EMBL-EBI

# A Short Trip around the EBI

www.ebi.ac.uk

EMBL-EBI

# Sequence Retrieval System

The Sequence Retrieval System (SRS) at the EBI is an implementation of the Software developed originally by EMBL and EBI, further developed by LION Biosciences AG and now owned by BioWisdom. It is made freely available to anyone who wishes to use it either through the EBI web portal, or as part of the EBI Webservices. SRS is also available at many sites around the world.

As its name suggests, the system allows retrieval of sequences and other information from any database linked to the system. At the EBI there are currently approximately 200 databases, 50 of which have been created, or are curated and maintained at the EBI. The SRS system offers a text searching interface using keywords to retrieve specific sets of information.

## Information Retrieval

There is one main nucleic acid sequence database and one main protein sequence database in widespread general use among the biological community. The nucleotide database is EMBL[1] (or GenBank or DDBJ) and the protein database is UniProt[2].

As these databases contain hundreds of thousands of sequences, searching through them requires the processing power of a computer search engine and the **Sequence Retrieval System (SRS)** has been designed to do just that at the EBI and many other places around the world.

### Nucleotide Sequence Retrieval

In order to investigate a gene, often the nucleotide sequence is a good starting point.

> Go to http://srs.ebi.ac.uk and click on the "Library Page " tab. Select the "EMBL" database by clicking in the box to the left of the database name. A green tick should appear.
>
> Proceed to "Standard Query Form" and use the first field to type in "**Cystic Fibrosis**". Type "**human**" in the second field and alter the pull down menu to read "Organism Name".

Each pull down menu in SRS is representative of a particular data field in the chosen database. In this case, we have chosen the EMBL nucleotide database and all the information in it (approx. 61 million entries) is stored, amongst other things, as flatfiles. Each flatfile is a text file and each piece of data is located on a separate line in the file, preceded by a two letter tag indicative of the type of data in that field. The default option in the SRS system for these menus is "All Text" which means that the computer will search the entire EMBL entry. By reducing the second search to "Organism Name" the computer will restrict its search in each entry to that field only.

The search produces over 11,000 results which is far too many to try and look through. In order to reduce the number of hits, there are a variety of options. We could return to the query form and add further search criteria, or we could look at a different database. As cystic fibrosis will have a multitude of hits, we will change to a different database for this example.

---

[1] Nucleic acid sequences may be submitted to any of the three databases, depending on whether you are resident in Europe (EMBL), America (Genbank) or Asia (DDBJ – DNA DataBase of Japan). Newly submitted sequences are transferred between the three databases on a daily basis.

[2] This database was released in December 2003 and is the amalgamation of SwissProt, TrEMBL and PIR to form a non-redundant set of all known protein sequences.

> Return to the "Library Page" and expand the "Gene Dictionaries and Ontologies" section. Select "Entrezgene".

This is an NCBI database focusing on those genes that have been completely sequenced and have an active research community. As this rules out many genes, we should get fewer hits in our results than we did searching through the EMBL database.

As we expect to find considerably fewer entries, we can use the "Quick Search " option on the SRS pages.

> Type "**cystic fibrosis human**" into the "Quick Search" field at the top of the library page and hit the yellow "Quick Search" button.

There are 77 resulting entries and it is obvious from the first one that they are not all from human. This is because the quick search field searches the entire entry and as it only searches for keywords, whatever context the word "human" appears in, it will add it to the list of possible hits.

Even though 77 entries is much reduced from our EMBL results it is still too many and so using the "Standard Query Form" to refine our search further is appropriate.

> Return to the "Library Page", ensure that the Entrezgene database is selected and proceed to the "Standard Query Form". Type "**Cystic Fibrosis**" into the top search field and select the "Description" menu option. The second field menu option should be "Organism" and type "**Human**" into the corresponding field.

This time only 3 entries result – much easier to go through! Look at the "Locus" information and only the first gene suggests itself as the entire gene sequence for cystic fibrosis. On the right of the results table are the "Reference_Acc" numbers – or the accession numbers of each database entry that as been involved in collating the information represented in Entrezgene. Each entry has an accession number which is unique both to the entry and the particular database it is in. Those starting NM_ (or any other letter instead of M) belong to the RefSeq database - and NCBI collection of reference sequence information taken from their own and other sequencing projects. It is data from this database that forms the basis for the human and mouse genome builds in genome browsers such as Ensembl. Those numbers starting with two or three letters belong to the EMBL/Genbank/DDBJ collaboration of nucleotide databases. The long list here suggests why we had so many results returned when searching EMBL.

