

# Working with biological sequence data

# We will use the Biopython package

## <http://biopython.org>



[Documentation](#)

[Download](#)

[Mailing lists](#)

[News](#)

[Biopython Contributors](#)

[Scriptcentral](#)

[Source Code](#)

[Biopython Source Code -](#)

[Redirection](#)

[GitHub project](#)

[Edit this page on GitHub](#)

## Biopython

See also our [News feed](#) and [Twitter](#).

### Introduction

Biopython is a set of freely available tools for biological computation written in [Python](#) by an international team of developers.

It is a distributed collaborative effort to develop Python libraries and applications which address the needs of current and future work in bioinformatics. The source code is made available under the [Biopython License](#), which is extremely liberal and compatible with almost every license in the world.

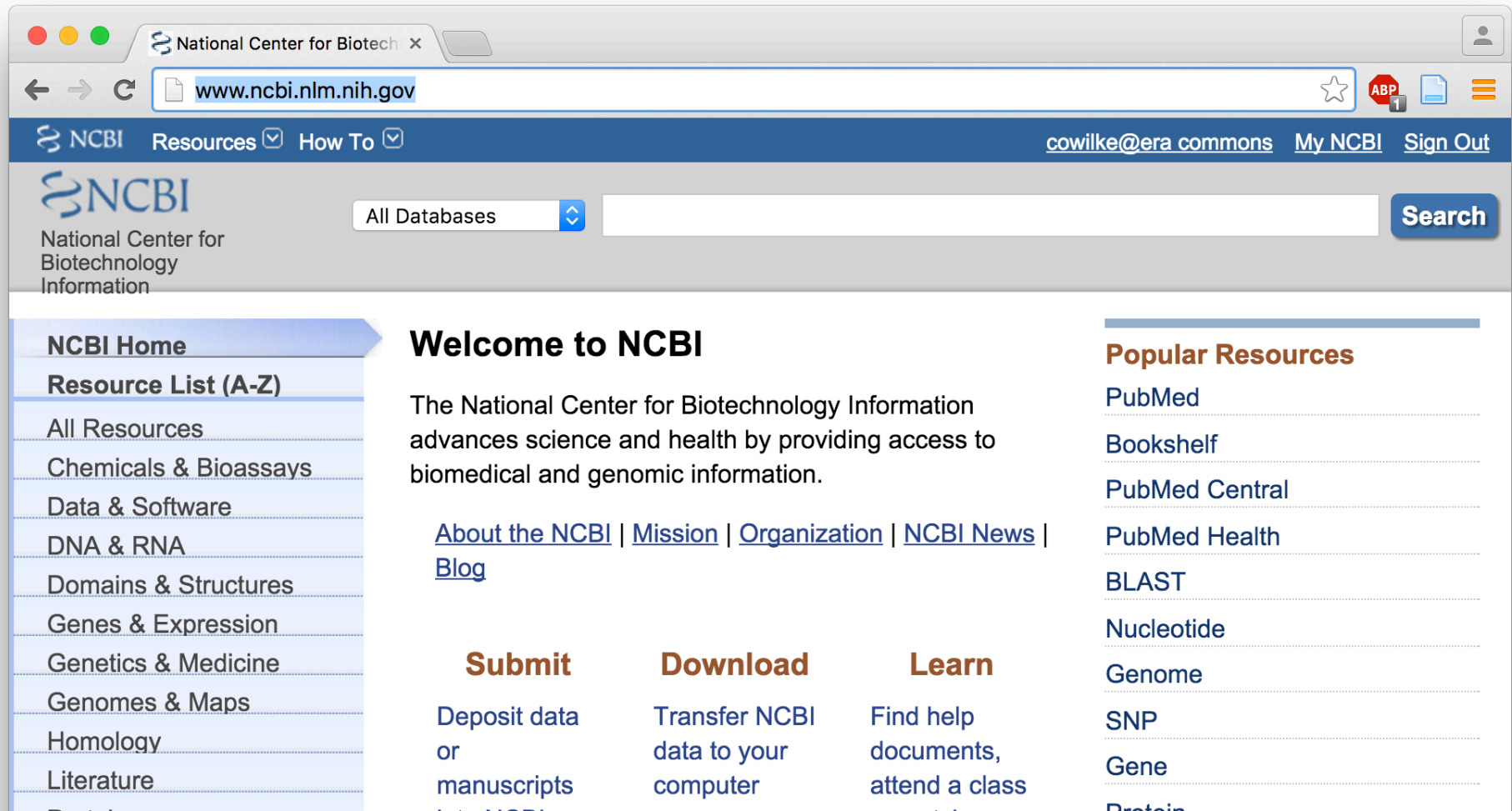
We are a member project of the [Open Bioinformatics Foundation \(OBF\)](#), who take care of our domain name and hosting for our mailing list etc. The OBF used to host our development repository, issue tracker and website but these are now on [GitHub](#).

