 William St., Cambridge</td>
<td>555-1212</td>
</tr>
<tr>
<td>Baggins, F.</td>
<td>109 Auburn Ct., Boston</td>
<td>555-1234</td>
</tr>
<tr>
<td>Bayford, A.</td>
<td>1154 William St., Newton</td>
<td>555-8934</td>
</tr>
</tbody>
</table>
Why are Databases Important?

• Data -> information -> knowledge
• Efficient manipulation of large data sets
• Integration of multiple data sources
• Adding crosslinks/references to other resources
Why is a Database Useful?

• If database systems simply manipulate data, why not use existing file system and spreadsheet mechanisms?

• “Baggins” Telephone No. Lookup:
  – Human: Look for B, then A, then G …
  – Unix: grep Baggins boston_directory.txt
  – DB: SELECT * FROM directory WHERE lastName="Baggins"

<table>
<thead>
<tr>
<th>Name</th>
<th>Address</th>
<th>Phone</th>
</tr>
</thead>
<tbody>
<tr>
<td>Babbitt, S.</td>
<td>38 William St., Cambridge</td>
<td>555-1212</td>
</tr>
<tr>
<td>Baggins, F.</td>
<td>109 Auburn Ct., Boston</td>
<td>555-1234</td>
</tr>
<tr>
<td>Bayford, A.</td>
<td>1154 William St., Newton</td>
<td>555-8934</td>
</tr>
</tbody>
</table>
What is the Advantage of a Database?

• Find all last names that contain “th” but do not have street address that begin with “th”.
  – Human: a lot of careful reading....
  – Unix: Write a directory parser and a filter.
  – DB: SELECT lastName FROM directory WHERE lastName LIKE “%th%” AND street NOT LIKE “Th%”
Why Biological Databases?

- To access and manipulate lots of data
- To manage experimental results
- To improve search sensitivity
- To improve search efficiency
- To merge multiple data sets
Microarrays: a practical application

- The typical Excel spreadsheet of microarray data

<table>
<thead>
<tr>
<th>Affy</th>
<th>lung</th>
<th>heart</th>
<th>gall_bladder</th>
<th>pancreas</th>
<th>testis</th>
</tr>
</thead>
<tbody>
<tr>
<td>92632_at</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>94246_at</td>
<td>20</td>
<td>71</td>
<td>122</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>93645_at</td>
<td>216</td>
<td>249</td>
<td>152</td>
<td>179</td>
<td>226</td>
</tr>
<tr>
<td>98132_at</td>
<td>135</td>
<td>236</td>
<td>157</td>
<td>143</td>
<td>145</td>
</tr>
</tbody>
</table>

- Find all of the genes that have at least 2-fold higher expression in the gall bladder compared to the testis, and sort by decreasing RNA abundance in the heart
Course Goals

• Conceptualize data in terms of relations (database tables)
• Design relational databases
• Use SQL commands to extract data from (mine) databases
• Use SQL commands to build and modify databases
Session Outline

• Session 1
  – Database background and design
• Session 2
  – SQL to data mine a database
• Session 3
  – SQL to create and modify a database

• Hands-on sessions after each lecture
Supplemental Information

• Links to class information:
  http://jura.wi.mit.edu/bio/education/
  bioinfo2006/db4bio/

• MySQL documentation:
  http://dev.mysql.com/doc/

• Books:
  – *MySQL* – Paul DuBois
  – and many others
Flat vs. Relational Databases

• Flat file databases use identity tags or delimited formats to describe data and categories without relating data to each other
  – Most biological databases are flat files and require specific parsers and filters

• Relational databases store data in terms of their relationship to each other
  – A simple query language can extract information from any database
Fasta format sequence file

>gi|2137523|pir||59068 MHC class I H2-K-b-alpha-2 cell surface glycoprotein - mouse (fragment)
AHTIQVISGCEVGSDGRLLRGYQQYAYDGCDYIALNEDLKTWTADAAMAILTKHKWEQAGEAERLRAYLE
GTCVEWLRRYLKNGNATLLRT

