

# Retrieve the geographic coordinates associated with DNA sequences

Luis Amador

2022-05-17

## Overview

This is a small tutorial to retrieve the geographic information of published DNA sequences in the form of geographic coordinates: Latitude and Longitude (e.g., Latitude = 35.0775, Longitude = -106.6625).

This work is part of the NSF project awarded to professor Lisa Barrow: *Determinants of amphibian genomic diversity: Integrating traits, phylogeny and geography*. Specifically, we are trying to retrieve geographic coordinates of mitochondrial DNA (mtDNA) sequences of the Cytochrome B gene belonging to amphibian species.

We recommend some steps to follow:

## Steps

1. We have a directory with sequences of amphibian species in **FASTA** format. In the figure 1 we can see a typical alignment, in this case we see an alignment of nine sequences of the Toncheek spiny-chest frog (*Alsodes gargola*). Each sequence is coded with the scientific name and the accession number of **GenBank** (e.g., Alsodes.gargola.AY843787). This is our input, since we can extract from the alignment the information that we will need to find the geographic coordinates associated with a certain DNA sequence.

The alignment can also be opened in any program that reads text files, for example Microsoft Notepad or Microsoft WordPad, and BBEdit if you are using macOS. In addition, there are several free software to edit and visualize alignments, such as BioEdit, MEGA, Aliview or Jalview.

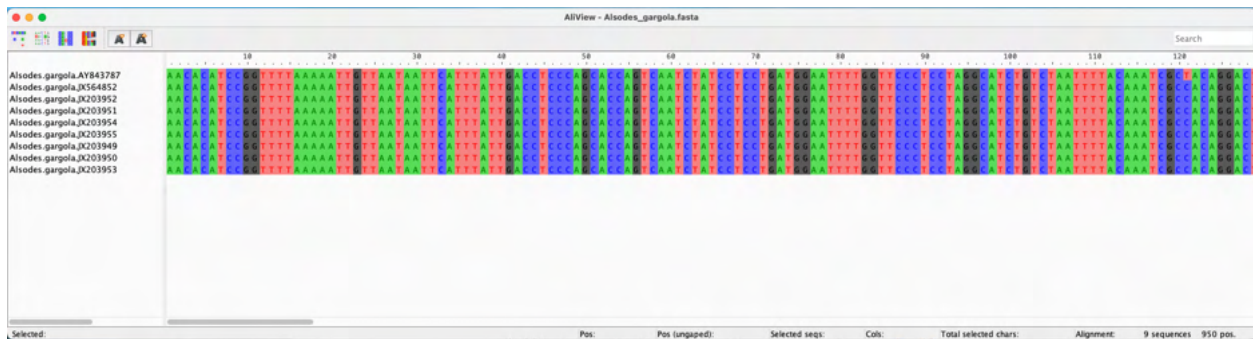


Figure 1: View of a typical alignment in the software Aliview.

2. We extract the accession number of each sequence (e.g., AY843787) of the alignment and go to the website to The National Center for Biotechnology Information (NCBI) <https://www.ncbi.nlm.nih.gov/nucleotide/>. Then we can search for the information of the sequence in the NCBI portal by write the accession number in the browser (Figure 2).