This wiki will help you download and install Biopython, and start using the libraries and tools.

<a href="#">Get Started</a>	<a href="#">Get help</a>	<a href="#">Contribute</a>
<a href="#">Download Biopython</a>	<a href="#">Tutorial (PDF)</a>	<a href="#">What's being worked on</a>
<a href="#">Installation help (PDF)</a>	<a href="#">Documentation on this wiki</a>	<a href="#">Developing on Github</a>
	<a href="#">Cookbook (working examples)</a>	<a href="#">Google Summer of Code</a>

# Getting biological data: The NCBI databases

<http://www.ncbi.nlm.nih.gov/>



The screenshot shows the NCBI website homepage. At the top, there's a browser window with the address bar showing 'www.ncbi.nlm.nih.gov'. Below the browser window, the NCBI logo is on the left, and a navigation bar contains 'Resources' and 'How To' dropdown menus. On the right of the navigation bar, there's a user profile icon, a star icon, and a red 'ABP' icon. Below the navigation bar, there's a search bar with a dropdown menu set to 'All Databases' and a 'Search' button. On the left side, there's a 'NCBI Home' section with a 'Resource List (A-Z)' and a list of resources: All Resources, Chemicals & Bioassays, Data & Software, DNA & RNA, Domains & Structures, Genes & Expression, Genetics & Medicine, Genomes & Maps, Homology, and Literature. In the center, there's a 'Welcome to NCBI' section with a paragraph about the center's mission and a list of links: About the NCBI, Mission, Organization, NCBI News, and Blog. Below this, there are three columns: 'Submit' (Deposit data or manuscripts), 'Download' (Transfer NCBI data to your computer), and 'Learn' (Find help documents, attend a class). On the right side, there's a 'Popular Resources' section with a list of resources: PubMed, Bookshelf, PubMed Central, PubMed Health, BLAST, Nucleotide, Genome, SNP, Gene, and Protein.

National Center for Biotech x

← → ↻ [www.ncbi.nlm.nih.gov](http://www.ncbi.nlm.nih.gov) ☆ ABP

NCBI Resources ▾ How To ▾ cowilke@era commons My NCBI Sign Out

NCBI  
National Center for  
Biotechnology  
Information

All Databases ▾ Search

**NCBI Home**  
**Resource List (A-Z)**  
All Resources  
Chemicals & Bioassays  
Data & Software  
DNA & RNA  
Domains & Structures  
Genes & Expression  
Genetics & Medicine  
Genomes & Maps  
Homology  
Literature

**Welcome to NCBI**  
The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.  
[About the NCBI](#) | [Mission](#) | [Organization](#) | [NCBI News](#) | [Blog](#)

**Submit**  
Deposit data or manuscripts

**Download**  
Transfer NCBI data to your computer

**Learn**  
Find help documents, attend a class

**Popular Resources**  
PubMed  
Bookshelf  
PubMed Central  
PubMed Health  
BLAST  
Nucleotide  
Genome  
SNP  
Gene  
Protein

# Try search for "KT220438"

The screenshot shows a web browser window with the address bar displaying `www.ncbi.nlm.nih.gov/gquery/?term=KT220438`. The page title is "KT220438 - GQuery: Glob: x". The NCBI logo and navigation links (Resources, How To) are visible in the header. The user is logged in as `cowilke@era commons` with links for "My NCBI" and "Sign Out".