>gi|25054197|ref|XP_193866.1| histocompatibility 2, K region [Mus musculus]
MSRGRGGWSRRGSGSIGSRHHRKPRAMSRSVSEWTLRTLGGYNQSKGSHTIQVISGCEVGSDGRLLRGYQ
QYAYDGCDYIALNEDLKTWTADAAMAILTKHKWEQAGEAERLRAYLEGTCEVWLRRLRYLKNHGNATLLRTDS
PKAHTHHERSRPDKVTISRWALGFPADITLTWQLNGEELIQDMELVETRPAHGTFQKWASVVKVPGKE
QYYTCHVYHQGLPEPLTRWEPPSTVSNMATVAVLVLGAAIVTGAVVAVFVMKMRRTNGKGDGDYALA
PGSQTSDLSDLPDCKVMVHDPHSLA

>gi|25032382|ref|XP_207061.1| similar to histocompatibility 2, K region [Mus musculus]
MVPCTLLLLAAALAPTQTRAPGHSLRYFVTAVSRPGLEGRHEPRMEVGYVDDTEFVRFSDAENPRYEPRA
RWMEQEGPEYWERETQQAKGNEQSFRVDLRRTLLGGYNQSKGSHTIQVISGCEVGSDGRLLRGYQQYAYD
GCDYIALNEDLKTWTADAAMAILTKHKWEQAGEAERLRAYLEGTCEVWLRRLRYLKNHGNATLLRTDSPK
THHSRPEDKVTISRWALGFPADITLTWQLNGEELIQDMELVETRPAHGTFQKWASVVKVPGKEQYYTC
HVYHQGLPEPLTRWEPPSTVSNMATVAVLVLGAAIVTGAVVAVFVMKMRRTNGKGDGDYALAPGSQT
SDLSDLPDCKVMVHDPHSLA
GenBank format sequence file

LOCUS       NM_205137                  1111 bp    mRNA    linear    VRT 16-APR-2005
DEFINITION  Gallus gallus homeobox protein Nkx-2.8 (NKX2.8), mRNA.
ACCESSION   NM_205137 XM_444649
VERSION     NM_205137.1 GI:49170097
SOURCE      Gallus gallus (chicken)
ORGANISM    Gallus gallus
             Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
             Archosauria; Aves; Neognathae; Galliformes; Phasianidae;
             Phasianinae; Gallus.
FEATURES    Location/Qualifiers
            source  1..1111
            /organism="Gallus gallus"
            /mol_type="mRNA"
            /db_xref="taxon:9031"
            gene     1..1111
            /gene="NKX2.8"
            /db_xref="GeneID:396037"
            CDS      32..613
            /gene="NKX2.8"
            /codon_start=1
            /product="homeobox protein Nkx-2.8"
            /protein_id="NP_990468.1"
            /db_xref="GI:49170098"
            /db_xref="GeneID:396037"
            /translation="MLPTFPSVEDILSELQSSAPGAPGVRPSVEEEEPSGQCLLSQPLAQDQQQTDPCHHPKQPRKPRVLFQTSQVQLELERRFKQQYLSALEREHLANLQLQTSTQVKGQNRYYCRRQRQDRSLEMATYPLPPKAVPVPLVRNGKPCFEQSQPHLAPYGITVSYSPSYSTYSAYGVGSYGVGYTGVLTP"
ORIGIN
1 cagggagctc acacggatcc cccccggag gatgctgccc acccctttct ccgctgagga
61 tatcctcagc ctggagcaga gcagcgctcc cggagccccc ggggtctgcc gcagcccttc

Relational Databases for Biologists © Whitehead Institute, 2006
The Relational Database

