Tutorial section

Entrez: Making use of its power

Entrez[1] is a data retrieval system developed by the National Center for Biotechnology Information (NCBI) that provides integrated access to a wide range of data domains, including literature, nucleotide and protein sequences, complete genomes, three-dimensional structures, and more. Entrez includes powerful search features that retrieve not only the exact search results but also related records within a data domain that might not be retrieved otherwise and associated records across data domains. These features enable us to gather previously disparate pieces of an information puzzle for a topic of interest.

Effective and powerful use of Entrez requires an understanding of the available data domains, the variety of data sources and types within each domain, and Entrez's advanced search features.

This tutorial uses the human *MLH1* gene, implicated in colon cancer, to demonstrate the wide variety of information that we can rapidly gather for a single gene. The numbers noted in the search results will of course change over time as the databases grow. The same techniques shown here can be used for any topic of interest.

The search goals are to:

- separate the wheat from the chaff – identifying a representative, well-annotated mRNA sequence record;
- retrieve associated literature and protein records;
- identify conserved domains within the protein;
- identify similar proteins;
- identify known mutations within the gene or protein;
- find a resolved three-dimensional structure for the protein or, in its absence, identify structures with homologous sequence;
- view genomic context and download the sequence region.

SEPARATE THE WHEAT FROM THE CHAFF

An Entrez data domain usually encompasses data from several different source databases. The goal is to identify a representative, well-annotated mRNA sequence record among the many available in the Entrez Nucleotide data domain.

The Entrez Nucleotide domain includes sequence records from the archival GenBank database, the curated RefSeq[2] database, nucleotide sequences extracted from Protein Data Bank (PDB)[3] records, and a new Third-Party Annotation (TPA) database. As a result, an unrefined search can retrieve records of varying quality (in both sequence and annotation), and there can be a high degree of redundancy in search results, depending upon how many labs have submitted sequence data for a gene or its fragments.

For example, an unqualified search of Entrez Nucleotide for **colon cancer** currently retrieves >10,000 hits. The results include archival and curated records, characterised sequences and lower-quality sequences such as expressed sequence tags (ESTs), contigs from the genome project and more.

A 'Limits' option allows us to restrict our search, if desired, to a specific data subset, such as the curated, non-redundant RefSeq database. It also allows us to limit searches to specific data fields, retrieve records with certain attributes, such as molecule type, and exclude sometimes unwanted records such as ESTs, which are typically numerous and of lower sequence and annotation quality than characterised genes.

In this case, if we use the Limits page to restrict our 'colon cancer' search to the Title field and then only to records from RefSeq, our retrieval narrows to 31 hits. If we then do a new search for human in the Organism field and use the 'History' option to combine the two searches with a Boolean AND, we retrieve 13 hits – far fewer and far more specific results than our original >10,000.

In addition, because each RefSeq record presents an encapsulation of the knowledge about a single gene or splice variant, rather than the work of an individual laboratory, each hit is similar to a review article.

For this example, we will more closely examine NM_000249: *Homo sapiens* mutL homologue 1 (*MLH1*), and the additional information we can retrieve for that gene in Entrez. Of course, a search for *MLH1*, rather than colon cancer, would have worked as well, and the same techniques could have been used to narrow the search results. Gene symbol searching, however, can sometimes be less reliable if a gene has been known by numerous aliases. Although curated RefSeq records include the official gene symbol as well as the aliases, archival records, such as those in GenBank, include only the gene symbol that the authors used at the time of submission or last update.

TRAVERSE THE DATA DOMAINS

The Links menu for each record allows us to retrieve directly associated records from the other Entrez data domains. For example, the PubMed[4] link for NM_000249 retrieves the 12 references cited in the RefSeq record. They represent a set of articles selected by curators (if a record is in a 'Reviewed' rather than 'Provisional' state) that discuss salient research on the gene, such as mapping, characterisation and phenotype.

Returning to the nucleotide record for NM_000249, we can just as easily traverse from the nucleotide data domain to protein. By simply selecting 'Protein' from the 'Links' menu box, we can view the corresponding amino acid sequence record. We will only see the record for NP_000240, which contains the sequence that was extracted from the Features/CDS translation field of NM_000249. Additional, similar protein sequences that were identified by the BLAST[5] algorithm can be retrieved by following the 'Related Sequences' link.

Similarly, the Links menu for NM_000249 lists all other Entrez data domains that contain associated information and can be used to easily access that additional data.

