Bioinformatics Databases Spring 2016 Class Project

Document Purpose: describe project requirements (as seen in the Lecture 1 Notes and partially described in Homework 5, 6 and 7).

Construct a database that houses next-generation sequence data that a topic you selected with sequence data. Once you have a focus area (scientific domain), decide on 5-10 specific questions that interest you – they should not be simple. To set up for the topic and the scope of the database, you must pose at least 5 scientific questions, obtain the data that will be required (note that this might be in a preliminary form – a series of transformation and filters may be required, but the starting data must be available) then design the database using the relational model and instantiate it and populate it using an RDBMS (MySQL unless you have experience with a different application) to allow you to answer them. After testing the database for correct structure and content you will formulate the queries that represent the SQL equivalent of your scientific questions and run them against your database. You will indicate how you have tested the queries for accuracy. You will be required to demonstrate a functioning instance of your database and give it context in an oral presentation to the class, and you will also write a paper in a journal format to explain your process and interpret your results.

Responsibilities

You may work in pairs, but you do not have to. It is preferred that you do not. Team members can work within the same general domain of interest but must have a set of unique questions that are not trivially different from their partners. In practice team members often have 1-2 tables of data that they use and their partner does not. Each team member must be prepared to

1. Explain the focus area and its importance
2. Present to the class a live demo of the database (which will be the same for both team members) but in the presentation use independently designed slides for an independent set of 5 questions of scientific interest that the database can answer.
3. Produce an individual project report – the overall work flow and schema will be the same for both members of the team, and testing the database schema and population, but each team member should explain what part of that work was his/her responsibility and of course the Introduction, Results, Discussion and some of the References will differ.

Setting the Scientific Domain

You are to show that you can manage biological sequence data and relevant annotations in order to answer a number of questions of interest to chloroplast biologists. Some things you could consider as a focus area include:

1. Are there positions that vary within one organism (you can find this out because NGS data covers each position in the genome many times, on average).
2. What genome positions vary between organisms?
3. Are there sequence changes to any of the chloroplast genes regulated by microRNAs?
4. Do any sequence changes result in amino acid changes?
5. Are amino acid changes likely to affect the structure or function of the protein, including protein-protein or protein-nucleic acid interaction sites?
6. If you look at gene expression data (from qRT-PCR or microarrays or RNA-Seq assays) are there changes in regulatory regions that correlate with the expression changes?
7. If there are gene expression changes are they shared by all of the genes in a network and if so is there a transcription factor whose site has been modified?
8. Are there changes in the phenotype of the organism that might be related to changes in the chloroplast DNA sequence? Sugar use, for example with the Golden-2 like transcription factor is known to be correlated to a truncated nuclear gene that changes chloroplast expression in tomato.

Scientific Questions (examples)

Once you have decided on a general domain to study (for example cp protein-binding interactions) then write out a number of questions within that domain. You are required to have 5. At early stages you might want a few more, in case you can't find data to support some of them. Usually people end up with a lot more questions that 5 that can be answered, however. Some examples could be

1. Which chloroplast proteins have demonstrated protein-protein interactions?
2. Which nuclear proteins have demonstrated protein-protein interactions with chloroplast proteins
3. Which nuclear proteins have been shown to bind to chloroplast DNA?
4. Are there sequence changes to chloroplast DNA that occur in the regions where nuclear proteins bind?
5. Are there sequence changes in the chloroplast proteins that bind to other chloroplast proteins and are they in the region where the interaction is supposed to occur?
6. Etc.

Data Requirements:

1. Sufficient raw data to show the queries work. A good working sample size for most studies is around 1500. Although this many samples is not required for this topic do consider using more than 10.
2. Include next-generation sequence data or protein sequence data from at least one public project.
3. Tables that contain the data lineage- chain of custody within the schema.
4. Integrate at least one public ontology where the attribute values are meaningful (the Gene Ontology and Sequence Ontology are pretty obvious, but you could also use a plant anatomy ontology, or a plant metabolite ontology).
5. You will have domain-specific data needs – for my protein binding component I will have to look at protein interaction and/or plant protein datasets.

Note: at least two of your tables must have >10,000 rows of data (pretty easy with NGS sequence or microarray data). If you plan to have more than 1,000,000 rows please be aware that you may have performance issues that require tuning.

Note: Access to some databases requires that you register as a user, and while most of them give immediate access there is sometimes a delay. For some human genetic databases/data you must have your advisor co-sign responsibility with you. Do not assume that once you have been granted an account the database will contain the data you expect – always check as soon as you can that the data is there, available to you and can be downloaded in a usable form and volume.

Modeling (include as appendices to the project, these are completed as homework assignments):

1. Produce an annotated conceptual model (the process flow) using correct conventions.
    a. This will include provenance of the data sources and transformations, so that anyone could find the original data, format it as you did, carry out any transformations required to get the datasets that are ultimately loaded in the database.
2. Produce an annotated logical model, specialized to the entity-relational framework, using correct conventions.
3. Produce a schema using the relational model, and normalize it; produce a test plan to be applied to the instance that you create.
4. Produce the SQL that creates the database instance, populate the database and provide a browser-screen shot of each table.

Note: the schema must include at least 10 data tables but not more than 20. Linking tables are not included in this count.

Query Requirements

Once your scientific questions have been formulated they must be translated to SQL in order to retrieve the data and carry out any functions/filters that are involved.

1. Your presentation queries must include at least 4 that require a 2-table join (or the equivalent if you use nested queries, for example) and 1 that requires a 3-table join.
2. As noted above, your queries must be distinct from each other, with changes beyond a table name or attribute name.
3. Testing: You are required to develop a testing plan in one homework assignment. Include that plan and its outcomes in an Appendix to your project report.

Presentation

1. You have 15 minutes to cover it all. You must have slides to cover the introduction and discussion. You must have a live demo ready to show.
2. Ahead of time: check that your laptop plugs into the podium computer and that slides display properly. Check that your live demo works and that none of the queries take so long to process that you will not have enough time. If they are taking to long to demo you need to clear them with Dr. Carr before the presentation.
3. Demo Preparation: one functional database, preloaded on a laptop or server, be ready to cut and paste the queries from a text file.
4. Context Preparation: 5-10 slides (submitted to Dr. Carr the day of your presentation) that cover the following:
    a. Background: scientific context, current data repositories with their strengths and weaknesses.
    b. Introduction: what are the 5 questions, why are they important.
    c. Methods: logical model and decisions made when transforming into the schema, data sources, upload and validation method
    d. Results: run the queries live, show the results – you can also provide summaries or graphs if they inform the outcome.
    e. Discussion: why are the results significant?
5. The class will assess each presentation, delivery and style will count along with content, and ability to answer 2-3 audience questions.

Note: Please see the Presentation Assessment Rubric – the class will use this form to provide peer evaluations.

Project Report

Your report must adhere to professional standards of writing for both style and format.

1. Due April 28th, 5pm, to Dr. Carr, submit via Email.
2. 8-10 pages (single spaced), written and formatted for a journal like the Oxford 'Database' (see http://database.oxfordjournals.org/) journal.
3. You must have:
    a. Title/Author page
    b. Abstract
    c. Introduction and Background
    d. Materials and Methods
    e. Results
    f. Discussion and Conclusions
    g. References (at least 10)
    h. Appendices – diagrams, schemas, test plans as noted above.
    i. Supplementary Materials (on a CD: data, scripts, SQL, etc.)
4. You should give me enough material that I can replicate your database and your results.
   Note: Please see the Rubric for Project Report for points granted on aspects of the report.