

Exercise 1: SRS

The Sequence Retrieval System is a database system that works with flat-files. In addition, many bioinformatics tools are incorporated and can be combined with the databases searches.

We will have a look at two SRS servers:

<http://srs.bioinformatics.nl/> (running SRS Release 7.1.3.1)

<http://srs.ebi.ac.uk/> (running SRS Release 7.1.3.2)

Many servers exist that have their own databases and SRS look. See “Public SRS servers worldwide” on the EBI SRS home page.

- 1) Go the SRS servers linked above and have a look at the homepages.
- 2) Check the available databases by of the servers by clicking on the “Databanks” tab (SRS@EBI). Which tab you’d have to click on to get this info from SRS@bioinformatics? You can see that the available databases differ per SRS host. Which main categories of databases are present?
- 3) On the Library Page/Select Databanks you have an overview of available database groups. Also this differs among sites.

We will continue with SRS@EBI. You can search in three ways:

- 1) use the “Quick Text Search” on the home page/ “Quick Search” page.
- 2) use the “Quick Search” on the library page.
- 3) use the “Query Form” for extended searches.

Searching for a protein

We are interested in pyruvate kinase, a member of the phosphor-transferase family. Search for these proteins.

- 1) select a database on the “Library Page” (UniprotKB)
- 2) use the “Query Form” and search “Description” field for “pyruvate kinase”. How many entries did you retrieve?
- 3) have a look at the different Display Options. Check the “*Complete Entries* view and “FastaSeqs”. Complete entries in this case shows the plain text SwissProt entry. What is the sequence length and where do you find this in the complete entry?

Linking to related information

SRS allow us to link from databases entries to related information in other databases.

- 1) Select the first entry from the previous search (KPYA_RICCO) and click on the “Link” button on the left.
- 2) You will see the Library Page where you have to select a database to link to.
 - 2a) Choose GOA and Search. What is GOA? Here you can see all GO(A) annotations linked to this pyruvate kinase.
 - 2b) Alternatively, choose InterPro from the “Protein function, structure and interaction databases” category and search. The InterPro database (Integrated Resource of Protein Domains and Functional Sites) is an

integrated documentation resource for protein families, domains and functional sites.

- 3) You will see four entries; select one. Have a look at the record and follow the “Overview” link at the bottom. You see the location of different domains, present in InterPro, on different proteins. When you now click on the protein name, you get a detailed view. If you click on the domain name, you will see information on the domain, such as a domain description, GO terms, references and proteins that have this domain.

Linking from proteins to DNA

Many databases in SRS are linked with others. We will have a look at the protein ↔ DNA links.

- 1) Search for the protein: “104K_THEAN”. How many entries do you get?
- 2) We wish to retrieve the corresponding DNA record. SRS allows this via “Link”. Link to the DNA. Which database should you select?
- 3) How many DNA entries did you find?
- 4) Now we can link back to a protein databases. Select UniParc.
- 5) How many proteins did you retrieve?
- 6) Why did you find more than one protein?

Exercise 2: SRS Tools

An example of a multi-domain protein is pyruvate kinase.

You have quite some options to search:

- Quick search proteins
- Library page, select database to search
- Query Form
- BookMarkLets links on some SRS pages (lower left)

- 1) Search on “pyruvate kinase” or retrieve your previous results (“Results” tab).
- 2) Launch an analysis tool of your choice. Or select Iep to calculate the iso-electric point (only the first entry is taken).
- 3) Select a sequence and BLAST this sequence. The results of SRS tools are stored in databases. This means you can use these results in your database queries. For example, you can link the BLAST results to SwissProt.
- 4) We wish to retrieve the corresponding DNA record. SRS allows this via “Link”. Link to the DNA. Which database should you select?
- 5) How many DNA entries did you find?

Exercise 3: SRS Multiple alignment

This exercise demonstrates the use of multiple alignments with a simple example (from Arthur. Lesk). Multiple alignments can be used to find conserved regions or domains and are used as input for phylogeny programs. One has to be sure that the sequences are related. In this case, the sequences are indeed homologues.

Determine from the sequences of pancreatic ribonuclease from horse (*Equus caballus*), minke whale (*Balaenoptera acutorostrata*) and red kangaroo (*Macropus rufus*), which two of these species are most closely related. What do you think and what do the sequences tell you?

1. Go to srs.ebi.ac.uk and retrieve the sequences.
2. Go to “Results Options” and select ClustalW and launch the application.
3. The results will be available soon. Meaning of symbols: “*” position same in all sequences; “:” or “.” indicate positions at which residues are very similar or somewhat similar. You see that the sequences are very similar. With pair-wise comparison you can count the number of identical residues. Which pair has the most identical residues?
4. Which species are most closely related? Is this what you expect?

The SRS server at EBI also has a database containing multiple alignments of nucleotide and protein sequences. It was developed since the amount of multiple alignments submitted to the EMBL database increased considerably. In which group is this database placed and how is it called? (publication: <http://bioinformatics.oxfordjournals.org/cgi/reprint/18/5/763.pdf>).

Exercise 4: KEGG pathway database (www.genome.ad.jp/kegg/)

KEGG, the Kyoto Encyclopedia of Genes and Genomes, provides, for example, pathway information.

1. Click KEGG2, ‘KEGG Table of Contents.’
2. Click ‘KEGG PATHWAY.’
3. Have a look at the pathways and click ‘fatty acid metabolism.’ (under 1.3: Lipid metabolism).
4. Select ‘Homo sapiens’ and click ‘Go.’
5. All green coloured enzymes are organism specific. Clicking these will give extra information. KEGG pathways are linked to human diseases in OMIM.
6. KEGG also contains a list of human diseases with links to the metabolic maps. (this is linked from the KEGG PATHWAY page (<http://www.genome.ad.jp/kegg/pathway.html>)). In this case the enzyme causing the disease is clearly indicated. Select e.g. “type I diabetes mellitus” to see a metabolic map. Which proteins are implicated in diabetes I and II?

Exercise 5: Retrieving sequences from GenBank

1. Go to NCBI’s Entrez and select Nucleotide. (www.ncbi.nih.gov/)
2. Retrieve a sequence (just type a query). Alternatively, retrieve the RefSeq entry for APOE mRNA ([NM_000041](http://www.ncbi.nlm.nih.gov/RefSeq/record/NM_000041)) and browse through the record.
3. Look at the record using the different Display views. For example, Graph, GenPept, FASTA, Summary, GI list. The latter is useful if one needs the accession numbers of a large number of records.
4. Also click the ‘Links’ link. Here you can go to PubMed, OMIM, SNPs etc. The links available depends on the entry. Can you think which kind of accession may have the most links?
5. Keyword searches generally return a large number of records. In case you are looking for a specific sequence, you could build a specific query. Click Limits

and check out the possibilities. Limits only present a few limitations. Follow Preview/Index. Here you see two things: your three most recent queries and a section to add terms to your query. From the drop-down menu ('All Fields') you can select different terms. If you click Index you see the terms within a field.

6. Build an extended query with any gene or extract only the *mRNAs* of the *human APOE gene* with a precise query and write down this query.
7. If you are only interested in a specific *gene* you could also search a different database. Which?
8. Search this database and look at the info (for human only). The Reference into Function (GeneRIF) provides a 'short' overview of what is known about the gene in literature. What disease(s) is APOE connected with?