Relational Databases for Biologists: Efficiently Managing and Manipulating Your Data

Session 1
Data Conceptualization
and Database Design

Robert Latek, Ph.D.
Sr. Bioinformatics Scientist
Whitehead Institute for Biomedical Research

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?

Babbitt, S.	38 William St., Cambridge	555-1212
Baggins, F.	109 Auburn Ct., Boston	555-1234
Bayford, A.	1154 William St., Newton	555-8934

Why are Databases Important?

- Data -> Information -> Knowledge
- Efficient Manipulation of Large Data Sets
- Integration of Multiple Data Sources
- Cross-Links/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 lName=“Baggins”

Babbitt, S.	38 William St., Cambridge	555-1212
Baggins, F.	109 Auburn Ct., Boston	555-1234
Bayford, A.	1154 William St., Newton	555-8934

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: Good Luck!
  - UNIX: Write a directory parser and a filter.
  - DB: SELECT lName FROM directory WHERE lName LIKE “%th%” AND street NOT LIKE “Th%”

Why Biological Databases?

- Too Much Data
- Managing Experimental Results
- Improved Search Sensitivity
- Improved Search Efficiency
- Joining of Multiple Data Sets

Still Not Convinced?

- The Typical Excel Spreadsheet of Microarray Data

Affy	lung	cardiac	gall_bladder	pancreas	testis
92632_at	20	20	20	20	20
94246_at	20	71	122	20	20
93645_at	216	249	152	179	226
98132_at	135	236	157	143	145

- Now Find All of the Genes that have 2-3 fold Over-Expression in the Gall Bladder Compared to the Testis

Mini-Course Goals

- Conceptualize Data in Terms of Relations (Database Tables)
- Design Relational Databases
- Use SQL Commands to Extract/Data 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
- Demonstration and Lab

Supplemental Information

- http://jura.wi.mit.edu/bio/education/bioinfo-mini/db4bio/
- http://www.mysql.com/documentation/
- A First Course In Database Systems. Ullman and Widom .
  - ISBN:0-13-861337-0

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

GenBank Report

```
LOCUS       H2-K                1585 bp    mRNA    linear   ROD 18-NOV-2002
DEFINITION  Mus musculus histocompatibility 2, K region (H2-K), mRNA.
ACCESSION   XM_193866
VERSION     XM_193866.1  GI:25054196
KEYWORDS    .
SOURCE      Mus musculus (house mouse)
ORGANISM  Mus musculus.
REFERENCE   1  (bases 1 to 1585)
AUTHORS   NCBI Annotation Project.
TITLE     Direct Submission
JOURNAL   Submitted (13-NOV-2002) National Center for Biotechnology
COMMENT     GENOME ANNOTATION REFSEQ
FEATURES             Location/Qualifiers
     source          1..1585
                     /organism="Mus musculus"
                     /strain="C57BL/6J"
                     /db_xref="taxon:10090"
                     /chromosome="17"
     gene            1..1585
                     /gene="H2-K"
                     /db_xref="LocusID:14972"
                     /db_xref="MGI:95904"
     CDS             223..1137
                     /gene="H2-K"
                     /codon_start=1
                     /product="histocompatibility 2, K region"
                     /protein_id="XP_193866.1"
                     /translation="MSRGRGGWSRRGPSIGSGRHRKPRAMSRVSEWTLRT...
BASE COUNT     350 a   423 c   460 g   352 t
ORIGIN
        1 gaagtcgcga atcgccgaca ggtgcgatgg taccgtgcac gctgctcctg ctgttggcgg
```

NCBI NR Database File

>gi|2137523|pir||I59068 MHC class I H2-K-b-alpha-2 cell surface glycoprotein - mouse (fragment)
AHTIQVISGCEVGSDGRLLRGYQQYAYDGCDYIALNEDLKTWTAADMAALITKHKWEQAGEAERLRAYLE
GTCVEWLRRYLKNGNATLLRT

>gi|25054197|ref|XP_193866.1| histocompatibility 2, K region [Mus musculus]
MSRGRGGWSRRGPSIGSGRHRKPRAMSRVSEWTLRTLLGYYNQSKGGSHTIQVISGCEVGSDGRLLRGY
Q
QYAYDGCDYIALNEDLKTWTAADMAALITKHKWEQAGEAERLRAYLEGTCVEWLRRYLKNGNATLLRTDS
PKAHVTHHSRPEDKVTLRCWALGFYPADITLTWQLNGEELIQDMELVETRPAGDGTFQKWASVVVPLGKE
QYYTCHVYHQGLPEPLTLRWEPPPSTVSNMATVAVLVVLGAAIVTGAVVAFVMKMRRRNTGGKGGDYALA
PGSQTSDLSLPDCKVMVHDPHSLA

