EMBL-EBI Tutorial

Jennifer McDowall (Dec 2011)

Searching the Non-redundant Patent Sequence Database

This tutorial provides an introduction to EBI resources and the different sequence search methods available for doing prior art searches for patent and non-patent sequences. Exercises are provided to help you practice what you learn in this tutorial.

Contents:

Text Searching

There are several ways to query the EBI website for patent documents and sequence. The simplest method is to use the **EB-eye** tool at the top of every EBI webpage. EB-eye is a full-featured text search that provides very fast access to the EBI's data resources, allowing the user to search globally across all EBI databases. Therefore, you can enter a patent publication number, a sequence accession number or a general text term such as 'kinase' to perform an EBI-wide search that will list all the databases that have information relating to your query. These databases cover a wide range of biological and chemical areas, including:

- Nucleotide sequences
- Protein sequences
- Genomes
- Macromolecular structures
- Small molecules
- Gene expression
- Molecular interactions
- Reactions pathways
- Protein families
- Enzymes
- Literature
- Ontologies

For more complex searching, **SRS** (http://www.ebi.ac.uk/srs/) allows you to select specific databases from a list of over a hundred. Once you have selected the database(s) you wish to search, you can create tailored text searches using a series of drop-down boxes that restricts the search to specific fields. In this way, you can cross-reference multiple databases at once with complex queries that are easily built.

You can also create a text search in the **ENA Browser** (http://www.ebi.ac.uk/ena/) to search for nucleotide sequences, and in the **UniProt Browser** (http://www.uniprot.org/) to search for protein sequences. With these browsers, text searches will provide a list of relevant entries, while a search using a specific sequence accession number will take you directly to that entry.

Note: SRS is the only text search facility available for the non-redundant patent sequence databases.

Exercise 1:

SRS Search

- **Navigate to EBI's SRS page at http://www.ebi.ac.uk/srs/.**

- **Select the header tab '*Library Page*' at the top of the page.**

You now have a list of all the databases it is possible to search using SRS.

- **Scroll down to the section called '*Other protein sequence databases*', and expand it by clicking on the '+' icon at the side.**

We are going to look at two databases:

- Non-redundant patent protein database level-1
  - Each entry contains protein sequences that are 100% identical
- Non-redundant patent protein database level-2
  - Each entry contains protein sequences that are 100% identical AND occur in the same patent family (patent equivalents)

- **Click on the linked words '*Patent Protein NRL1*'.**

This page provides a summary of the database, including the total number of entries in the database. Remember, this is the total number of non-redundant patent protein sequences.

- **Take note of the total number of entries for the NRL1 Database.**

- **Use the *back-button* to return to the previous page**

- **Now click on the linked words '*Patent Protein NRL2*'.**

- **Take note of the total number of entries for the NRL2 Database.**

? **Does NRL1 or NRL2 have more entries?**

- **Use the *back-button* to return to the previous page**

- **Select the boxes for <u>both</u> non-redundant patent databases: '*Patent Protein NRL1*' and '*Patent Protein NRL2*'.**

You now able to query both databases at once.

- **Select the header tab '*Query Form*'.**

You have up to 4 different query boxes that enable you to easily refine your search to 4 search criteria of your choosing, allowing easily tailored searching.

- **Click on the drop-down menu for the first box.**

You are able to query specifically on any of these fields.

- **Go back to the header tab '*Library Page*'.**

- **Select the database '*UniProtKB*' under the section '*UniProt Universal Protein Resources*' AND re-select the patent databases '*Patent Protein NRL1*' and '*Patent Protein NRL2*' under the section '*Other protein sequence databases*'** (Note: your previous selections are usually wiped when you return to the library page).

- **Select the header tab '*Query Form*'.**

- **Click on the drop-down menu for the first box.**

Note that the drop-down menu has changed. It now only contains the fields that are common to these 3 databases.

- **Leave the search field at '*All Text*'.**

- **Search for the patent publication number 'EP0242329'.**

- **Press the '*Search*' button to submit your request.**

You should now have a list of entries related to this patent document in these 3 databases.