RETRIEVE RELATED RECORDS

Links from one Entrez data domain to another provide access only to data that are directly related to our original record of interest. However, the 'Related Records' option within most Entrez data domains allows us to instantly broaden retrieval to other relevant records in that domain that would not otherwise have been retrieved by the original query. For example, when viewing the 12 PubMed records above, the 'Display: Related Articles' option instantly retrieves hundreds of other PubMed records that were identified using a word weight algorithm, which finds records with similar words in their titles, abstracts, and Medical Subject Headings (MeSH).

Similarly, the display for protein record NP_000240 includes a link to 'Related Sequences' that were identified using the BLAST algorithm. The related sequences are shown in decreasing order of similarity to the original sequence and can provide valuable insights into the possible function of the original sequence if it has not yet been characterised.

In the Entrez Protein domain, the BLink (short for BLAST Link) option provides a graphical overview of the top 200 similar sequences, showing the regions of similarity to the original sequence of interest. BLink also provides great flexibility in filtering and customising the view of the complete set of similar sequences (not just the top 200). It allows us, for example, to see the best hit from each organism, only the hits that have associated 3D structure records, a phylogenetic tree of our hits (in which we can choose to exclude organisms or organism groups), and more.

IDENTIFY CONSERVED DOMAINS

Conserved domains, like similar sequences, can shed light on a protein's function as well as its organisation. Each protein sequence in Entrez has been compared against NCBI's Conserved Domain Database (CDD).[6]

Returning to the original protein record for NP_000240, we can follow the 'Domains' link to view the conserved domains that have been identified in the sequence. This traverses to the CDD and, if 'Details' are viewed, shows the presence of the HATPase and DNA mismatch repair domains. In addition, the grey 'MUTL' bar represents the protein family with which NP_000240 is associated. Clicking on the graphic for any domain or protein family leads to more detailed information. The 'Show Domain Relatives' option retrieves protein sequences with similar domain architectures identified by the Conserved Domain Architecture Retrieval Tool (CDART).[7]

IDENTIFY KNOWN MUTATIONS

Variations within the human *MLH1* gene can be identified through the 'SNP' link or through the 'Allelic Variants' section of the corresponding Online Mendelian Inheritance in Man (OMIM)[8] record. The SNP link will retrieve records for variations submitted by individual labs to dbSNP[9] and aligned to the corresponding mRNA using the BLAST algorithm. A graphic summary for each SNP indicates whether it is in a locus region, transcript or coding region and gives additional information about mapping consistency, heterozygosity, validation status and more.

An OMIM record, on the other hand, describes (if available) allelic variants that have been reported in the literature and summarised by the OMIM editorial staff. For example, one interesting mutation reported in MIM entry number 120436 is allelic variant .0011 (Gly67Trp), in which the smallest amino acid has been substituted by the largest amino acid. A corresponding structure record, as described in the next section, can shed light on the possible significance of that substitution.

FIND THREE-DIMENSIONAL STRUCTURES

As noted by Mullan,[10] finding a resolved structure for a protein is the exception rather than the rule. This is true because the currently available $>$2.7 million protein sequence records far exceeds the available number of individual structure records, currently $\sim$20,250 in Entrez's Molecular Modeling Database (MMDB).[11] However, the presence of a homologous structure can assist in the analysis of protein function.

The 'Links' menu for NP_000240 does not include 'Structure', indicating that this sequence record is not directly associated with a 3D protein structure record. Several options exist to find possible homologous structures:

- retrieve the approximately 600 related sequences for NP_000240 and then display the 'Structure Links' for the complete set;

- use BLink to graphically view the related sequences and then view only the subset that has 3-D structures; and

- use the BLAST system to compare the NP_000240 protein sequence against all the protein sequences from PDB.

In this case, all three options retrieve the same set of six structures, although retrieval can sometimes vary because of the differences in the three systems. For example, BLAST might retrieve additional sequences, depending on the cutoff score used. BLink, on the other hand, might retrieve fewer sequences because it uses a non-redundant set of proteins, and it shows only the top 200 hits. We will use the first option in this example.

The first three structure links (1B62, 1BKN, 1B63) are from *Escherichia coli*, and the last three (1H7S, 1H7U, 1EA6) are from human. The latter were deposited by the Guarne[12] lab and represent a free protein, a protein bound to ATPγS, and a protein bound to ADP, respectively. For this example, we will look at 1H7U to see what we might be able to discern from that structure about the sequence in NP_000240.