>gi|25032382|ref|XP_207061.1| similar to histocompatibility 2, K region [Mus musculus]
MVPCTLLLLLAAALAPTQTRAGPHSLRYFVTAVSRPGLGEPRYMEVGYVDDTEFVRFDSDAENPRYEPRA
RWMEQEGPEYWERETQKAKGNEQSFRVDLRTLLGYYNQSKGGSHTIQVISGCEVGSDGRLLRGYQQYAY
D
GCDYIALNEDLKTWTAADMAALITKHKWEQAGEAERLRAYLEGTCVEWLRRYLKNGNATLLRTDSPKAHV
THHSRPEDKVTLRCWALGFYPADITLTWQLNGEELIQDMELVETRPAGDGTFQKWASVVVPLGKEQYYTC
HVYHQGLPEPLTLRWEPPPSTVSNMATVAVLVVLGAAIVTGAVVAFVMKMRRRNTGGKGGDYALAPGSQT
SDLSLPDCKVMVHDPHSLA

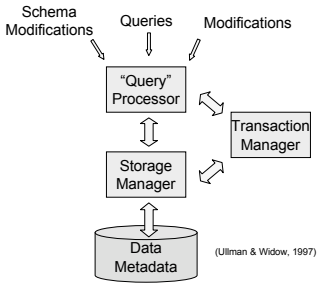
The Relational Database

- Data is Composed of Sets of Tables and Links
- Structured Query Language (SQL) to Query the Database
- DBMS to Manage the Data

DBMS

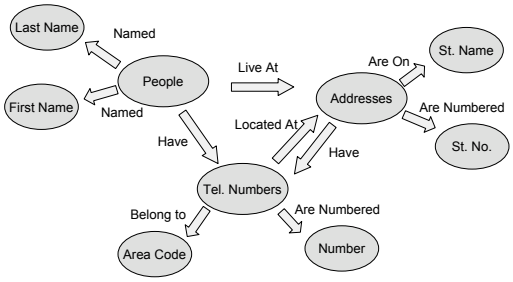
- Database Management System (ACID)
  - Atomicity: Data independence
  - Consistency: Data integrity and security
  - Isolation: Multiple user accessibility
  - Durability: Recovery mechanisms for system failures

DBMS Architecture



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 (Tuple)

Entity = People

First Name	Last Name
Frodo	Baggins
Samuel	Babbitt
Andrea	Bayford

⇐ Attributes

⇐ Tuples

Relational Database Specifics

- Tables are Relations
  - You perform operations on the tables
- No Two Tuples (Rows) can 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 (Record)
- Must be Absolutely Unique
  - SSN ?
  - Phone Number ?
  - ISBN ?

First Name	Last Name	SSN
Frodo	Baggins	332-97-0123
Frodo	Binks	398-76-5327
Maro	Baggins	215-01-3965

Find the Keys

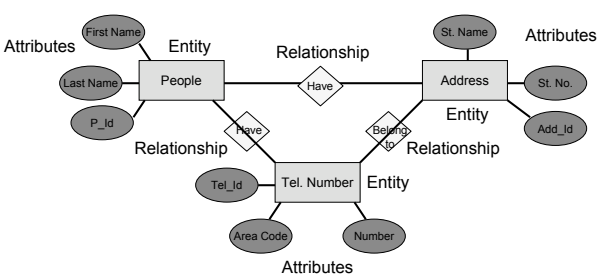
First Name	Last Name	SSN	Phone Number	Address
Frodo	Baggins	321-45-7891	123-4567	29 Hobbitville
Aragon	Elf-Wantabe	215-87-7458	258-6109	105 Imladris
Boromir	Ringer	105-91-0124	424-9706	31 Hobbitville
Bilbo	Baggins	198-02-2144	424-9706	29 Hobbitville
Legolas	Elf	330-78-4230	555-1234	135 Imladris

Design Principles

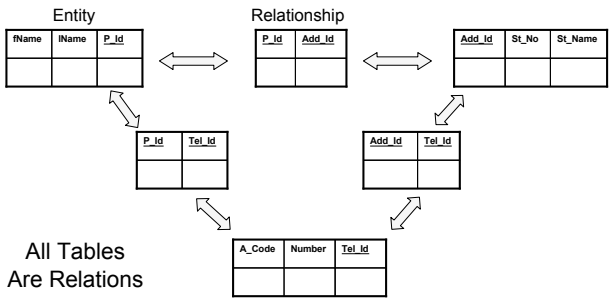
- Conceptualize the Data Elements (Entities)
- Identify How the Data is Related
- Make it Simple
- Avoid Redundancy
- Make Sure the Design Accurately Describes the Data!

