Bioinformatics Resources
Exercise Sheet 4
due on June 5th, 9 a.m.
Total number of points: 15 P

Task 1
Set up a local instance of the MYSQL Server on your private computer. If you have access to a server elsewhere that is fine as well, you should not need administrator privileges for this exercise. The server version is irrelevant, if you are on Linux, install whatever is in your distribution’s repositories. Perform all the following tasks using the mysql client on the commandline, don’t use GUIs. Give every SQL statement you use, not just the results. First, create a database called pfam. (1P)

Task 2
Pfam is available for download as an SQL database. Both the .sql.gz and .txt.gz files were created with the commandline tool mysqldump. Familiarize yourself with its functionality, download the files for the pdb table (see links on homepage) and restore it in your database instance. Give the commands you used to do so. (1P)

Task 3
How can you list all tables currently present in the database? How can you get more information about the definition of the columns in table pdb? (1P)

Task 4
Give SQL statements for the following questions (6P):

What is the pdb_id with the highest resolution
How many entries have a resolution of less than 0.8. The result should be the number, not all rows.
From all entries with a resolution smaller than 0.7, which are the PDB IDs of the 3 entries with a resolution closes to 0.7Å.
How many different experimental methods are present in column method
How many structures involve Wayne Hendrickson (Hendrickson, W.A.) as an author
How many structures involve Wayne Hendrickson as an author and Zhang, Z. as the first author

Task 5
5a) Analogous to the table pdb also load the tables pdb_pfamA_reg and pfamA into your database. List all PDB IDs and chain identifiers which are part of the pfam family 'G-alpha'. First, identify the relevant column in pfamA, then get the information about the pdb_ids and chains. (3P)
5b) Use MYSQL's EXPLAIN statement to check how your query is optimized. What is the order in which the query is executed. Explain why the steps are performed in the given order by the DBMS. Hint: Identify the most expensive operation(s) in your query, then see what MYSQL is doing with it, to mitigate runtime (3P).

Good luck,
J. Reeb & L. Richter