

Lecture 3 – MySQL and other languages, NoSQL, online databases, APIs & version control

Extra exercises

NoSQL

Exercises

- Transform these SQL statements to mongo statements
 - `INSERT INTO users(user_id, age, status) VALUES ("bcd001", 45, "A");`
 - `SELECT * FROM users WHERE status != "A";`
 - `SELECT * FROM users WHERE age > 25;`
 - `SELECT * FROM users WHERE status = "A" ORDER BY user_id DESC;`
 - `SELECT COUNT(*) FROM users;`
 - `SELECT COUNT(user_id) FROM users;`
 - `SELECT DISTINCT(status) FROM users;`
 - `SELECT * FROM users LIMIT 1;`
 - `SELECT * FROM users LIMIT 5 SKIP 10;`
 - `UPDATE users SET age = age + 3 WHERE status = "A";`
 - `DELETE FROM users WHERE status = "D";`

NoSQL

Exercises

- Import the protein coding genes from file protCodingGenes.bson
\$ mongorestore --collection protCodingGenes --db test
protCodingGenes.bson
- Return the first 10 genes (alphabetically) on chromosome 22
- Return the last but one group of 10 genes (by position) on chromosome 12
- Return the number of unique gene names
- Return the 50 most common genes and their number of occurrences
- Return a sorted list of the number of genes per chromosome

Databases in bioinformatics

Exercices – part 2

- How many nucleotide sequences for collagen genes from nematode worms are there in the NCBI Database?
- How many mRNA sequences for collagen genes from nematode worms are there in the NCBI Database?
- How many protein sequences for collagen proteins from nematode worms are there in the NCBI database?
- What is the accession number for the *Trypanosoma cruzi* genome in NCBI?
- How many fully sequenced nematode worm species are represented in the NCBI Genome database?
- Find the accession number of human beta-globin mRNA sequence. What is the accession number of the encoded protein ? How many amino acids does it contain?

Databases in bioinformatics

Exercices – part 3

- Find the publication in PubMed with the following ID: 8663200
 - Use the search record of this publication in PubMed to obtain this data entry in the Nucleotide database
 - Download the GenBank formatted flatfile
- Find the corresponding record from the previous exercise in the ENA database
 - Compare the EMBL format to the GenBank format you downloaded before
- How many alternative transcripts are known for *Drosophila melanogaster* eIF-4E
- Find the human hemoglobin alpha protein in UniprotKB. What is the entry name?
- How many genes are associated with Huntington Disease (HD), with Alzheimer's disease (AD) and with Parkinson's disease (PD)?
- How many transcripts does Ensembl predict for the human gene ACHE?
- Find the mouse orthologue of the human SSBP4. Does this gene have paralogues?