Entity-Relationship Diagrams

- Expression of a Database Table Design



E-R to Table Conversion



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       IL2RG              1451 bp    mRNA    linear   PRI 17-JAN-2003
DEFINITION  Homo sapiens interleukin 2 receptor, gamma (severe combined immunodeficiency) (IL2RG), mRNA.
ACCESSION   NM_000206
VERSION     NM_000206.1  GI:4557881
ORGANISM  Homo sapiens
REFERENCE   1  (bases 1 to 1451)
AUTHORS   Takeshita,T., Asao,H., Ohtani,K., Ishii,N., Kumaki,S., Tanaka,N.,Munakata,H., Nakamura,M. and Sugamura,K.
TITLE     Cloning of the gamma chain of the human IL-2 receptor
JOURNAL   Science 257 (5068), 379-382 (1992)
MEDLINE   92335883
PUBMED   1631559
REFERENCE   2  (bases 1 to 1451)
AUTHORS   Noguchi,M., Yi,H., Rosenblatt,H.M., Filipovich,A.H., Adelstein,S., Modi,W.S., McBride,O.W. and Leonard,W.J.
TITLE     Interleukin-2 receptor gamma chain mutation results in X-linked severe combined immunodeficiency in humans
JOURNAL   Cell 73 (1), 147-157 (1993)
MEDLINE   93214986
PUBMED   8462096
CDS             15..1124
                /gene="IL2RG"
                /product="interleukin 2 receptor, gamma chain, precursor"
                /protein_id="NP_000197.1"
                /db_xref="GI:4557882"
                /db_xref="LocusID:3561"
                /translation="MLKPSLPFTSLLFLQLPLLGVGLNTTILTPNGNEDTTADFFLTT…"
BASE COUNT    347 a    422 c    313 g    369 t
ORIGIN
        1 gaagagcaag cgccatgttg aagccatcat taccattcac atccctctta ttcctgcagc
```

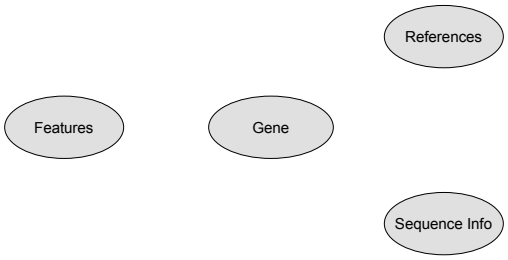

Identify Attributes

- 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

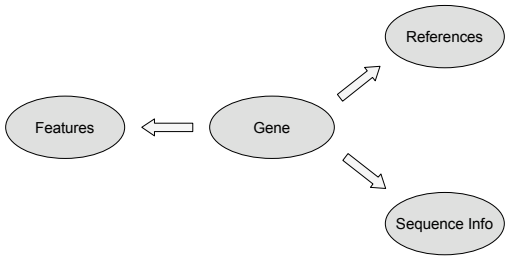
Identify Entities by Grouping

- 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