The main heading is "Search NCBI databases" with a "Help" link. A search input field contains "KT220438" and a "Search" button. Below the search bar, the results are summarized as "Results found in 2 databases for 'KT220438'".

The search results are displayed in a light blue box:

- [Influenza A virus \(A/NewJersey/NHRC\\_93219/2015\(H3N2\)\) segment 4 hemagglutinin \(HA\) gene, complete cds](#)
- 1,701 bp cRNA.
- Lab\_host: MDCK. Country: USA: New Jersey. Segment: 4. Isolation\_source: nasopharyngeal swab. Collection\_date: 17-Jan-2015.
- Accession: **KT220438.1** GI: 887493048
- [GenBank](#) [FASTA](#) [Graphics](#)

Below the search results, there are two sections: "Literature" and "Genes".

Literature		Genes	
<b>Books</b>	0	<b>EST</b>	0
<b>MeSH</b>	0		

Under "Literature", there are links for "books and reports" and "ontology used for PubMed indexing". Under "Genes", there are links for "expressed sequence tag sequences" and "collected information about".

# Direct link to search results

<http://www.ncbi.nlm.nih.gov/gquery/?term=KT220438>

A genbank record is just a simple text file

LOCUS KT220438 1701 bp cRNA linear VRL 20-JUL-2015  
 DEFINITION Influenza A virus (A/NewJersey/NHRC\_93219/2015(H3N2)) segment 4  
 hemagglutinin (HA) gene, complete cds.  
 ACCESSION KT220438  
 VERSION KT220438.1 GI:887493048  
 KEYWORDS .  
 SOURCE Influenza A virus (A/New Jersey/NHRC\_93219/2015(H3N2))  
 ORGANISM Influenza A virus (A/New Jersey/NHRC\_93219/2015(H3N2))  
 Viruses; ssRNA viruses; ssRNA negative-strand viruses;  
 Orthomyxoviridae; Influenzavirus A.  
 REFERENCE 1 (bases 1 to 1701)  
 AUTHORS Sitz,C.R., Thammavong,H.L., Balansay-Ames,M.S., Hawksworth,A.W.,  
 Myers,C.A. and Brice,G.T.  
 TITLE GEISS Influenza Surveillance Response Program  
 JOURNAL Unpublished  
 REFERENCE 2 (bases 1 to 1701)  
 AUTHORS Sitz,C.R., Thammavong,H.L., Balansay-Ames,M.S., Hawksworth,A.W.,  
 Myers,C.A. and Brice,G.T.  
 TITLE Direct Submission  
 JOURNAL Submitted (29-JUN-2015) Operational Infectious Diseases, Naval  
 Health Research Center, 140 Sylvester Rd., San Diego, CA 92106, USA  
 COMMENT ##Assembly-Data-START##  
 Sequencing Technology :: Sanger dideoxy sequencing  
 ##Assembly-Data-END##  
 FEATURES Location/Qualifiers  
 source 1..1701  
 /organism="Influenza A virus (A/New  
 Jersey/NHRC\_93219/2015(H3N2))"  
 /mol\_type="viral cRNA"  
 /strain="A/NewJersey/NHRC\_93219/2015"  
 /serotype="H3N2"

## FEATURES

source

Location/Qualifiers

1..1701

/organism="Influenza A virus (A/New  
Jersey/NHRC\_93219/2015(H3N2))"

/mol\_type="viral cRNA"

/strain="A/NewJersey/NHRC\_93219/2015"

/serotype="H3N2"

/isolation\_source="nasopharyngeal swab"

/host="Homo sapiens"

/db\_xref="taxon:1682360"

/segment="4"

/lab\_host="MDCK"

/country="USA: New Jersey"

/collection\_date="17-Jan-2015"

gene

1..1701

/gene="HA"

CDS

1..1701

/gene="HA"

/function="receptor binding and fusion protein"

/codon\_start=1

/product="hemagglutinin"

/protein\_id="AKQ43545.1"

/db\_xref="GI:887493049"

