Sample User Question

I am studying the regulation of cancer genes and would like to retrieve all human sequence records associated with cancer that contain a promoter region.

Comments / Analysis

Sequence records can be annotated with a wide variety of biological features. The GenBank sample record, http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html shows some examples and points to several sources that provide a list of features that can be annotated on records.

In this case, the researcher is interested in the upstream regions of genes that regulate their expression (i.e., that turn the genes on and off). This exercise demonstrates how to retrieve records annotated with a promoter feature. (See important caveat about biological feature annotations under additional tips.)

Step By Step Guide

Two approaches of answering the user's question are shown below; both use the Preview/Index page to build your query one search term at a time. Approach A begins in the **Entrez Nucleotide** database while Approach B begins in the **Entrez Gene** database. Note the approaches are complementary, not redundant -- see comments about the net difference for details. The searches are also shown as a complex Boolean queries in additional tips.

Approach A:

Open **Entrez Nucleotide** - retrieve human nucleotide sequence records that contain the term "cancer" and that have a promoter feature annotated on them.

1. select the **Preview/Index** option beneath the search box.

At the bottom of the Preview/Index page:

2. select the **Organism** field from the pop-up menu and enter **human** in the text box next to the search field menu.

Then press the **AND** button to add that term and search field to the active query at the top of the page

3. select the **Text Word** field from the pop-up menu and enter **cancer** in the adjacent text box.

Press the **AND** button to add that term to the active query.

4. select the **Feature Key** field from the pop-up menu and enter **promoter** in the adjacent text box.

Press the **AND** button to add that term to the active query.

5. press **Go**

Adapted from: NCBI Advanced Workshop for Bioinformatics Information Specialists
http://www.ncbi.nlm.nih.gov/Class/NAWBIS/index.html

Notes: Because the Text Word field in selected in step 3 actually represents a number of text-containing fields, the term cancer might appear in various contexts in the records that our search retrieves. For example, it might be in the definition line (title) of a sequence record, in the title of a published or unpublished article cited in the references field, in an author's affiliation information, etc.

You can try to limit your search for cancer to the Title field instead of the Text Word field. However, that will miss potentially relevant records that only have gene names in the definition line and have the term "cancer" in relevant contexts in other fields of the record.

Approach B:

Open **Entrez Gene** - retrieve human loci associated with cancer, then retrieve the associated nucleotide sequence records and limit to those that are annotated with a promoter feature.

I. Retrieve Entrez Gene records for human loci that are associated with cancer:

1. select the **Preview/Index** option beneath the search box.

At the bottom of the Preview/Index page:

2. select the **Organism** field from the pop-up menu and enter **human** in the text box next to the search field menu.

Then press the **AND** button to add that term and search field to the active query at the top of the page

3. select the **Text Word** field from the pop-up menu and enter **cancer** in the adjacent text box.

Press the **AND** button to add that term to the active query.

4. select the **Filter** field from the pop-up menu and enter **gene_nucleotide** in the adjacent text box. (This limits your search results to records in the Entrez Gene database that have links to records in the Entrez Nucleotide database.)

Press the **AND** button to add that term to the active query.

5. press **Go**

II. Retrieve the associated sequence records from Entrez Nucleotide and limit the set to only those records which have a promoter feature annotated on them

1. using the **Display** options near the top of the Entrez Gene search results page, select **Nucleotide Links** from the pop-up menu and press the **Display button**. Entrez then displays the associated sequence records in the Nucleotide database.

2. On the Entrez Nucleotide page, select the **History** option.

3. In the text box at the top of the History page, type the number of the search that represents the nucleotide links for Entrez Gene and followed by a Boolean AND plus the term "promoter[fkey]". For example, the text box would contain the query: **#2 AND promoter[fkey]**

4. press **Go**

Adapted from: NCBI Advanced Workshop for Bioinformatics Information Specialists
http://www.ncbi.nlm.nih.gov/Class/NAWBIS/index.html

What was the net difference between the two approaches? Were they redundant or complementary?

To see which additional records you retrieved, go to the Entrez Nucleotide History page and create a Boolean query that will show the difference of what you retrieved in Approach A but not in Approach B, and vice versa. For example, let's say the **Entrez Nucleotide History** page shows:

Search	Most Recent Queries	Result
#5	Search #3 NOT #1 (unique hits from Approach B: Entrez Gene to nucleotide)	329
#4	Search #1 NOT #3 (unique hits from Approach A: straight to Entrez nucleotide search)	214
#3	Search #2 AND promoter[fkey]	380
#2	Nucleotide Links for Gene (Search human[Organism] AND cancer[Text Word] AND gene_nucleotide[Filter]) Approach B: Entrez Gene search then follow links to nucleotide	65604
#1	Search human[Organism] AND cancer[Text Word] AND promoter[Feature key] Approach A: Entrez Nucleotide search	265

Search #1 is the result of Approach A. Search #3 is the result of Approach B. To find out what records are retrieved by B but not A, enter the following query at the top of the Entrez Nucleotide History page: **#3 NOT #1**.

Additional Tips

Complex Boolean query

The search can be done in a single step by entering the search as a complex Boolean query. For example, the Boolean query for the Entrez Nucleotide search (approach #A, steps 2-4) shown in the step-by-step guide is:

human[orgn] AND cancer[word] AND promoter[fkey]

The Boolean query for the Entrez Gene search (approach #B, steps 2-4) shown in the step-by-step guide is:

human[orgn] AND cancer[word] AND gene_nucleotide[filter]

Important caveat about biological feature annotations

It is very important to note that the biological annotations present in GenBank records are those that **were provided by the submitter of the sequence** record. In some cases, it is possible that a feature is present on a sequence, but that the submitter did not did not annotate it or did not yet identify it. In the latter case, additional biological features are identified through further study of the gene, at which time the submitter might update their record with the newly discovered information.

Records in curated databases, such as RefSeq, oftain contain additional biological annotations that were not present in the source GenBank records on which they are based. However, they, too, can only reflect what is currently known about a sequence, and will be enhanced over time with new information as molecular biology research continues to progress.

The study of gene regulatory regions, for example, is an active area of research and the promoter region for many genes has not yet been identified. However, if users are interested in a particular gene or genes, they can use **Map Viewer** to download the genomic sequence for a gene of interest plus any desired number of bases upstream. They can then study the upstream region in an attempt to identify the promoter.

Synonymous biological features

Some biological features can be annotated in various ways by submitters. For example, if a user is looking for human sequence records that contain **intron/exon splice junctions**, they might want to consider the following query:

> **(CDS[fkey] OR mRNA[fkey] OR exon[fkey] OR intron[fkey]) AND biomol_genomic[prop] AND human[orgn]**

This is because some submitters have annotated introns and/or exons explicitly, while others have just annotated a CDS or an mRNA feature on genomic sequence. In the latter case, the CDS or mRNA base spans (e.g., CDS join 1087..1200,1633..1758, etc.) imply the intron/exon junctions.

(For genomes that are represented in Map Viewer, intron/exon splice junctions can also be viewed graphically on the Genes_sequence map.)