If the Cn3D[13] program is already installed on the computer, the 'View Structure' button will automatically open Cn3D. One window will display the 3D structure of 1H7U, and a second window will display the corresponding protein sequences for protein chains A and B (referred to as 1H7U_A and 1H7U_B, respectively). From here, Cn3D offers a wide range of features that enable us to label residues, zoom in or out, render the structure in different styles, colour the structure by various features, import and align a protein sequence from Entrez Proteins, and more.

In this example, use the 'Style' menu to render the structure as 'tubes' and change the colouring shortcut to 'domains'. The resulting pink and blue regions of 1H7U_A represent the compact 3D domains that in this case correspond closely to the HATPase and DNA mismatch repair domains, respectively. The brown and green regions represent the same domains in 1H7U_B. These colours correspond to the graphic summary of 1H7U in the Entrez Structure database. The Cn3D sequence alignment window also now colours the residues in 1H7U_A and 1H7U_B by domains.

Because we are interested in the relationship between the protein sequence in 1H7U and that in NP_000240, we can now import NP_000240 (gi 4557757) and align it to 1H7U_A. That is the protein chain identified by BLAST and BLink as being similar to NP_000240. The steps to import and align NP_000240 are provided in Table 1. Now, we can see

Table 1: Steps to import and align NP_000240 with 1H7U_A

While viewing 1H7U in Cn3D 4.1:

- In the Sequence/Alignment Viewer window, select the menu item 'Imports/Show Imports'. This will cause the Import Viewer window to appear.
- In the Import Viewer window, select the menu item 'Edit/Import Sequences'.
- In the Select Chain dialogue box, select 1H7U_A and click OK.
- In the Select Import Source dialogue box, select 'Network via GI/Accession' and click OK.
- In the Input Identifier dialogue box, enter the accession NP_000240 and click OK. The new sequence will appear in the Import Viewer window.
- Select 'Algorithms/BLAST single' and, using the crosshair, click anywhere on the sequence for NP_000240 to align it to 1H7U_A using the BLAST algorithm.
- To make the alignment appear in the Sequence/Alignment Viewer window, select the menu item 'Alignments/Merge All' in the Import Viewer window.
- The alignment should now appear in the Sequence/Alignment Viewer window, and the colouring scheme changes to show the aligned residues in red. Dismiss the Import Viewer window, if desired.
- Reset the 'Style/coloring shortcut' in the structure window to 'domains', and set the mouse mode in the Sequence/Alignment Viewer window to 'select rectangle'.

the high degree of sequence alignment between NP_000240 and the pink-coloured residues of the HATPase domain in 1H7U_A.

Given this alignment, how might the observed Gly67Trp substitution in NP_000240 affect its structure, based on the view of the homologous structure? In the sequence alignment window, mouse over the residues of NP_000240 until the grey footer bar shows 'gi 4557757, loc 67' (Glycine). Click on the corresponding Glycine residue in 1H7U_A (loc 74) to highlight it. In the structure window, use the left mouse button to spin the 3D structure until you can clearly see and identify the highlighted residue. Is it possibly in the active site? For example, is it within 5 Å of the ATPγS molecule? To find out, remove the highlighting from residue #74 of 1H7U_A by clicking on any residue in NP_000240 in the sequence alignment window. Going back to the structure window, double click on the Mg-complexed ATPγS to highlight it. Then use the menu bar option called 'Show/Hide|Select By Distance|Residues Only' to highlight all residues within 5 Å(or other desired distance) of the ATPγS. Indeed, the Glycine at position #74 is within 5 Å and is likely part of the active site for this energy-producing domain. This hints at the possible problems a Gly → Trp mutation might cause at that position.

VIEW GENOMIC CONTEXT AND DOWNLOAD THE SEQUENCE REGION

To further study the *MLH1* gene, it may be useful to identify the chromosome region that contains the gene, download the corresponding genomic sequence data, and order the relevant clones. The Map Viewer[14] link for NP_000240 can provide such information. It leads to a graphical view of the appropriate chromosome region. The 'seq' link for MLH1 allows us to download the genomic sequence data for that gene and to adjust the region to download upstream and/or downstream data, if desired. The Maps & Options dialogue box can then be used to add the Component map, which shows the GenBank records used to assemble that chromosome region. The GenBank records contain corresponding clone source information, which can be used to order those clones from their distributors for further study.