> Follow the "Entrezgene 1080" link to see the information held in this database.
>
> Now use the yellow "link" button on the left hand side and select the UniProtKB/SwissProt database from the resulting library page. Hit the yellow search button to retrieve results.

## UniProt – the Universal Protein Resource

The UniProt knowledgebase is the universal protein resource, resulting from an amalgamation of the SwissProt, TrEMBL and PIR databases and funded by grants from both the EU and National Institute of Health in the USA. SwissProt is a database designed to house curated protein sequences. These are sequences that have either been submitted directly from the lab which sequenced them, or translated from open reading frames and corroborated against information in the literature. As curation is a slow process (depending on the entry it can take from a couple of hours, to several weeks), the TrEMBL database was devised to automatically translate and annotate open reading frames from the EMBL database. The PIR database was a separate protein database held in USA. Any PIR entries

concurrent with current SwissProt entries were merged and the rest were put into the UniProt/TrEMBL part of the database to be annotated at a later date.

> Follow the link on the left hands side of the results table to the single protein result and access the information contained in the SwissProt entry.

Each entry in the UniProt database has both an entry name and unique accession number. The entry name immediately defines the protein as being from the Swissprot (i.e. curated) part or the TrEMBL (i.e. automated) part of the database. SwissProt entries have a name made up from the name of the gene, bound by an underscore to the name of the organism. Generally if the organism is a common mammal, the name will be in English. For less common organisms, the first three letters of the genus followed by the first two letters of the species are used. Thus ***Dro****sophila* ***me****lanogaster* becomes DROME.

The unique accession number is often a letter followed by five digits. This number should always be used in conjunction with the same entry. If several TrEMBL entries are amalgamated into a single SwissProt entry, their original accession numbers will be kept as secondary accession numbers.

> Make a note of the accession number for this protein.

> Scroll down to the database cross references section and note down the PDB identification number. Do not include the top one as this is a theoretical model and has not been created using experimental data.

This section cross references the CFTR protein to several databases. Some are held at the EBI and others are not. We will be visiting the intact and interpro databases later on, so if you want to note down the accession numbers now you can do.

> Scroll down to the "Features" section.

This lists all the information known about the protein. It has been annotated by the curators and includes all the data from the various references cited in the section above. The first few features suggest potential topological and transmembrane domains across the protein. Below these are specific domains, nucleotide phosphate binding, modified residues and carbohydrate binding regions. Other proteins may have the same or different categories of features.

Further down the feature list is the "variant". These have been taken from several literature sources and define variance information that different scientists have discovered whilst studying this protein.

Finally, as there is structural information, the residues relating to the protein secondary structure are also noted.

> Scroll down to the variants at position 508 in the protein. This position is occupied by a phenylalanine residue and deletion of this residue results in cystic fibrosis in 72% of cases.

## BLAST – Sequence Homology Searching

In order to retrieve other proteins that are similar to the human product of the cystic fibrosis gene BLAST can be used through SRS.

> Scroll to the top of the protein entry and notice the "Entry Options" on the left hand side bar. The "analysis tool" option is already set to "BlastP" which is the program we want. As the protein entry is already open, the SRS program knows which sequence to base it's search on.

The default database to search is the "SwissProt (Release)" database. Both UniProtKB and EMBL have two databases, the "release" database and the "updates" database. As both of these databases are huge it would be impossible to add new entries on a daily basis and manipulate the data in such a way that is searchable for pieces of software such as SRS (a process known as indexing). Thus new entries are collated in a new, "updates" database and then amalgamated with the main "Release" database to form a new "Release". For EMBL a new release occurs approximately every three months, for UniProtKB this is approximately every two weeks.

Hit the yellow "Launch" button underneath the BlastP selection. Leave the options as default and hit the yellow "Launch" button at the top right of the screen.

Click on the blue "results" hyperlink to proceed to the results page. Check the status of this job. If there is an egg timer still in view, wait a couple of minutes and refresh the page. Repeat this until the egg timer is replaced by a green tick.

Click on this tick to see the first 50 results of the search (one of the parameters we accepted before launching the program).