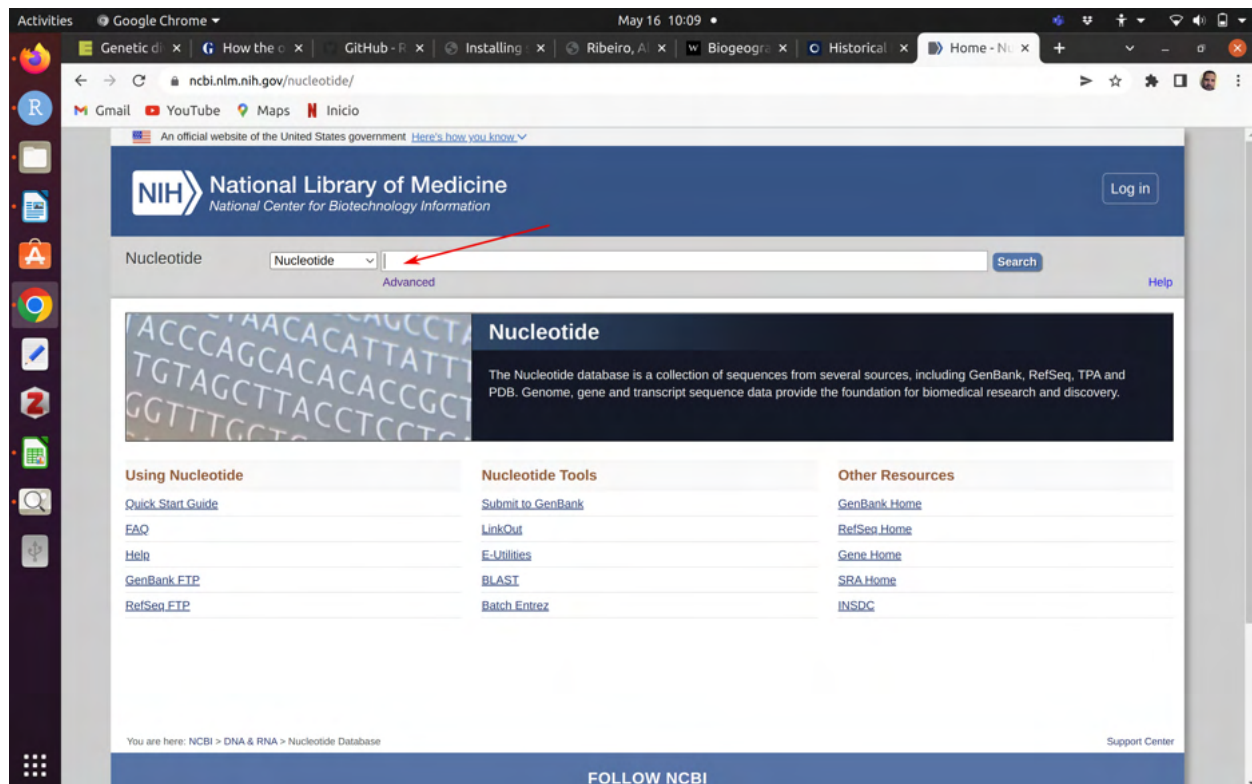


Figure 2: Window of NCBI portal to search in the nucleotide database.

3. We can see the information of the GenBank submission that consist primarily of the nucleotide sequence data, source organism information, and sequence features. If the geographic coordinates are not part of the features of the GenBank record, two key parameters to start the search of the geographic coordinates are: 1) “TITLE” that is the name of the publication where the sequenced was released, and 2) feature source such as the locality “country” and which are shown as shaded text (Figure 3).
4. In some cases, we can find in the GenBank submission page a link (Figure 4A) to the specimen voucher information deposited in databases (e.g., museum databases like arctos). Luckily, the geographic coordinates in latitude and longitude format can be directly extracted (copied) from this database (Figure 4B).
5. There are several ways to retrieve geographic coordinates from the information of the GenBank submission. Here we are showing two ways using as example the GenBank accession number KJ418850. In this example, the article associated to the sequence appears as unpublished (Figure 5) in “TITLE”, in addition, we can found information about the locality of the specimen voucher (Region de O’Higgins, Rinconada de Idahue).

First we can do a Google search with the title of the article (maybe it is published already!) and check within the article, for the information (i.e., geographic coordinates) related to the accession number/specimen, that could be presented wrote in the text of the article, in a table in the main text of the article, or as part of the supplementary material of the article. In this example, we do not find the geographic coordinates within the article, but we can use other information that could help to get the geographic coordinates (Figure 6).