Let's look at the non-redundant databases first.

- **Select the entry for the '*Patent Protein NRL1*' database.**

? **How many identical patent sequences are contained in this entry?**
? **Which patent has the earliest publication date?**

- **Select the accession number '*EPOP:A00210*', which is the one associated with our query patent, namely EP0242329.**

From here you can gain access to the patent literature.

- **Scroll down to the '*References*' section.**

Note that you have a selection of databases in which to view the patent document, including (1) in SRS, (2) in the EPO Esp@cenet, (3) in CiteXplore and (4) in Patent Lens.

- **Select the link to the patent publication in CiteXplore.**

Here you have the patent abstract and another link to the full patent in Esp@cenet.

- **Use the *back-button* (3x) to return to the SRS results page.**

- **Select the entry for the '*Patent Protein NRL2*' database.**

The level-2 database groups sequences by patent families.

? **What is the Priority Date for this patent sequence?**
? **Is it earlier than the publication date you recorded from the level-1 database?**

- **Select the 'Family' number (26290626).**

This provides a list of all the patent equivalents. Note the additional (different) sequences in this family.

- **Use the *back-button* (2x) to return to the SRS results page.**

- **Select the entry for the '*UniProtKB*' database.**

UniProt knowledgebase contains non-patent sequences. The database is subdivided into TrEMBL entries that receive automatic annotation, and SwissProt entries that receive high-quality manual curation. This entry is SwissProt (this information is under 'General Information').

- **Scroll down to the 'Reference' section.**

? **Can you find the patent reference for EP0242329?**

This patent document was used to annotate this protein.

- **Scroll back up to the '*General Information*' section.**

? **What is the date this entry was created?**
? **Was it before or after the patent publication date?**

Note that the dates of any sequence updates and annotation updates are also recorded.

****

Sequence Searching

There are several ways to search sequences at the EBI. Both the **ENA Browser** (http://www.ebi.ac.uk/ena/) and the **UniProt Browser** (http://www.uniprot.org/) have their own sequence search engines – these are basic search tools that are restricted to searching only these databases using default settings.

You can also use the **Similarity & Homology Tools** from EBI webservices (http://www.ebi.ac.uk/Tools/sss/), where you have a choice of search algorithms, including NCBI-BLAST, WU-BLAST, FASTA, SSEARCH and iterative searches (PSI-BLAST and PSI-SEARCH). In addition, you have the choice of using default settings or tailoring the search to best suit your search criteria.

Tips for sequence searching

1) *Use the tool that best fits your needs.*
In general, use the faster BLAST searches for querying large databases (e.g. UniParc), the slightly slower but more sensitive FASTA for medium-sized databases (e.g. UniProt/SwissProt or non-redundant patent protein databases) and the slow but accurate SSEARCH for small databases (e.g. taxonomic division of a UniProt/SwissProt).

2) *Wherever possible, search at the protein level rather than at the nucleotide level.*
Because of codon degeneracy, where multiple codons can encode for the same amino acid, homology can be more difficult to detect. For instance, two proteins could have the identical amino acid sequence, but vary considerably in their nucleotide sequence. Because of this, the % identity between two genes is usually lower than between their corresponding proteins, and therefore, their e-values will be less significant (i.e. higher).

3) *Search the smallest database that is likely to contain the sequence(s) of interest.*
E-values reflect database size. The larger the database, the less significant the match, the higher the e-value. Searching too large a database can result in losing good matches when their e-values go above the threshold.

4) *Use e-values over % identity or % similarity.*
E-values provide a statistical measure of the significance of a match, and are an estimate of the number of matches one can expect to see by chance. For instance, an e-value of 2 means that in a database of the current size, one would expect to see two matches with a similar score purely by chance.

5) *Consider using different gap penalties and scoring matrices.*
To restrict the search to the closest relatives, use high gap penalties and a strict matrix (decrease ratio of match/mismatch score for nucleotides; use a high BLOSOM or low PAM for proteins). To search for more divergent relatives, use lower gap penalties and more lenient matrices (increase ratio of match/mismatch score for nucleotides; use a low BLOSOM or high PAM for proteins). For very short query sequences, use high gap penalties and strict matrices to force conserved matches; you may need to increase the e-value threshold is you fail to get matches.