• Data is composed of sets of tables and links
• Structured Query Language (SQL) to query the database
• Database management system (DBMS) to manage the data
DBMS ACID properties

• ACID properties/model
  – Atomicity: “All or nothing” transactions
  – Consistency: Only valid data can be input
  – Isolation: Multiple user independence
  – Durability: Recovery mechanisms for system failures
Selected DBMSs

• MySQL
  – “The world's most popular open source database”, probably for biology too
  – Free; open source; small application; quick to learn
  – DBMS for this class

• PostgreSQL
  – “The world's most advanced open source database”
  – Free; open source; somewhat larger application

• Oracle
  – “The worlds #1 database”
  – A lot more features but takes longer to learn
  – Expensive (but of course, many feel it’s worth it)

• All three are great choices and have the same core SQL functionality.
Data Conceptualization

- Data and Links (For a Phonebook)
Data Structure

- Data stored in tables with multiple columns ("attributes").
- Each record is represented by a row (a "tuple")

Entity = People

<table>
<thead>
<tr>
<th>First Name</th>
<th>Last Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frodo</td>
<td>Baggins</td>
</tr>
<tr>
<td>Samuel</td>
<td>Babbitt</td>
</tr>
<tr>
<td>Andrea</td>
<td>Bayford</td>
</tr>
</tbody>
</table>

Attributes

Tuples
Relational Database Specifics

• Tables are relations
  – You perform operations on the tables
• No two tuples (rows) should be identical
• Each attribute for a tuple has only one value
• Tuples within a table are unordered
• Each tuple is uniquely identified by a primary key
Primary Keys

• Primary Identifiers (IDs)
• Set of attributes that uniquely define a single, specific tuple (row)
• Must be absolutely unique
  – SSN ?
  – Phone Number ?
  – ISBN ?

<table>
<thead>
<tr>
<th>First Name</th>
<th>Last Name</th>
<th>SSN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frodo</td>
<td>Baggins</td>
<td>332-97-0123</td>
</tr>
<tr>
<td>Frodo</td>
<td>Binks</td>
<td>398-76-5327</td>
</tr>
<tr>
<td>Maro</td>
<td>Baggins</td>
<td>215-01-3965</td>
</tr>
</tbody>
</table>
## Find the Keys

<table>
<thead>
<tr>
<th>First Name</th>
<th>Last Name</th>
<th>SSN</th>
<th>Phone Number</th>
<th>Address</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frodo</td>
<td>Baggins</td>
<td>321-45-7891</td>
<td>123-4567</td>
<td>29 Hobbitville</td>
</tr>
<tr>
<td>Aragon</td>
<td>Elf-Wantabe</td>
<td>215-87-7458</td>
<td>258-6109</td>
<td>105 Imladris</td>
</tr>
<tr>
<td>Boromir</td>
<td>Ringer</td>
<td>105-91-0124</td>
<td>424-9706</td>
<td>31 Hobbitville</td>
</tr>
<tr>
<td>Bilbo</td>
<td>Baggins</td>
<td>198-02-2144</td>
<td>424-9706</td>
<td>29 Hobbitville</td>
</tr>
<tr>
<td>Legolas</td>
<td>Elf</td>
<td>330-78-4230</td>
<td>555-1234</td>
<td>135 Imladris</td>
</tr>
</tbody>
</table>
Design Principles

• Conceptualize the data elements (entities)
• Identify how the data is related
• Make it simple
• Avoid redundancy
• Make sure the design accurately describes the data!
Entity-Relationship Diagrams

- Expression of a database table design
E-R to Table Conversion

Entity

<table>
<thead>
<tr>
<th>fName</th>
<th>lName</th>
<th>P_Id</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>P_Id</th>
<th>Add_Id</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Add_Id</th>
<th>St_No</th>
<th>St_Name</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Relationship

<table>
<thead>
<tr>
<th>P_Id</th>
<th>Tel_Id</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Add_Id</th>
<th>Tel_Id</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>A_Code</th>
<th>Number</th>
<th>Tel_Id</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