CONCLUSION

The techniques shown in this tutorial can be used to gather pieces of the information puzzle for any topic of interest. The types and quantities of information will vary by gene and organism, and will lead to other Entrez data domains in addition to those explored here. As the existing data domains grow and as new ones are added, Entrez will continue to provide a single access point to previously disparate data.

Renata C. Geer,
National Center for Biotechnology Information,
National Library of Medicine,
8600 Rockville Pike, Bldg 38A,
Bethesda, MD 20894, USA
Tel: +1 301 435 5947
Fax: +1 301 480 9241
E-mail: renata@ncbi.nlm.nih.gov
and
Eric W. Sayers,
The KEVRIC Co., Inc.,
8484 Georgia Ave,
Silver Spring, MD 20910, USA
Tel: +1 301 402 4039
Fax: +1 301 480 9241
E-mail: sayers@ncbi.nlm.nih.gov

References

1. Ostell, J. (2002), 'The Entrez search and retrieval system', in 'The NCBI Handbook' [Internet], National Library of Medicine (US), National Center for Biotechnology Information, Bethesda, MD, Chapter 14 (URL: http://ncbi.nlm.nih.gov/entrez/query.fcgi?db=Books).

2. Pruitt, K. D., Tatusova, T. and Maglott, D. R. (2003), 'NCBI Reference Sequence project: Update and current status', *Nucleic Acids Res.*, Vol. 31(1), pp. 34–37.

3. Westbrook, J., Feng, Z., Chen, L. *et al.* (2003), 'The Protein Data Bank and structural genomics', *Nucleic Acids Res.*, Vol. 31(1), pp. 489–491.

4. Canese, K., Jentsch, J. and Myers, C. (2002), 'PubMed: The Bibliographic Database', in 'The NCBI Handbook' [Internet], National Library of Medicine (US), National Center for Biotechnology Information, Bethesda, MD, Chapter 2 (URL: http://ncbi.nlm.nih.gov/entrez/query.fcgi?db=Books).

5. Altschul, S. F., Madden, T. L., Schäffer, A. A. *et al.* (1997), 'Gapped BLAST and PSI-BLAST: A new generation of protein database search programs', *Nucleic Acids Res.*, Vol. 25(17), pp. 3389–3402.

6. Marchler-Bauer, A., Anderson, J. B., DeWeese-Scott, C. *et al.* (2003), 'CDD: a curated Entrez database of conserved domain alignments', *Nucleic Acids Res.*, Vol. 31(1), pp. 383–387.

7. Geer, L. Y., Domrachev, M., Lipman, D. J. and Bryant, S. H. (2002), 'CDART: protein homology by domain architecture. Conserved Domain Architecture Retrieval Tool', *Genome Res.*, Vol. 12(10), pp. 1619–1623.

8. Hamosh, A., Scott, A. F., Amberger, J. *et al.*(2002), 'Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders', *Nucleic Acids Res.*, Vol. 30(1), pp. 52–55.

9. Sherry, S. T., Ward, M. H., Kholodov, M. *et al.* (2001), 'dbSNP: The NCBI database of genetic variation', *Nucleic Acids Res.*, Vol. 29(1), pp. 308–311.

10. Mullan, L. J. (2002), 'Protein 3D structural data – where it is, and why we need it', *Brief Bioinform.*, Vol. 3(4), pp. 410–412.

11. Chen, J., Anderson, J. B., DeWeese-Scott, C. *et al.* (2003), 'MMDB: Entrez's 3D-structure database', *Nucleic Acids Res.*, Vol. 31(1), pp. 474–477.

12. Guarne, A., Junop, M. S. and Yang, W. (2001), 'Structure and function of the N-terminal 40 kDa fragment of human PMS2: A monomeric GHL ATPase', *EMBO J.*, Vol. 20(19), pp. 5521–5531.

13. Wang, Y., Geer, L. Y., Chappey, C., *et al.* (2000), 'Cn3D: sequence and structure views for Entrez', *Trends Biochem Sci.*, Vol. 25(6), pp. 300–302.

14. Dombrowski, S. M. and Maglott, D. (2002), 'Using the Map Viewer to explore genomes', in 'The NCBI Handbook' [Internet], National Library of Medicine (US), National Center for Biotechnology Information, Bethesda, MD, Chapter 19 (URL: http://ncbi.nlm.nih.gov/entrez/query.fcgi?db=Books).