The first hit is of course our CFTR_Human protein. BLAST works by breaking the query sequence down into smaller "words" of approximately 3 adjacent residues in length. Each possible word from the sequence is then indexed using a scoring matrix and that score then compared to all other scores achieved by a three residue block of sequence. All these blocks are then used to search the database for sequences that match – with particular preference given to the high scoring blocks. AS our highest scoring blocks were obtained from the CFTR_Human sequence, it follows that a match to this sequence should be the highest scoring, and thus appears first in our list of results.

The second hit is a CFTR protein from *Pan troglodytes* – or the chimpanzee. This is the closest organism to *Homo sapiens* and thus this hit is not surprising either.

There are many hits with an "E" or "Expect" value of zero. The "E" value is an indication of the quality of the alignment and thus the likelihood of a true positive hit. It represents the probability of finding an alignment of the length and quality of the hit in a random BLAST of a sequence and database with the same composition as was used to obtain the hit. Thus an "E" value of zero suggests there would be no random matches.

We can use SRS to extract the hits that have an "E" value of zero so we can align them.

Return to the "Library Page" tab in the SRS interface and explode the "Application Results Data" tab. Select the "BlastP" option by checking the box to the right hand of it and proceed to the "Extended Query Form".

Select the "E Value" and type "0" into both boxes to ensure that we collect data just from the hits displaying an "E" value of zero. Hit the yellow "Search" button at the top right of the page.

## ClustalW – Multiple Sequence Alignment

This time instead of the 50 entries that were listed by default, only 24 proteins have been returned. These proteins can be multiply aligned using a program called ClustalW. This program is relatively old now, and whilst it is still relatively accurate much in use, newer algorithms have been developed to more accurately reflect the clustering process that goes into creating a multiple alignment and thus represent evolutionarily more distant sequences.

In the "Apply Options to" field, ensure that the application will apply to "unselected results only" and alter the "Launch analysis tool" menu to "ClustalW" and hit the

yellow "Launch" button. Accept the default parameters and hit the "Launch" button again. Retrieve the results as you did for the BLAST program.

All 24 sequences are aligned against each other in blocks wrapped to fit the page. Whilst this view groups all sequences together, it does not give a residue count so finding residue 508 could be tricky.

Alter the "View" menu at the top right of the page to "ClustalWAli" and hit the yellow "view" button.

A table appears with the multiple alignment coloured. Sadly the residues are still not numbered and the colouring makes it hard to pick out our F508 residue.

To the far right hand side of the table is a brown "Jalview" button. Click this to open a java alignment viewer. Scroll along the alignment until position 508 is displayed. Note that this is conserved across all 24 proteins.

## Interpro - Protein Domain Database

Whilst a residue which is fully conserved across all proteins tends to be particularly important in the protein, you will have noted that a sizeable majority of the residues are fully or almost fully conserved. To investigate the location of this residue further, we can use the Interpro database. A database of protein domains using information from a variety of different domain databases.

Go to http://www.ebi.ac.uk/interpro and type "**CFTR**" into the search box and hit the "Search Interpro" button.

Three entries are found using this search term.

Click on each of these in turn and look at the "Type" of protein they represent.

Only one of the entries represents a domain, whilst the other two entries represent a family of proteins.

Follow the link back to the IPR001140 domain entry and scroll down the information.

There is a variety of information on the ABC transporter domain, including what it is used for in the body and a variety of other protein families that include this domain. The taxonomic coverage indicates the number of proteins in UniProKBt/SwissProt that also incorporate this domain as part of the protein.

Scroll back to the top of the entry and select ""For all matching proteins" from the "Table" option in the "Matches" toolbar. Alter the "select" menu to "Protein accession(s) and type in the accession number of the CFTR_Human protein in the field next to the menu. Click the "view" button at the bottom of this section.

The table view of our protein indicates the domains involved in that protein and where they occur in terms of residues. The first domain in the table is the entry we were looking at, namely IPR001140 the ABC transporter. There are two such domains in this protein, but neither of them cover the F508 residue, so it is not this domain that gets damaged when the phenylalanine residue is missing.

Scroll across the table and make notes of any domains that include the residue we are interested in.

There are two domains, IPR003439 (also found as a Pfam entry PF00005) which spans the region 451 – 622 (and residues 423 – 646 is Prosite PS 50893) and IPR003593 (also found in the SMART database SM00382) and spanning residues 450 – 639.

