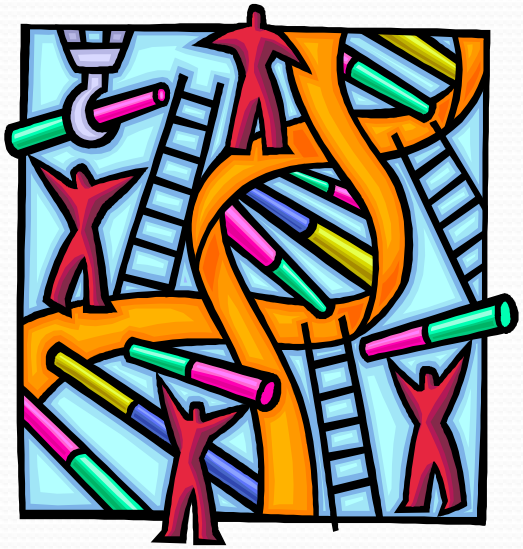
Computing Concepts for Bioinformatics

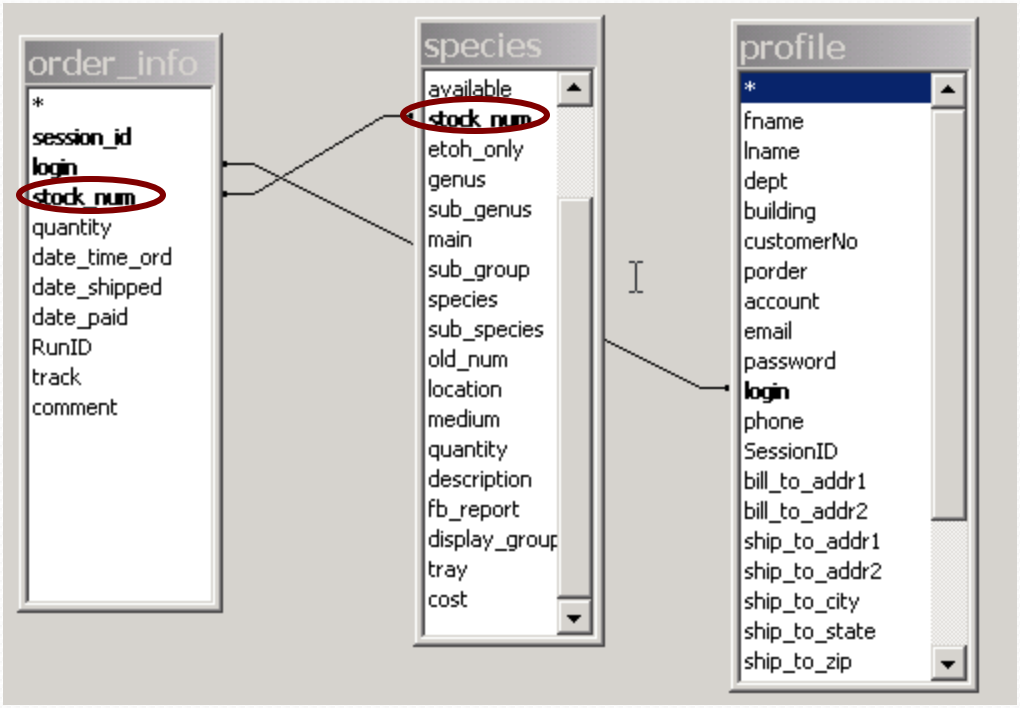


- Introduction to databases
- Using sqlite
- More database concepts

Relational Database

- A relational database is a collection of data items organized as a set of formally-described tables from which data can be accessed or reassembled in many different ways without having to reorganize the database tables
- The relational database was invented by E. F. Codd at IBM in 1970.
- A relational database is a set of tables containing data fitted into predefined categories
- Each table contains one or more data categories in columns.
- Each row contains a unique instance of data for the categories defined by the columns.
- The standard user and application program interface to a relational database is the structured query language (SQL)

Tables and Relations



order_info : Table

session_id	login	stock_num
f9a1eee27700106dc790cd5	lujing	14030-0781.4
f9a1eee27700106dc790cd5	lujing	14030-0771.16
f9a1eee27700106dc790cd5	lujing	14030-0801.0
be95d3245bde56a4ce1375:	gailphilli	14028-0681.3
be95d3245bde56a4ce1375:	gailphilli	14028-0681.5
be95d3245bde56a4ce1375:	gailphilli	14030-0811.0
be95d3245bde56a4ce1375:	gailphilli	15010-1051.0
be95d3245bde56a4ce1375:	gailphilli	15010-1051.9
be95d3245bde56a4ce1375:	gailphilli	15085-1641.0
be95d3245bde56a4ce1375:	gailphilli	15085-1641.4
128dedef9738280fb7818cb	Li Xu	14021-0251.68
128dedef9738280fb7818cb	Li Xu	14021-0251.68
3f446453dd77a383b579b3E	MCJB	15010-1051.13
c4911dde2e0ceca2f6cc91E	hsmalik	15010-0951.0
c4911dde2e0ceca2f6cc91E	hsmalik	15010-0951.14

