

## Übungen zur Vorlesung „Molekularbiologische Datenbanken“

### Aufgabenblatt 3: Data Load

---

Abgabetermin: **03.06.2003 bis 17:00 Uhr** über Goya oder in RUD25 IV.102

Bitte geben Sie die SQL Queries und deren Ergebnisse  
im PDF-Format oder auf Papier ab.

Maximal: 20 Punkte, Erforderlich: 15 Punkte ab.  
Namen nicht vergessen!

---

Leider senden uns die Mitarbeiter in den Laboren die bereits erworbenen Daten nicht in XML sondern in CSV. Die Aufgabe besteht nun darin, diese Daten in unsere Datenbank zu schreiben. Die Dateien enthalten jeweils eine Header-Zeile, in der die Feldnamen angegeben sind, und darunter die Daten.

Angaben zu den einzelnen Dateien:

#### Platform

Array-Submitter.csv

Angaben über Array Submitter

Array.csv

Alle bis jetzt genutzten Arrays, wobei die ID = GPLxx ist.

GPLxx.csv

Jede Datei steht für einen Array. In diesen Dateien ist angegeben, welche Proben auf dem Array sind und noch die zusätzlichen Daten über GenBank-Accession und GeneOntology.

#### Samples

Organisation.csv

Daten über die beteiligten Labors

Submitter.csv

Daten über den Submitter von Samples und Meßergebnissen

Sample.csv

Daten über die Samples, wobei ID = GSMxxxx.csv ist.

Sample-Protokoll.csv

Welche Probe ist welchem Protokoll behandelt worden

GSMxxxx.csv

Jede Datei steht für ein Sample. In diesen Dateien sind die ID der Probe und die Meßwerte für jede Probe auf dem Array angegeben.

#### Protokolle

Protokolle.csv

Daten über Protokoll-ID und die dazugehörigen Angaben.

Aufgabe:

### SQL-Queries

Schreiben Sie SQL-Queries für folgende Fragestellungen:

1. Bestimmen Sie von jeder Tabelle die Anzahl der Einträge (count(\*)) und geben Sie diese an. **1 P**
2. Es kam eine Anfrage von Stephen Welle, welche Samples er schon an uns übermittelt hat. **1 P**
3. Er wollte auch wissen, welche Proben-ID's mit der GeneOntology – Nummer GO:6997 (nuclear organization and biogenesis) für den biologischen Prozeß in der Datenbank schon gespeichert sind und auf welchem Array sie sich befinden. **2 P**
4. Ein Mitarbeiter eines Labores will wissen, wieviele Proben auf jedem unserer Arrays zu finden sind. **2 P**
5. Das Gen für den oncostatin M receptor hat die GenBank Accession-Number U60805. Einer unserer Partner möchte nun wissen, welche Proben davon existieren und auf welchem Array diese sind. **2 P**
6. Dieses Gen ist auf dem Chromosom 5 lokalisiert. Er möchte nun wissen, welche Gene ebenfalls an der Stelle 5p13.1 sind. Er will neben den GenBank Accession-Numbers auch noch vorhandene Proben mit dem dazugehörigen Array wissen. **2 P**
7. Er will auch noch wissen, auf welchen Arrays die drei Gene mit den GenBank Accession-Numbers D26561, L15409 und M58600 gemeinsam vertreten sind. ? **2 P**
8. Für welche Samples und Arrays wurde Protokoll P-EMBL-11 für das labeling verwendet? **1 P**
9. Von Sample GSM3719 sind einige Average Difference Intensities (Value) über 120.000. Von diesen würden wir gerne wissen, welche molekulare Funktion diese haben. **2 P**
10. Microarrays können in Serien erstellt werden, um beispielsweise eine Entwicklung über das Wachstum von Zellen zu dokumentieren. Serie GSE80 hat 2 Samples in der Datenbank bereits gespeichert. Welche Werte sind 10-fach erhöht von Sample GSM2393 zu GSM2391? **3 P**
11. Mitarbeiter eines Labors haben erwähnt, dass sie von dem Web-interface gerne die Information hätten, welcher Submitter welche Samples bereits an die Datenbank übermittelt hat. Um die Informationen in einer Table zu vereinigen, erstellen wir einen View mit dem Namen des Submitters, dem Namen der Organisation, der Sample-ID, dem Titel des Samples und des dafür verwendeten Arrays. **2 P**