Following the links to any of these entries will display further information on them. However, we can also display a graphical view of these domains.

> Return to the original page for IPR001140 and again in the "Matches" toolbar follow the link to "sort by AC" in the "Detailed" section.
>
> Once again select "protein accession(s)", type in the relevant number and hit the "view" button.

This time a graphical view of each of the domains appears. By mousing over each domain it is possible to see which residues in the protein they span. In addition, entries in other databases are also much more obvious than they were in the table view.

> Follow the link to the appropriate domains. What are they? How does this relate back to a possible effect of a missing residue at position 508 in the protein?

The domains cover an ABC transporter like protein domain, and an AAA+ ATPase domain. The description found in the UniProt entry suggests that perhaps it is the disruption of the ABC transporter that causes the defect.

## Ensembl Genome Browser

There are many ways to look at the same data, and the EBI offers this a variety of different resources treat the same genes an proteins in a different way. Whereas EMBL and UniProt are designed to record individual sequences and their annotations as found in the literature, Ensembl takes its data from the appropriate project to annotate the entire genome. In the case of Homo sapiens this data is taken from the NCBI build and is the same set of data as used by the UCSC Genome Browser (at University California Santa Cruz) and the NCBI MapViewer.

Ensembl takes the raw data, and the re-predicts the locations of genes along the genome. This information is then corroborated with protein evidence taken from a variety of protein databases.

> Go to http://www.ensembl.org and follow the link on the home page to the *Homo sapiens* genome.

This is similar to the homepages for each of the metazoan genomes represented in Ensembl. The karyogram of the chromosomes is visible on the left hand side and this is linked to the relevant area of the chromosome in a more detailed ContigView within the Ensembl genome browser. On the right there is a brief history of the genome, and below any new additions to the data and the genome statistics.

> Use the search box in the top right hand corner of the page to search for "**CFTR**". Alter the search terms to "Gene" to restrict the search.

The search produced 11 hits all relating to different genes with CFTR in their description somewhere.

> Use the knowledge acquired above to select the correct Ensembl gene entry for the human CFTR gene. Follow this link to see the GeneView.

Each Ensembl webpage is called a "View" prefixed by the thing that is being displayed. Thus this page displays information on the gene and is known as "GeneView". Information on the

contig would be known as "ContigView", on the Exon as "ExonView" and so on. In addition, each of these features has a unique Ensembl accession number. It starts with ENS followed by a letter denoting the feature (thus Gene would be "G", Exon would be "E", protein would be "P") and eleven digits.

Scroll down the information to see where the gene is located in the genome and how many exons it has.

Move further down to the orthologue predictions and use the "align" link at the side of each of them to confirm the presence of the important F508 residue. (There are no numbers here, but each block is 60 residues in length).

Scroll back to "Transcripts" and follow the link to "Exon Info".

The exon information details all 27 exons involved in the CFTR transcript. The length of each exon is recorded, along with its start and end point within the chromosome. The purple coloured text indicates sequence within the untranslated region and black text indicates sequence that will eventually be translated to protein.

The sequence is once again wrapped in blocks of 60 bases and the information here indicates that the first Methionine residue in the protein is at bases 133-136.

Use this information to work out which exon contains the phenylalanine residue at position 508 in the protein by adding up the lengths of the subsequent exons as far as necessary[3].

You should see the TTT codon for the phenylalanine at position 132 of the exon.

Return now back to the "GeneView" and this time follow the top link in the "Genomic Location" section towards the top of the screen.

The overview demonstrates the CFTR gene within its own neighbourhood on chromosome 7. The detailed view displays all 27 exons extended over the area of genome. The transcript can be seen to be on the forward strand of the genome (as depicted by Ensembl) and extend across two contigs (the blue lines in the middle). The numbers on these contigs refer to EMBL entries and the sequences could be accessed in this manner.

Scroll down to the detailed view and use the mouse to rubberband[4] the vertical block that represents exon 11 – the exon you should have discovered contains the F508 residue. Follow the link on the subsequent menu that offers to "Zoom into this region in detailed view".

Use the features menu at the top of detailed view to turn on the SNPs track (if it is not already). To do this, check the box to the left of the SNP option, close the menu using the option at the bottom of the list and wait for the display to be redrawn.