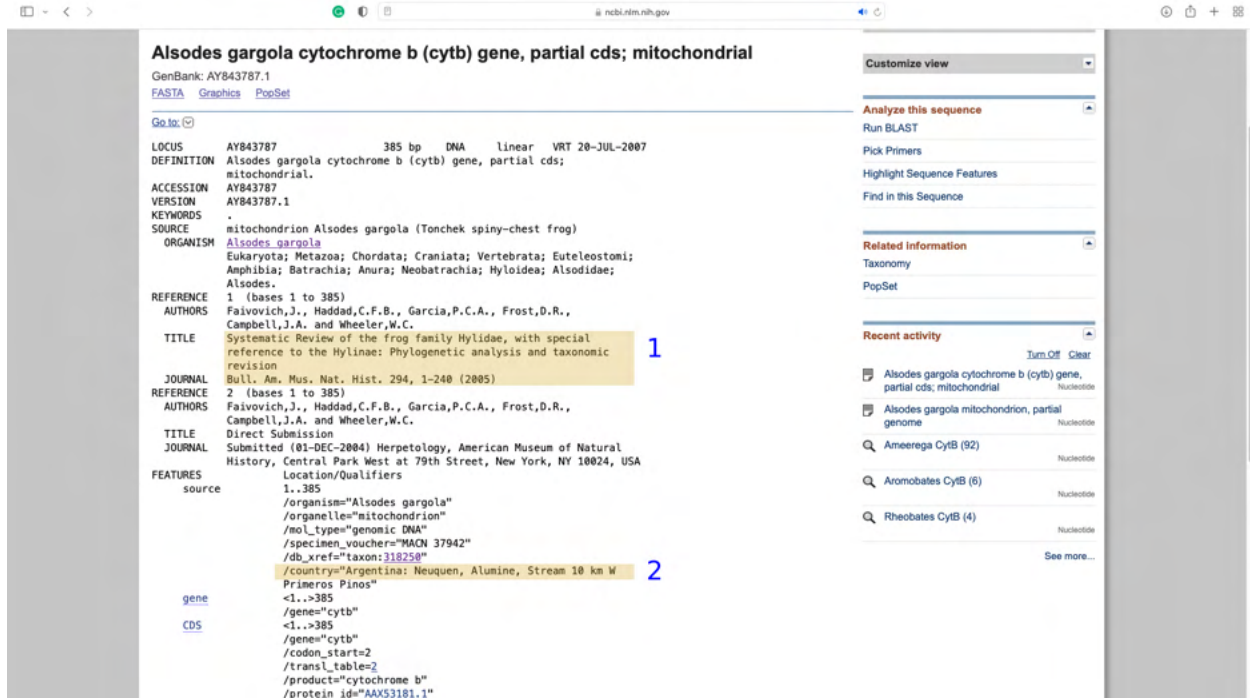


Figure 3: Window of NCBI portal with the information of the GenBank submission.

6. Another way to retrieve the geographic coordinates is to search for the name of the locality in **Google Earth** (e.g., Rinconada de Idahue), and obtain the geographic coordinates directly from there (Figure 7). Google Earth gives the coordinates in DMS (degrees, minutes and seconds) but for our analyses we are using DD (decimal degrees) latitude and longitude format for geographic coordinates (Table 1).

Table 1. Geographic coordinates of a locality in DMS and DD.

Locality	DMS coordinates	Latitude	Longitude
Rinconada de Idahue	34°17'28"S; 71°08'42"W	-71.145	-34.291

Fortunately, there are several websites to change the format of coordinates, we are using <https://www.gps-coordinates.net/> which is easy to use and with an acceptable interface (Figure 8).



The screenshot shows the NCBI GenBank submission page for the sequence **Alsidodes nodosus voucher DBGUCH 0905001 cytochrome b (cytb) gene, partial cds; mitochondrial**. The page is divided into several sections:

