Lab Worksheet 14 Solutions

The web interface to BLAST is available here: http://blast.ncbi.nlm.nih.gov/Blast.cgi (http://blast.ncbi.nlm.nih.gov/Blast.cgi)

Let's search for proteins related to the following query sequence, which is the human chemokine receptor 4 (a receptor that plays a fundamental role in the immune system):

```
>human
MSIPLPLLQIYTSDNYTEEMGSGDYDSMKEPCFREENANFNKIFLPTIYSIIFLTGIVGN
GLVILVMGYQKKLRSMTDKYRLHLSVADLLFVITLPFWAVDAVANWYFGNFLCKAVHVIY
TVNLYSSVLILAFISLDRYLAIVHATNSQRPRKLLAEKVVYVGVWIPALLLTIPDFIFAN
VSEADDRYICDRFYPNDLWVVVFQFQHIMVGLILPGIVILSCYCIIISKLSHSKGHQKRK
ALKTTVILILAFFACWLPYYIGISIDSFILLEIIKQGCEFENTVHKWISITEALAFFHCC
LNPILYAFLGAKFKTSAQHALTSVSRGSSLKILSKGKRGGHSSVSTESESSSFHSS
```

Problems

Problem 1:

Download the blast results from the NCBI website in XML format and store them as `cxcr4_BLAST.xml`. Extract the genbank identifiers (written as gb|string|, where string is the actual identifier, consisting of letters, numbers, and the period symbol) for all matches with a score greater than or equal to 1600 and less than or equal 1800, and store them in a python list. For matches that list multiple genbank identifiers, only extract the first one.

In [1]:
```
from Bio.Blast import NCBIXML
import re

# open the downloaded file and parse with NCBIXML.read()
blast_handle = open("cxcr4_BLAST.xml")
blast_record = NCBIXML.read(blast_handle)
blast_handle.close()

gb_list = []
for alignment in blast_record.alignments:
    for hsp in alignment.hsps:
        if hsp.score >= 1600 and hsp.score <= 1800:
            match = re.search(r'gb\|([\w\d\.]+)\|', alignment.title)
            if match:
                gb_id = match.group(1)
                gb_list.append(gb_id)

print(gb_list)
```

```
['AAF89355.1', 'AAF89362.1', 'ABX55951.1', 'AAF42991.1', 'AAF42992.1', 'AAF4299
0.1', 'ABX55952.1', 'AAF37288.1', 'EFB23364.1', 'ABA28309.1', 'KFO22725.1', 'EH
B03245.1', 'ELR58312.1', 'AAF89363.1', 'ACH54079.1', 'AAZ32767.1', 'AAF89359.1'
, 'AAC48852.1']
```

Problem 2:

Using the list of genbank identifiers obtained in the previous exercise, download the corresponding sequences from genbank and print them out in FASTA format.

Hints:

- You will have to specify the database as "protein" for this to work, since the previous exercise generated identifiers for protein sequences.
- Use the function `SeqIO.write()` to output your results in FASTA format, and use `sys.stdout` from the `sys` module as your output handle.