species : Table

genus_group	species_sub	strain	stock_num	genus	sub_genus	main
11010	21	0	11010-0021.0	Scaptodrosophi	scaptodrosophil	victoria
11010	31	0	11010-0031.0	Scaptodrosophi	scaptodrosophil	victoria
11010	41	0	11010-0041.0	Scaptodrosophi	scaptodrosophil	victoria
11010	41	1	11010-0041.1	Scaptodrosophi	scaptodrosophil	victoria
11010	45	0	11010-0045.0	Scaptodrosophi	scaptodrosophil	victoria
11020	51	0	11020-0051.0	Scaptodrosophi	scaptodrosophil	coracina
11030	61	0	11030-0061.0	Scaptodrosophi	scaptodrosophil	latifasciaeformis
11030	61	1	11030-0061.1	Scaptodrosophi	scaptodrosophil	latifasciaeformis

Buzz words you must know

- Schemas or conceptual view
  Describes the overall organization / structure of the database
- Domains
  Describes what values can be stored in the column of a given table
- Constraints
  Rules that govern what values can be stored in a column

Many Many more to follow !!

Structured Query Language (SQL)

- Standard interactive and programming language for getting information from and updating a database
- SQL is both an ANSI and an ISO standard
- Was a non procedural language but from SQL:1999 onwards it became procedural
- SQL can be considered a special purpose language it needs a wrapper to talk to database i.e Perl, C, Java
- Every vendor has its own unique implementation of SQL, even though they all follow the SQL standard there are subtle variances and supported/unsupported calls.
- You Query a database using SQL, if a match is found the data is returned

SQL components

- Data Definition Language (DDL)
  Deals with structural aspect of the database: creation, modification, deletion of tables
- Data Manipulation Language (DML)
  This allows modification of the data contained in the tables: insertion, deletion, selection,changing (even aggregation i.e count,sum,average)
- Data Control Language (DCL)
  This deals with maintaining the integrity of the database using permissions, transactions etc.

Getting to know "sqlite"

- Log on to your account on login.hpc.arizona.edu
- Lets get a sample database http://ccp.arl.arizona.edu/dthompso/sql_workshop_files/genotypes.sqlite
- Now lets open the genotype.sqlite with sqlite3
  sqlite3 genotype.db
- Type .help
- Type .tables
  what do you see ?

Some SQL basics

- To store data the database uses tables
- Tables consists of rows and columns
- Column names have to be unique
- CREATE is for generating tables
- ALTER for making changes to the tables
- DROP for deleting the tables
- SELECT is for ?
- UPDATE
- JOIN
- DELETE

Some Common Column types (SQLite)

- Check: http://www.hwaci.com/sw/sqlite/datatype3.html

For details

- **NULL**. The value is a NULL value.
- **INTEGER**
- **REAL**. The value is a floating point value,
- **TEXT**.
- **BLOB**.

Your first query !

- When writing a SQL query, it is common practice to write SQL commands in uppercase.
- The -- command indicates a comment, and the database ignores everything else on the rest of the line.
- The SELECT command tells the database which data fields to retrieve.
- The FROM command tells the database which table to fetch the data from.
- Some databases care about table and column name case, but others don't, so it's best to always use the correct case when referencing tables and columns.
- The end of a query is always marked with a semicolon ;.
- -- This query selects the data in all columns
  -- from the table 'loci'
  SELECT * FROM loci;

Having fun with SELECT

- Lets jump to a good resources created by David Thompson
- http://ccp.arl.arizona.edu/dthompso/sql_workshop/sql/select.html

End the torture …give me a GUI

- You can use many different GUI for sqlite
- SQLite Database browser

http://sourceforge.net/projects/sqlitebrowser/files/sqlitebrowser/1.3/

- Mike T's SQLite admin tool

http://saxmike.com/MySoftware/MySoftware.asp?Menu=MYSOFTWARE

- Both are installed in the BLC lab.