- LOCUS:** KJ418850 933 bp DNA Linear VRT 27-JAN-2015
- DEFINITION:** Alsidodes nodosus voucher DBGUCH 0905001 cytochrome b (cytb) gene, partial cds; mitochondrial.
- ACCESSION:** KJ418850
- VERSION:** KJ418850.1
- KEYWORDS:** mitochondrial Alsidodes nodosus (black spiny-chest frog)
- SOURCE:** Alsidodes nodosus
- ORGANISM:** Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Amphibia; Batrachia; Anura; Neobatrachia; Myloidea; Alsididae; Alsidodes.
- REFERENCE:** 1 (bases 1 to 933) Correa, C. NISden phylogenetic diversity in anurans from Central Chile: a new species of Alsidodes (Anura, Alsididae) from Alto de Cantillana. Unpublished.
- JOURNAL:** 2 (bases 1 to 933) Correa, C. Direct Submission. Submitted (09-FEB-2014) Departamento de Zoología, Universidad de Concepción, Barrio Universitario Sin Numero, Concepción, Biobío 160-C, Chile.
- COMMENT:** ##Assembly-Data-START## Sequencing Technology :: Sanger dideoxy sequencing ##Assembly-Data-END##
- FEATURES:**
  - source** 1..933
    - /organism="Alsidodes nodosus"
    - /organism="mitochondrion"
    - /mol\_type="genomic DNA"
    - /isolate="An0905001R1"
    - /specimen\_voucher="DBGUCH 0905001"
    - /db\_xref="taxon:83398"
    - /country="Chile: Región de O'Higgins, Rincónada de Idahué"
  - gene** <1..>933
    - /gene="cytb"
  - cds** <1..>933
    - /gene="cytb"
    - /codon\_start=2
    - /transl\_table=2
    - /product="cytochrome b"
    - /protein\_id="AJ218850.1"
    - /translation="SFTDLFAPALSSAMFGLGICLIGIATGLFAMHYADTSMAFSSIAHCRDNNAGLNNLHANGASFFICITLHSGRTYIGSYLFKETHNIGVILFLMATAFGVYLPNGQPSMGATYTNLLSAAPYISGTELVMKNGGFSVQATLTHTFTTFPLPIIAGASHLLFLHGTSSDPFGLSNGKRVFFHAYYSYNDYDFIVLLALLAILSAFAPNIGDPDNFTPAPPLVTPPKPEWYFFAYAILRSIPNKLGGVLLAFSLILFLMPLFHLSHORSIFRPLAKLLFWT"
- ORIGIN**

```

1 ctcgtttatt gactctccag caccagctaa ctatctcttc tpatggaact ttggttcctt
61 cctaggcacc tgcctaattt tacaatctgc tacagpctta ttttttagta tacattatgc
121 tctctatacc taatatacat tctcatctcat cctctacatt totctaatcaca taataaacaa

```

On the right side, there is a sidebar with options like "Customize view", "Analyze this sequence", "Run BLAST", "Pick Primers", "Highlight Sequence Features", "Find in this Sequence", "Related information", "Protein", "Taxonomy", "PopSet", and "Recent activity".

Figure 5: Window of NCBI portal with the information of the GenBank submission.