All Tables Are Relations
Steps to Build an E-R Diagram

- Identify data attributes
- Conceptualize entities by grouping related attributes
- Identify relationships/links
- Draw preliminary E-R diagram
- Add cardinalities and references
Developing an E-R Diagram

• Convert a GenBank File into an E-R Diagram

```
LOCUS NM_205137  1111 bp  mRNA  linear  VRT 16-APR-2005
DEFINITION Gallus gallus homeobox protein Nkx-2.8 (NKX2.8), mRNA.
ACCESSION NM_205137 XM_444649
VERSION NM_205137.1  GI:49170097
SOURCE Gallus gallus (chicken)
ORGANISM
   Gallus gallus
   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
   Archosauria; Aves; Neognathae; Galliformes; Phasianidae;
   Phasianinae; Gallus.
FEATURES Location/Qualifiers
    source  1..1111
       /organism="Gallus gallus"
       /mol_type="mRNA"
       /db_xref="taxon:9031"
    gene  1..1111
       /gene="NKX2.8"
       /db_xref="GeneID:396037"
    CDS  32..613
       /gene="NKX2.8"
       /codon_start=1
       /protein_id="NP_990468.1"
       /db_xref="GI:49170098"
       /db_xref="GeneID:396037"
       /translation="MLPTFSSVEDILSLEQSSAGAPGVPVRRSPSVVVEFEPQCSSQCLLSQPQLADVQWTDFCHIPQQRKPVRLFSQIQLELRFKQQKYSLYLDHTEHRANQLTPLGTVKIFQRYCKQEQDRSLEMATYPLPRKVAVFVPVRNGKPCFEQGQPHLAPYGNITVSPYSGTYSSAYGVGTVGVTVLTP"
ORIGIN
    1 cagggagctc acaccgatcc ccccccggag gatgctgccc acccctttct ccgtcgagga
     61 tatcctcagc ctggagcaga gcagcgctcc cggagcccgg ggggctccgc gcagcccttc
```
Identify Attributes

- Locus, Definition, Accession, Version, Source Organism
- Authors, Title, Journal, Medline Id, PubMed Id
- Protein Name, Protein Description, Protein Id, Protein Translation, Locus Id, GI
- A count, C count, G count, T count, Sequence
Identify Entities by Grouping

- Gene
  - Locus, Definition, Accession, Version, Source Organism
- References
  - Authors, Title, Journal, Medline Id, PubMed Id
- Features
  - Protein Name, Protein Description, Protein Id, Protein Translation, Locus Id, GI
- Sequence Information
  - A count, C count, G count, T count, Sequence
Conceptualize Entities

- Features
- Gene
- References
- Sequence Info
Identify Relationships

- Features
- Gene
- References
- Sequence Info
Cardinalities and References

Features
- Name
- LocusId
- Trans.
- Descr.
- ID
- GI
- Acc.
- Version
- Source
- Defin.
- Locus

Gene
- Acc. ID
- Locus
- Defin.
- Acc.
- Source
- Version
- Acc. Seq.
- Acc.
- PubMed
- Title
- Journal
- Authors
- Reference
- PubMed
- Medline
- A_count
- G_count
- T_count
- C_count

Sequence Info
- Sequence
- A_count
- G_count
- T_count
- C_count

Relational Databases for Biologists © Whitehead Institute, 2006
Apply Design Principles

- Faithful, non-redundant
- Simple element choice
Summary

- Databases provide ACID
- Databases are composed of tables (relations)
- Relations are entities that have attributes and tuples
- Databases can be designed from E-R diagrams that are easily converted to tables
- Primary keys uniquely identify individual tuples and represent links between tables
Next class

- Using structured query language (SQL) to data mine databases

- SELECT a FROM b WHERE c = d
Database design example:

