GenBank Introduction

一月 26, 2013, 3:04 a.m. by Rosalind Team

Topics: Bioinformatics Tools, Sequence Analysis

> ### The World's Biggest Genetic Database
>
> The most comprehensive database available for molecular biologists is **GenBank**, an open access resource that contains an annotated collection of all publicly available sequenced DNA and its translation into proteins. GenBank was founded by NCBI in 1982, and over the last three decades, the data it houses has grown exponentially, doubling every 18 months. As of August 2012, GenBank contained over 143 **billion** nucleobases.
>
> Every sequence has a unique GenBank identifier that will directly retrieve its full sequence record. Here are some examples of database IDs:
>
> ```
> CAA79696
> NP_778203
> 263191547
> BC043443
> NM_002020
> ```
>
> You can also search by submitter/author name in the format `last_name first_initials` (e.g., `Smith JR`).
>
> To search for an exact match, enclose it in quotation marks.
>
> ```
> "contactin associated protein"
> "duchenne muscular dystrophy"
> ```
>
> You can also restrict your search by using Boolean operators (AND, OR, NOT) as well as use specific subsets of records.

Problem

GenBank comprises several subdivisions:

- Nucleotide: a collection of nucleic acid sequences from several sources.
- Genome Survey Sequence (GSS): uncharacterized short genomic sequences.
- Expressed Sequence Tags, (EST): uncharacterized short cDNA sequences.

Searching the Nucleotide database with general text queries will produce the most relevant results. You can also use a simple query based on protein name, gene name or gene symbol.

To limit your search to only certain kinds of records, you can search using GenBank's Limits page or alternatively use the `Filter your results` field to select categories of records after a search.

If you cannot find what you are searching for, check how the database interpreted your query by investigating the `Search details` field on the right side of the page. This field automatically translates your search into standard keywords.

For example, if you search for `Drosophila`, the `Search details` field will contain (`Drosophila[All Fields]`), and you will obtain all entries that mention *Drosophila* (including all its endosymbionts). You can restrict your search to only organisms belonging to the *Drosophila* genus by using a search tag and searching for `Drosophila[Organism]`.

Given: A genus name, followed by two dates in YYYY/M/D format.

Return: The number of Nucleotide GenBank entries for the given genus that were published between the dates specified.

Sample Dataset

```
Anthoxanthum
2003/7/25
2005/12/27
```

Sample Output

```
7
```

Programming Shortcut

NCBI's databases, such as PubMed, GenBank, GEO, and many others, can be accessed via Entrez, a data retrieval system offered by NCBI. For direct access to Entrez, you can use Biopython’s `Bio.Entrez` module.

The `Bio.Entrez.esearch()` function will search any of the NCBI databases. This function takes the following arguments:

- `db` : The database to search. For example, this field can be `nucleotide` for GenBank or `pubmed` for PubMed.
- `term`: The search term for the "Query" field. You can use search tags here.

We will now demonstrate a quick search for the rbcL gene in corn (*Zea mays*):

```
>>>from Bio import Entrez
>>>Entrez.email = "your_name@your_mail_server.com"
>>>handle = Entrez.esearch(db="nucleotide", term='"Zea mays"[Organism] AND rbcL[Gene]')
>>>record = Entrez.read(handle)
>>>record["Count"]
'6' # Surely this value will change over time because GenBank is constantly updated
```

Note that when you request Entrez databases you must obey NCBI's requirements:

- For any series of more than 100 requests, access the database on the weekend or outside peak times in the US.
- Make no more than three requests every second.
- Fill in the Entrez.email field so that NCBI can contact you if there is a problem.
- Be sensible with your usage levels; if you want to download whole mammalian genomes, use NCBI's FTP.

Congratulations You solved this problem (attempt #2).