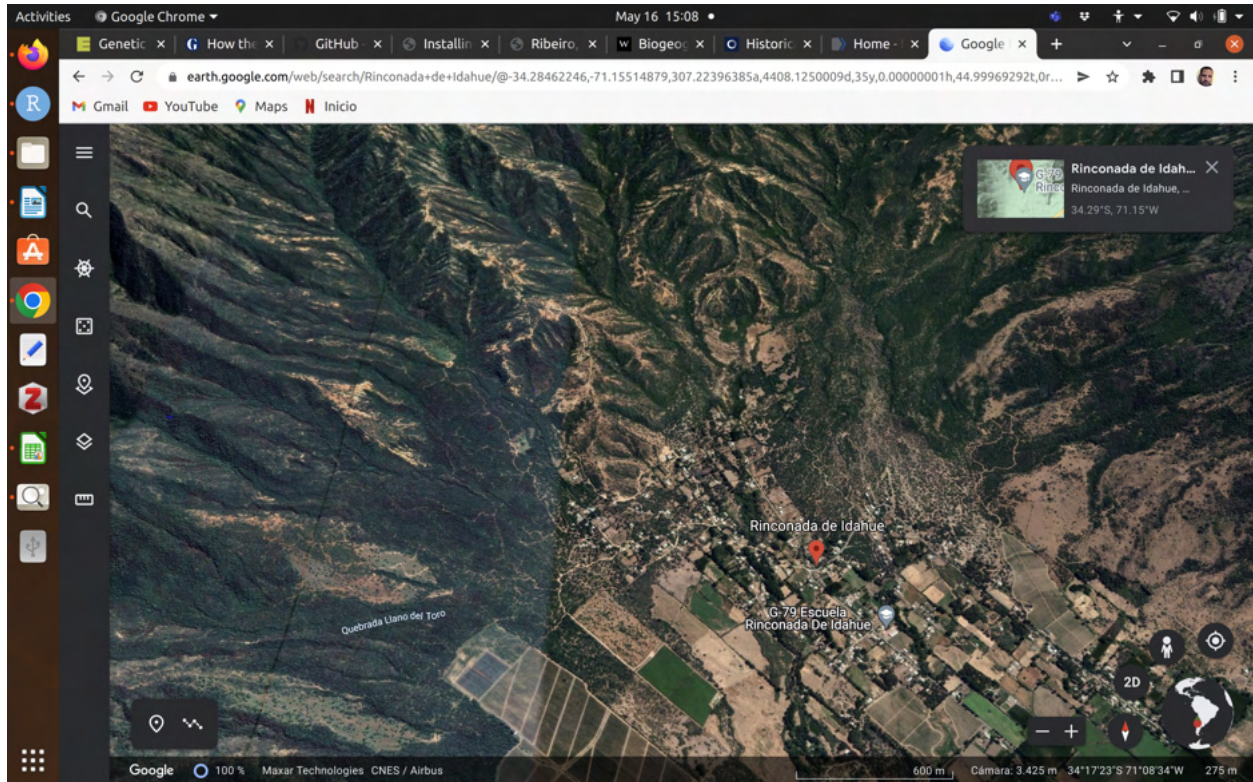


Figure 7: Google Earth interface, geographic coordinates can be found in the bottom right of the screen.

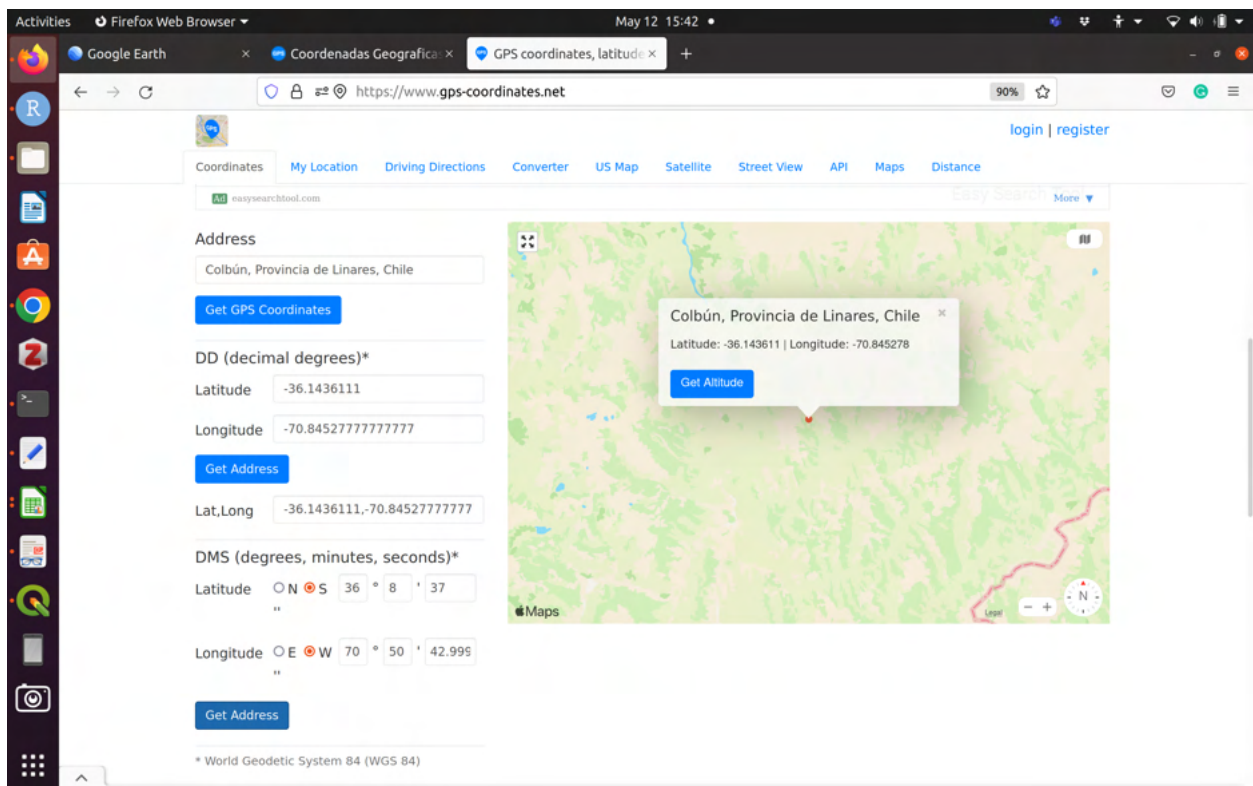


Figure 8: Example of tool where we can format the geographic coordinates.