/translation="MKTIIALSYILCLVFAQKIPGNDNSTATLCLGHHAVPNGTIVKT  
ITNDRIEVTNATELVQNSSIGEICDSPHQILDGENCTLIDALLGDPQCDGFQNKKWDL  
FVERSKAYSNCYPYDVPDYASLRSLVASSGTLEFNNE SFNWTGVTQNGTSSACIRRSS  
SSFFSRLNWLTHLNYTYPALNVTMPNNEQFDKLYIWGVHHPGTDKDKQIFLYAQSSGRI  
TVSTKRSQQAVIPNIGSRPRIRDIPSRSISYWTIVKPGDILLINSTGNLIAPRGYFKI  
RSGKSSIMRSDAPIGKCKSECITPNGSIPNDKPFQNVNRITYGACPRYVKHSTLKLAT  
GMRNVPEKQTRGIFGAIAGFIENGWEGMVDGWYGFRHQNSEGRGQAADLKSTQAAIDQ  
INGKLNRLIGKTNEKFFHOIEKEEFSEVEGRIODLEKYVEDTKIDLWSYNAELIVALENO



# ORIGIN

```

1  atgaagacta tcattgcttt gagctacatt ctatgtctgg ttttcgctca aaaaattcct
61  ggaaatgaca atagcacggc aacgctgtgc cttgggcacc atgcagtacc aaacggaacg
121 atagtgaaaa caatcacaaa tgaccgaatt gaagttacta atgctactga gctggttcag
181 aattcctcaa taggtgaaat atgcgacagt cctcatcaga tccttgatgg agaaaactgc
241 acactaatag atgctctatt gggagaccct cagtgtgatg gctttcaaaa taagaaatgg
301 gacctttttg ttgaacgaag caaagcctac agcaactgct acccttatga tgtgccggat
361 tatgcctccc ttaggtcact agttgcctca tccggcacac tggagtttaa caatgaaagc
421 ttcaattgga ctggagtcac tcaaaacgga acaagttctg cttgcataag gagatctagt
481 agtagtttct ttagtagatt aaattggttg acccacttaa actacacata cccagcattg
541 aacgtgacta tgccaaacaa tgaacaattt gacaaattgt acatttgggg gggtcaccac
601 ccgggtacgg acaaggacca aatcttccctg tatgctcaat catcaggaag aatcacagta
661 tctacaaaaa gaagccaaca agctgtaatc ccaaatatcg gatctagacc cagaataagg
721 gatatcccta gcagaataag catctattgg acaatagtaa aaccgggaga catacttttg
781 attaacagca cagggaatct aattgctcct aggggttact tcaaaatacg aagtgggaaa
841 agctcaataa tgagatcaga tgcacccatt ggcaaatagca agtctgaatg catcactcca
901 aatggaagca ttcccaatga caaaccattc caaaatgtaa acaggatcac atacggggcc
961 tgtcccagat atgttaagca tagcactcta aaattggcaa caggaatgcg aaatgtacca
1021 gagaaacaaa ctagaggcat atttggcgca atagcgggtt tcatagaaaa tggttgggag
1081 ggaatggtgg atggttggtg cggtttcagg catcaaaatt ctgagggaag aggacaagca
1141 gcagatctca aaagcactca agcagcaatc gatcaaatca atgggaagct gaatcgattg
1201 atcgggaaaa ccaacgagaa attccatcag attgaaaaag aattctcaga agtagaagga
1261 agaattcagg accttgagaa atatgttgag gacactaaaa tagatctctg gtcatacaac
1321 gcggagcttc ttgttgccct ggagaaccaa catacarttg atctaactga ctcagaaatg
1381 aacaaactgt ttgaaaaaac aaagaagcaa ctgagggaaa atgctgagga tatgggaaat
1441 ggttgtttca aaatatacca caaatgtgac aatgcctgca taggatcaat aagaaatgga
1501 acttatgacc acaatgtgta cagggatgaa gcattaaaca accggttcca gatcaaggga
1561 gttgagctga agtcagggtg caaagattgg atcctatgga tttcctytgc catatcatgt
1621 tttttgcttt gtgttgcttt gttgggggtc atcatgtggg cctgccaaaa gggcaacatt
1681 aggtgcaaca tttgcatttg a

```