



Getting Coding Sequences for Proteins

A workflow to extract DNA coding sequences for specific protein families

National Center for Biotechnology Information • National Library of Medicine • National Institutes of Health • Department of Health and Human Services

Overview of Challenges

Sequence analysis is an important approach in elucidating and understanding of the biological function of genes and their evolutionary relationship. However, analysis at the protein sequence level may not provide enough resolution to allow separation of closely related organisms, biologists often need to analyze the coding sequences (CDS) of the proteins of interest. Since biological processes through which proteins are synthesized differ between eukaryotes and prokaryotes, sequence records available from NCBI, accurately reflecting the biological processes, also differ.

In sections below, we will describe the workflows needed to retrieve CDS for a family of proteins from eukaryotic organisms and prokaryotic organism, respectively. Both workflows require the use of the Entrez Direct package (EDirect), which operates within the NCBI's Entrez Programming Utilities (EUtils) framework. For more information see:

EUtils Help Manual <https://www.ncbi.nlm.nih.gov/books/NBK25501/>
EDirect Help Manual <https://www.ncbi.nlm.nih.gov/books/NBK179288/>

CDS Extraction for Eukaryotic Organisms

Eukaryotes produce their proteins from mRNAs, which are spliced together from transcribed precursors. The mRNA and their protein products have a one-to-one correspondence. The CDS subsequences can be readily extracted from the corresponding mRNA records. Given a protein accession, we can use the elink function of EUtils to link to the mRNA record, then retrieve it from the nucleotide database in **fasta_cds_na** format to get the CDS. In EDirect, this workflow can be streamlined using the Unix shell pipe ("|") function. Given a set of protein accessions, we can post the list to the history server using the epost function, then pass the information to the elink and efetch to retrieve the CDS.

Below is an example set of command in EDirect:

```
$ epost -db protein -input protein-acc.txt | \  
  elink -db protein -target nuccore | \  
  efetch -db nuccore -format fasta_cds_na > protein_cds.fna
```

Line by line, the above command set does the following:

- epost the protein accessions (protein-acc.txt) to the protein database, and pass the output on
- elink the protein accessions (now reside on the history server in the form of a WebEnv and key) to find the target nucleotide (mRNA) entries, and pass the output on
- efetch the linked nucleotide records in special **fasta_cds_na** format to extract the CDS, redirect the output to a file named protein_cds.fna

CDS Extraction for Prokaryotic Organisms

For prokaryotic organisms, there is no corresponding mRNAs for protein records. Instead, proteins are annotated products on a genomic nucleotide record. Given a specific genomic record, it is simple to get all the CDS annotated on it as shown by the example efetch command below:

```
$ efetch -db nuccore -id NC_000913 -format fasta_cds_na > NC_000913_cds.fna
```

For a set of accessions of prokaryotic genomic records (bacteria-genome-acc.txt), we can add an epost step and use the follow command set to get all the CDS annotated on them:

```
$ epost -db nuccore -input bacteria-genome-acc.txt | \  
  efetch -db nuccore -format fasta_cds_na > bacteria_cds.fna
```

To extract the CDS of a single protein, we will need to use the genomic sequence's accession and the subsequence range to fetch the sequence. The Identical Protein Group report of a prokaryotic protein provides the necessary information for this purpose. The following command set extracts the RefSeq genomic subsequences and displays only the first 7 columns of the tabular output more relevant to CDS extraction and record keeping:

```
$ efetch -db protein -id AKU48328.1 -format ipg -mode text | grep "RefSeq" | cut -f 1-7
```

87573571	RefSeq	NZ_LWQY01000013.1	24615	26597	-	WP_064580732.1
87573571	RefSeq	NZ_LWTD01000002.1	56725	58707	+	WP_064580732.1
87573571	RefSeq	NZ_LWTD01000015.1	25051	27033	-	WP_064580732.1
87573571	RefSeq	NZ_LXKJ01000056.1	24497	26479	-	WP_064580732.1

They represent protein gi, source, genomic accession.version, start, stop, orientation, and the protein accession.version.