Design the db4bio database
Build Your Own E-R Diagram

• Express the following annotated microarray data set as an E-R diagram

<table>
<thead>
<tr>
<th>AffyId</th>
<th>GenBankId</th>
<th>Name</th>
<th>Description</th>
<th>LocusLinkId</th>
<th>LocusDescrip</th>
<th>NT_RefSeq</th>
<th>AA_RefSeq</th>
</tr>
</thead>
<tbody>
<tr>
<td>U95-32123_at</td>
<td>L02870</td>
<td>COL7A1</td>
<td>Collagen</td>
<td>1294</td>
<td>Collagen</td>
<td>NM_000094</td>
<td>NP_000085</td>
</tr>
<tr>
<td>U98-40474_at</td>
<td>S75295</td>
<td>GBE1</td>
<td>Glucan</td>
<td>2632</td>
<td>Glucan</td>
<td>NM_000158</td>
<td>NP_000149</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>UnigeneliD</th>
<th>GO Acc.</th>
<th>GO Descr.</th>
<th>Species</th>
<th>Source</th>
<th>Level</th>
<th>Experiment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hs.1640</td>
<td>0005202</td>
<td>Serine Prot. Hs</td>
<td>Pancreas</td>
<td>128</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Hs.1691</td>
<td>0003844</td>
<td>Glucan Enz. Hs</td>
<td>Liver</td>
<td>57</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>
### Identify Attributes

<table>
<thead>
<tr>
<th>AffyId</th>
<th>GenBankId</th>
<th>Name</th>
<th>Description</th>
<th>LocusLinkId</th>
<th>LocusDescr</th>
<th>NT_RefSeq</th>
<th>AA_RefSeq</th>
</tr>
</thead>
<tbody>
<tr>
<td>U95-32123_at</td>
<td>L02870</td>
<td>COL7A1</td>
<td>Collagen</td>
<td>1294</td>
<td>Collagen</td>
<td>NM_000094</td>
<td>NP_000085</td>
</tr>
<tr>
<td>U98-40474_at</td>
<td>S75295</td>
<td>GBE1</td>
<td>Glucan</td>
<td>2632</td>
<td>Glucan</td>
<td>NM_000158</td>
<td>NP_000149</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>UnigenId</th>
<th>GO Acc.</th>
<th>GO Descr.</th>
<th>Species</th>
<th>Source</th>
<th>Level</th>
<th>Experiment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hs.1640</td>
<td>0005202</td>
<td>Serine Prot. Hs</td>
<td>Pancreas</td>
<td>128</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Hs.1691</td>
<td>0003844</td>
<td>Glucan Enz. Hs</td>
<td>Liver</td>
<td>57</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>
Identify Entities by Grouping

- Gene Descriptions
  - Name, Description, GenBank
- RefSeqs
  - NT RefSeq, AA RefSeq
- Ontologies
  - GO Accession, GO Terms
- LocusLinks
- Unigenes
- Data
  - Sample Source, Level
- Targets
  - Affy ID, Experiment Number, Species
Conceptualize Entities

- Gene Descriptions
- Ontologies
- Targets
- Data
- RefSeqs
- Unigenes
- LocusLinks
Identify Relationships

Unigenes → LocusLinks → Targets → RefSeqs → Ontologies → Gene Descriptions
Species → Data
Cardinalities and References
Apply Design Principles

Unigenes 1..1

Unisecs 1..1

Targets 1..1 0..n 1..1 0..n

Species 1..n

Data 1..1

Sources 1..1

Ontologies 1..n

LocusDescr 1..1

LocusLinks 1..1 0..n 0..n

RefSeqs 0..1

Descriptions 1..1

GO_Dscr 1..1

GBId

GoAcc

LinkId

UId

UniSeqs linkId

ntRefSeq

aaRefSeq

description

gbId

goAcc

level

species

exptId

name

gbId

exptId

source

Static Annotation

Experimental