Hands on exercise

- We will import data from a file into the database

http://amadeus.biosci.arizona.edu/~nirav/cds_product.txt

- Create database analysis.db using
  sqlite3 analysis.db
- Now create a table my_results to store analysis
  create table my_results (locus TEXT ,secondary_tag TEXT , start INTEGER , stop INTEGER);
- sqlite> .mode tabs
- sqlite> .import cds_product.txt my_results
- Have fun with SQL statements
  select distinct(locus) from my_results;
  select locus, start from my_results where start > 100;

Caveat

- Covering database design concepts in details is out of scope for this “introductory” section.
- Students wanting to learn more about database design are encouraged to pursue classes in the CS and MIS departments (with blessings from their advisors)
- CSc 460 DATABASE SYSTEMS
- MIS 535 Data Management: Technology and Applications

Some design concepts !

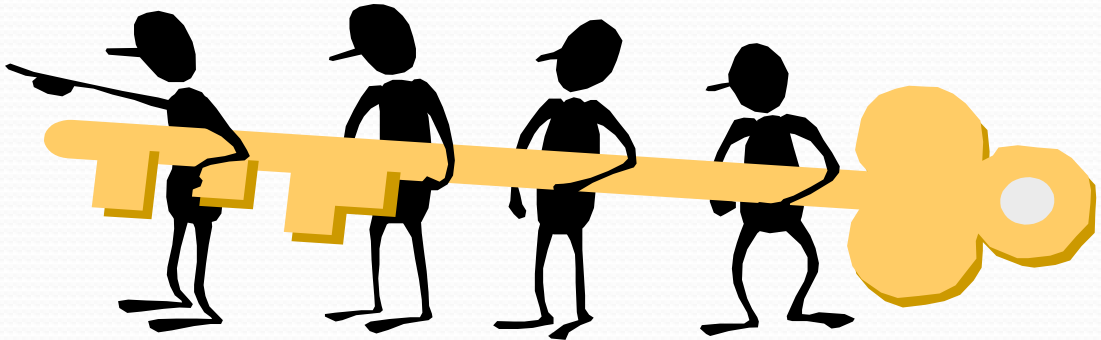
- Database design is not software or database specific
- Basic steps include:
  - **Defining the problem or objective**
  - **Researching the current database**
  - **Designing the data structures**
  - **Constructing relationships**
  - **Implementing rules and constraints**
  - **Creating views and reports**
  - **Implementing the design**

Normalization

- Is your database normalized ?
- Is that BCNF ?
- If you hesitate in answering you are not worthy !
  (BCNF: Boyce-Codd Normal Form)
- Normalization is a way to efficiently organizing data in your database (almost like closet cleaning)
- The goal is to:
  Eliminate redundancy in data
  Ensure data dependencies

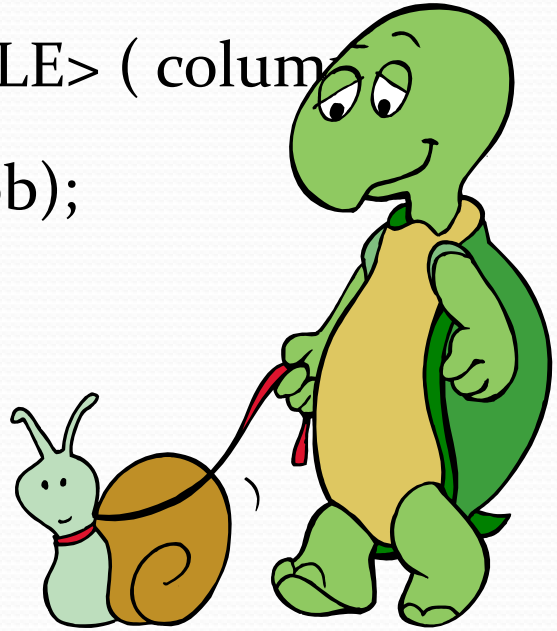
Keys

- A Key is a column or a collection of columns that uniquely identifies a row in a table
- 2 types of keys:
  Primary (composite key is a collection of columns)
  Foreign
- In many cases, data table keys are constructed by simply adding an additional field to function as the key
- Can primary key be NULL or have duplicate values ?
- Foreign key is a column or a collection of columns in a table that reference a primary key in another table



Index

- Data listed in a table is based on the order it was entered
- As the amount of data increases (number of rows), the database has to sort through more information (becoming slow)
- Index is supplementary to a table and keeps track of the corresponding rows
- Syntax:
  create index <index name> on <TABLE> (colum
  to index)
  create index by_id on patients (id,dob);



Typical Errors

- Spreadsheet design.
- Too much data.
- Compound fields.
- Missing keys.
- Bad keys.
- Missing relations.
- Unnecessary relationships.
- Incorrect relations.
- Duplicate field names.
- Cryptic field and table names.
- Missing or incorrect business rules.
- Referential integrity.
- Database security.
- International issues.