CDS Extraction for Prokaryotic Organisms (cont.)

Using the information from IPG, we can extract the subsequence for each of the CDS using `efetch`, one subsequence at a time. The corresponding `efetch` call for the first CDS entry is

```
$ efetch -db nuccore -id NZ_LWQY01000013.1 -seq_start 24615 -seq_stop 26597 -strand 2 -format fasta > test_cds.fa
```

For batch retrieval of CDS for a protein family across a broad spectrum of prokaryotic organisms, this approach will not scale. The redundancy in the IPG report (due to the redundant nature of the Protein database) also need to be eliminated. We will need alternative approach to address the challenge.

CDS Extraction for Prokaryotic Organisms Using Tools from blast+

In the example below, we will demonstrate the batch retrieval of CDS subsequences for a set of annotated gyrase B (`gyrB`) genes from the bacterial group *Enterococcus*, using a combination of `EDirect`, Perl script, plus the `makeblastdb` and `blastdbcmd` tools from standalone `blast+` package.

1. Search and retrieve IPG reports for complete `gyrB` sequences from Entrez Protein database

```
$ esearch -db protein -query "enterococcus[orgn] AND (gyrb[gene] OR gyrb[title]) NOT partial[title]" | \
  efetch -db protein -format ipg -mode text > gyrb_enterococcus_cds.txt
```

Since the protein database is redundant, there will be many duplicated rows in this output. We can easily remove the duplications using the `sort` and `uniq` commands. To further reduce duplication, we :

```
$ grep "RefSeq" gyrb_enterococcus_cds.txt | sort | uniq > gyrb_enterococcus_cds_refseq_nr.txt
```

2. This output tab-separated and we need the information from columns 3 - 7, with column 7 for reference purposes:

- Nature of the entry (in Step 1, we selected RefSeq and dropped INSDC)
- Genomic accession.version
- Start coordinate
- Stop coordinate
- Strand
- Protein accession.version

We can use the following inline perl script to extract columns 3 - 7 for use as input to downstream steps:

```
perl -e '@files=qw(gyrb_subset gyrb_genome gyrb_blast); @out=("","",""); while (<>) { @a=split(/\t/, $_); $out[0] .= join
("\t",@a[2..6])."\n"; $out[1] .= "$a[2]\n"; $out[2] .= "$a[2]\t$a[3]-$a[4]\t"; if ($a[5] eq "+") {$out[2] .= "\tplus\n"; } else {$out
[2] .= "\tminus\n";}} for ($c=0; $c<3; $c++){ open (O, ">$files[$c]"); print O $out[$c]; close (O);} exit;'
gyrb_enterococcus_cds_refseq_nr.txt
```

This inline script generates three files from the Step 1 output: a file named **gyrb_subset** with information from column 3 - 7 for our record, a file called **gyrb_genome** with column 3 for `efetching` the genomic records, and a third file named **gyrb_blast** with information from column 3 - 6, but transformed into `blastdbcmd` recognizable format for subsequence retrieval.

3. Use the `gyrb_genome` file as input to download the genomic records.

```
$ epost -db nuccore -input gyrb_genome | efetch -db nuccore -format fasta > gyrb_genome.fa
```

This post the genomic accession.version to history server and pass the list to `efetch` to get the FASTA sequences for these genomic records.

4. To retrieve the subsequences for the CDS we need, we first need to format the genome FASTA file into a BLAST database using `makeblastdb`:

```
$ makeblastdb -in gyrb_genome_cds.fa -dbtype nucl -parse_seqids -out gyrb_genomes
```

then extract the CDS using `blastdbcmd`:

```
$ blastdbcmd -db refseq_genomic -entry_batch gyrb_blast >gyrb_cds.fa
```

To correlate the extracted CDS back to the protein, we will need to refer to the `gyrb_subset` file's genomic sequence accession.version to get the protein accession.version (not shown).

Questions and Comments

Please send them to info@ncbi.nlm.nih.gov