There are two non-synonymous coding SNPs that fall within exon 11 and are displayed as yellow vertical blocks in the SNP track.

Click on one of these SNPs and follow the menu link to SNP properties.

We see that it refers to our F508 residue and labels it as an in-del. Ensembl takes genotype frequency data from the HapMap (www.hapmap.org) project via the dbSNP database and

[3] If this involves thinking too much, return to the GeneView page and follow the link to protein info. Ensure that the exons are coloured on the peptide sequence display and count the ones to the F508 residue.

[4] Drag the mouse over the area you wish to zoom into. You will see a red box appear around the chosen area.

displays it for relevant SNPs. In this case, however, this particular polymorphism has not been genotyped, so we are unable to look at potential diversity across human ethnic groups.

> Return to the GeneView – either by going back in your browser, or by following the link on the SNPView page to Genomic location. This will take you back to the ContigView. Once here, click on the red exon in the "detailed view" and follow the link to the "Gene".
>
> Scroll down to the "Gene DAS Report" section and note the entries within ArrayExpress. Explode this option by clicking on the plus sign.

ArrayExpress is the data repository for microarray experiments that conform to the standard MIAME requirements (www.mged.org). There are seven curated experiments, although none directly relate to cystic fibrosis.

## Macromolecular Structure Database

As was apparent in the UniProt entry for the CFTR protein, there have been several protein structures determined for parts of this protein. The Macromolecular Structure Database at the EBI takes the three dimensional, experimentally determined structures from the Protein Data Bank and puts them into a biological context using curation and assessment of

> Go to http://www.ebi.ac.uk/msd and follow the "msdlite" link from the "Services" menu to the right of the page.

All five PDB files listed in the UniProt entry have had their structures solved using X-ray crystallography and cover the same region which includes the part of the protein containing the F508 residue, although there are discrepancies in the number of chains involved in each of the structures, suggesting that a variety of crystals were produced using this protein.

There is no way to assess the quality of these entries other than to look at them one by one in the database.

> If you still have the UniProt page open, you can do this by following the "EBI" link to the right of each PDF cross-reference (don't forget the first one is a model and can be ignored). If not, then type each one of the PDB identification codes (one number followed by three digits) into he MSDLite search box and look at both the "Resolution" and the "Oligomeric State" of each protein.

The resolution is a measure of how close the x ray data allowed the scientists to get to the actual protein structure. It is measured in Angstroms, where one Angstrom (Å) is equivalent to $10^{-10}$ metres. Thus as the number of angstroms goes down, the resolution goes up. Hydrogen bonds spanning residues within a protein are approximately 2.5 – 3.5 Å in length, thus an ideal resolution for proteins would be 2.5Å or lower.

All of our structures would be of good enough resolution quality, but only the top one seems to display the five chains potentially necessary for this protein to function *in vivo*. This would be ideal for our search, except that the title of the paper indicates that the F508 residue has been replaced with an alanine residue.

> Return to MSD Lite and enter the PDB code 1XMI in the ID code field and hit the "Start search" button.
>
> Follow the "Assembly" link on the left hand side of the page to see the pentamer.
>
> Follow the "Sequence" link on the left hand side to compare the sequence used in the structure with that found in the UniProt database.

Check out the sequence at the 508 position and see that indeed the phenylalanine has been replaced. Unfortunately we want to see where the bonds are for the F508 residue and what is likely to be disrupted by its deletion, so this structure is not a good one to pursue further at this stage.

> Return to MSD Lite and enter 2BBO into the search field. Look once again at the "sequence" field to ensure that the phenylalanine residue is this time present.
>
> Follow the "visualisation" link on the left hand side and use the AstexViewer to "view" the protein structure.

Astex is a company in Cambridge that produces structural viewing software. It allows the EBI access to this viewer – and through the EBI to anyone who uses the MSD portal.

The structure display can be rotated using the left hand mouse button and can be manipulated either by using the menu of on the left hand toolbar, or by right clicking on the display to reveal a further menu of options.
The display is also linked to the sequence at the bottom of the viewer.

> Scroll along the sequence to the equivalent of the F508 residue. (Remember this protein only displays residues 388 – 638 so use the "Sequence" page to work out that the Phenylalanine residue will be at position 89 in this sequence).
>
> Once the residue has been located, click on it in the sequence and the display will zoom in to that residue.

Display as ribbons and point out location of phenylalanine in the structure.