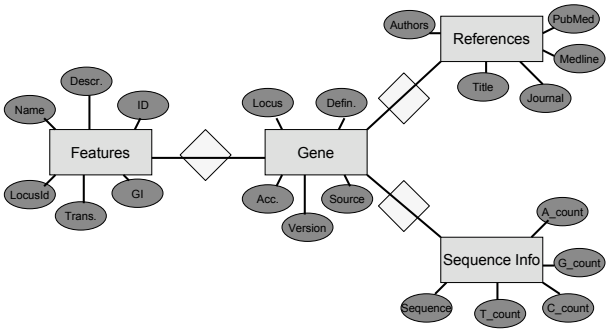
Conceptualize Entities



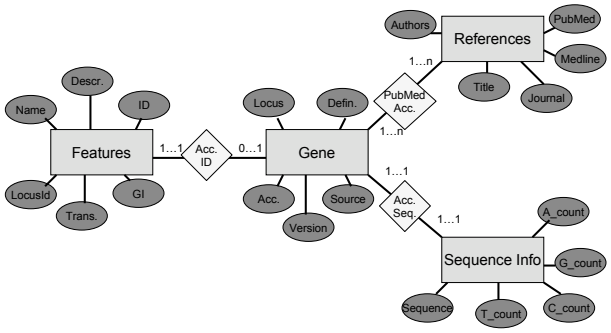
Identify Relationships



Preliminary E-R Diagram

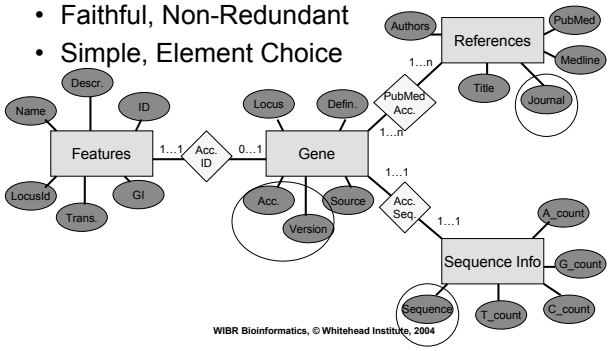


Cardinalities and References



Apply Design Principles

- Faithful, Non-Redundant
- Simple, Element Choice



Build Your Own E-R Diagram

- Express the Following Annotated Microarray Data Set as an E-R Diagram

AffyId	GenBankId	Name	Description	LocusLinkId	LocusDescr	NT_RefSeq	AA_RefSeq	\\
U95-32123_at	L02870	COL7A1	Collagen	1294	Collagen	NM_000094	NP_000085	\\
U98-40474_at	S75295	GBE1	Glucan	2632	Glucan	NM_000158	NP_000149	\\

UnigeneId	GO Acc.	GO Descr.	Species	Source	Level	Experiment
Hs.1640	0005202	Serine Prot.	Hs	Pancreas	128	1
Hs.1691	0003844	Glucan Enz.	Hs	Liver	57	2

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 Week

- Using Structured Query Language (SQL) to Data Mine Databases
- SELECT a FROM b WHERE c = d
- 5th Floor Conference Room on Monday, February 10.

Identify Attributes

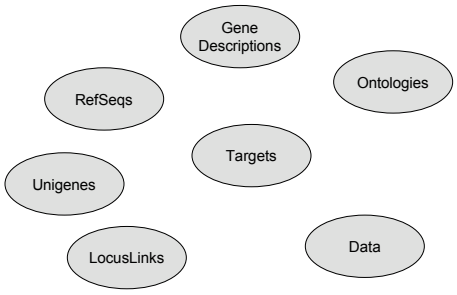
AffyId	GenBankId	Name	Description	LocusLinkId	LocusDescr	NT_RefSeq	AA_RefSeq	\\
U95-32123_at	L02870	COL7A1	Collagen	1294	Collagen	NM_000094	NP_000085	\\
U98-40474_at	S75295	GBE1	Glucan	2632	Glucan	NM_000158	NP_000149	\\

UnigeneId	GO Acc.	GO Descr.	Species	Source	Level	Experiment
Hs.1640	0005202	Serine Prot.	Hs	Pancreas	128	1
Hs.1691	0003844	Glucan Enz.	Hs	Liver	57	2

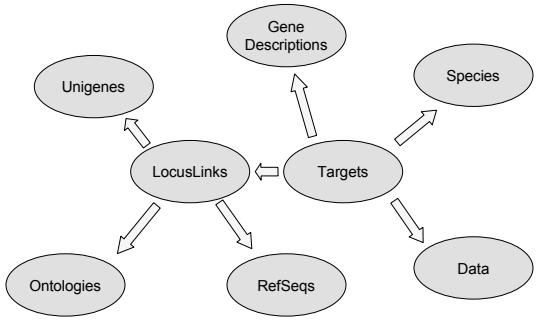
Identify Entities by Grouping

- 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

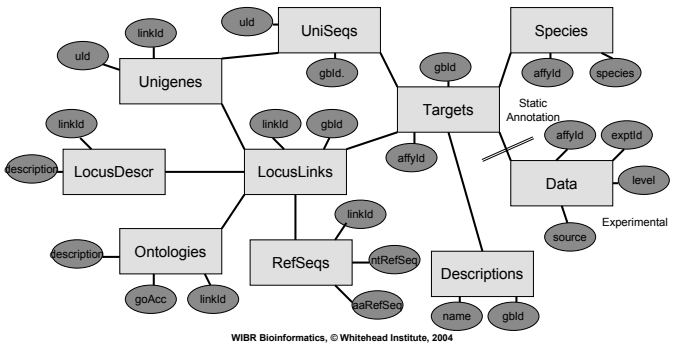
Conceptualize Entities



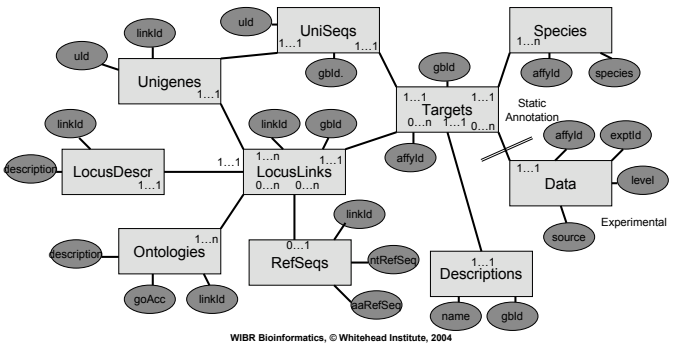
Identify Relationships



Preliminary E-R Diagram



Cardinalities and References



Apply Design Principles