6) *Use a filter to remove low complexity regions.*
CA repeats, poly-A tails or proline-rich regions can give spuriously high scores that reflect compositional bias rather than significant matches. Use the appropriate filter to mask the sequence before you search. However, be careful not to mask what you are looking for!

Exercise 2:

Sequence Search

- **Keep the tab with your SRS results as you will need to copy the sequence from the NRL2 entry.**

- **Open a new second tab in your browser and navigate to the '*Similarity & Homology Tools*' page at http://www.ebi.ac.uk/Tools/sss/.**

Note that you can also go to the EBI home page (http://www.ebi.ac.uk/) and navigate using the link to '*Sequence Similarity & Analysis*' under the section entitles '*Data Resources & tools*'.

- **Select 'FASTA' and 'Protein'.**

- **STEP 1:**
  - **Uncheck the protein database '*UniProt Knowledgebase*' that is set as default.**
  - **Check the database '*UniProtKB/Swiss-Prot*'** (smaller to search than all of UniProtKB, therefore faster for this exercise)
  - **Scroll down the list of databases and expand the section '*Patents*'.**
  - **Check the two databases '*NR Patent Proteins Level-1*' and '*NR Patent Proteins Level-2*'.**

You should now have 3 databases selected for Step 1.

- **From your SRS results tab, select the '*Patent Protein NRL2*' entry and copy the entire sequence (including the header):**

>NRPL2:NRP001827B8 PN:EP0242329 A2
MVVSEVDIAKADPAAASHPLLLNGDATVAQKNPGSVAENNLCSQYEEKVRPCIDLI
DSLRALGVEQDLALPAIAVIGDQSSGKSSVLEALSGVALPRGSGIVTRCPLVLKLKKL
VNEDKWRGKV

- **Return to the sequence search tab.**

- **STEP 2:**
  - **Paste the protein sequence (exactly as shown above) that you copied from the NRL2 entry (from the SRS results) into the sequence search form.**

Note that when searching, there is a limit of one sequence per search, even if you upload a sequence file.

- **STEP 3:**
  - **Leave the parameters set at default.**

- **STEP 4:**
  - **Submit your search request.**

You now have a list of patent (from NRPL1 and NRPL2) and non-patent (from UniProt/SwissProt) sequences that show similarity to sequence A00210 in EP0242329.

- **Take a look at the matches.**

? **How many matches show 100% identity to our sequence?**
? **How many are from the non-redundant patent sequence databases?**

Exercise #1 told us that in the NR level-1 database there were only three identical sequences.

? **Can you find the entries that correspond to those we found in exercise #1?**
? **How do you explain the remaining sequences with 100% identity?**

- **Take a look at the column labelled 'Length'.**

Note that our sequence is 124 residues long, which corresponds to the first 4 matches. The remaining matches from the non-redundant database are all longer than 124 residues, ranging from 508-682 residues. Therefore, our sequence shows a 100% identity to a *small region of these proteins*. This is why it is important to check what region matches the query sequence and not to rely on % identity.

? **Can you explain the variation in e-values for these other sequences showing 100% identity?**

NOTE: the lower the e-value, the better the match. E-values are a better measure of how good the match is over % identity.

- **Take a look at sequence 22, 'SP:MX1_HUMAN'.**

This is a match from the UniProt/SwissProt database of a non-patent sequence.

- **Look at some of the annotation links available for this protein.**

There is information on Gene Expression, Nucleotide Sequences, Genomes, Ontologies...and much more.

- **At the top of the page, click on the '*Function Predictions*' button.**

The resulting page shows the protein domain and family matches for the protein matches from the UniProt database (*Note:* the results from the non-redundant patent sequence database results are not shown). The *function predictions* provide information on the types of domains these proteins contain. This tool is valuable for providing additional annotation, as well as for helping to prioritize results and weed out poor or false matches.

This is the end of our short practical