Working with biological sequence data
We will use the Biopython package http://biopython.org

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.

<table>
<thead>
<tr>
<th>Get Started</th>
<th>Get help</th>
<th>Contribute</th>
</tr>
</thead>
<tbody>
<tr>
<td>Download Biopython</td>
<td>Tutorial (PDF)</td>
<td>What's being worked on</td>
</tr>
<tr>
<td>Installation help (PDF)</td>
<td>Documentation on this wiki</td>
<td>Developing on Github</td>
</tr>
<tr>
<td></td>
<td>Cookbook (working examples)</td>
<td>Google Summer of Code</td>
</tr>
</tbody>
</table>
Getting biological data: The NCBI databases

Try search for "KT220438"
A genbank record is just a simple text file
**DEFINITION** Influenza A virus (A/New Jersey/NHRC_93219/2015(H3N2)) segment 4 hemagglutinin (HA) gene, complete cds.

**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**JOURNAL** Unpublished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**

source 1..1701

/organism="Influenza A virus (A/New Jersey/NHRC_93219/2015(H3N2))"
/mol_type="viral cRNA"
/strain="A/New Jersey/NHRC_93219/2015"
/serotype="H3N2"
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"
/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="MKTIALSYILCLVFAQKIPGNNDNSTATLCLGHHAVPNGTIVKTI
TNDRIEVTNATELVQNSIGEICDSPHQILDGENCTLIDALLGDPQCDGFQNNKWDLF
VERSKAYSNCPYDVPDYASLRSLVASSGTEFNNESFNWTGVTQNGTSSACIFRRSS
SSFSSRLNWLTHLNYTYPALNVTPMNNEQFDKLHYGWGHQGTDKDOIFIQLFAQSSG"R
LTVSTKRSSQAVIPNIASRPRIDPRISIYTWITVKGHIILLINSTGNLIAKPRGFYKI
RSGKSIIMRSDAPIGCKCKSECITPNGSIPNDFQNVNRTYGACPFRVKHSTLK"LAT
GMRVPEKQTRGIFGAIAGFIEGMVGMDGWYGFRQHONSEGGRQAADLKSTQAAIDQ
INGKLNRLIGKTNKEHOTEKESEVEGRIODLEKYVEDETKIDLSWYNAELIL"VALENQ