In [2]:
```
from Bio import Entrez, SeqIO
import sys

Entrez.email = "dariya.k.sydykova@gmail.com" # put your email here

handle = Entrez.efetch(db="protein", id=gb_list, rettype="gb", retmode="text")
records = SeqIO.parse(handle, "genbank")

for record in records:
    SeqIO.write(record, sys.stdout, "fasta")

handle.close() # important, close the handle only after you have iterated over
the records. Otherwise you will get an error!
```

```
>AAF89355.1 chemokine receptor CXCR4, partial [Callithrix jacchus]
IYTSDNYTEEIGSGDYDSIKEPCFREENAHFNRIFLPTIYSIIFLTGIVGNGLVILVMGY
QKKLRSMTDKYRLHLSVADLLFVITLPFWAVDAVANWYFGKFLCKAVHVIYTVNLYSSVL
ILASISLDRYLAIVHATTSQRPPKLLAEKVVYVGVWIPALLLTIPDFIFANVSEADDRYI
CDRFYPNDLWVVVFQFQHIMVGLILPGIVILSCYCIIISKLSHSKGHQKRKALKTTVILI
LAFFACWLPYYIGISIDSFILLEIIRQGCEFENTVHKWISITEALAFFHCCLNPILYAFL
GAKFKTSAQHALTSVSRGSSLKILSKGKRGGHSSVSTESESSSFHSS
>AAF89362.1 chemokine receptor CXCR4, partial [Eulemur macaco]
IYTSDNYTEELGSGDYDSIKEPCFREENAHFNRIFLPTIYSIIFLTGIVGNGLVILVMGY
QKKLRSMTDKYRLHLSVADLLFVITLPFWAVDAVANWYFGKFLCKAVHVIYTVNLYSSVL
ILAFISLDRYLAIVHATNSQRPRKLSAEKVVYAGVWLPALLLTIPDFIFASVSEVDDRYI
CDRLYPNDLWVVVFQFQHIMVGLILPGIVILSCYCIIISKLSHSKGHQKRKALKTTVILI
LAFFACWLPYYIGISIDSFILLEIIKQGCEFENTVHKWISITEALAFFHCCLNPILYAFL
GAKFKTSAQHALSSVSRGSSLKILSKGKRGGHSSVSTESESSSFHSS
>ABX55951.1 chemokine receptor, partial [Oryctolagus cuniculus]
TSDNYTEELGSGDYDSIKEPCFREENAHFNRIFLPTIYSIIFLTGIVGNGLVILVMGYQK
KQRSMTDKYRLHLSVADLLFVITLPFWAVDAVANWYFGKFLCKAVHVIYTVNLYSSVLIL
AFISLDRYLAIVHATNSQKPRKLLAEKVVYVGVWIPALLLTIPDFIFANVREAEGRYICD
RFYPSDLWVVVFQFQHIMVGLILPGIVILSCYCIIISKLSHSKGHQKRKALKTTVILILA
FFACWLPYYIGISIDSFILLEIIKQGCEFENTVHKWISITEALAFFHCCLNPILYAFLGA
KFKTSAQHALTSVSRGSSLKILSKGKRGGHSSVSTESESSSFHSS
>AAF42991.1 CXC chemokine receptor 4, partial [Hylobates lar]
EEMGSGDYDSIKEPCFREENANFNKIFLPTIYSIIFLTGIVGNGLVILVMGYQEKLRSMT
DKYRLHLSVADLLFVITLPFWAVDAVANWYFGNFLCKAVHVIYTVNLYSSVLILAFISLD
RYLAIVHATNSQRPRKLLAEKVVYVGVWIPALLLTIPDFIFANVSEADDRYICDRFYPND
LWVVVFQFQHIMVGLILPGIVMLSCYCIIISKLSHSKGHQKRKALKTTVILILAFFACWL
PYYIGISIDSFILLEIIKQGCEFENTVHKWISITEALAFFHCCLNPILYAFLGAKFKTSA
QHALTSVSRGSSLKILSKGKRGGHSSVSTESESS
>AAF42992.1 CXC chemokine receptor 4, partial [Saguinus oedipus]
EEMGSGDYDSMKEPCFREENAHFNRIFLPTIYSIIFLTGIVGNGLVILVMGYQKKLRSMT
DKYRLHLSVADLLFVITLPFWAVDAVANWYFGKFLCKAVHVIYTVNLYSSVLILAFISLD
RYLAIVHATNSQRPRKLLAEKVVYVGVWIPALLLTIPDFIFANVSEADDRYICDRFYPND
LWVVVFQFQHIMVGLILPGIVILSCYCIIISKLSHFKGHQKRKALKTTVILILAFFACWL
PYYIGISIDSFILLEIIRQGCEFENTVHKWISITEALAFFHCCLNPILYAFLGAKFKTSA
QHALTSVSRGSSLKILSKGKRGGHSSVSTESESS
>AAF42990.1 CXC chemokine receptor 4, partial [Chlorocebus aethiops]
EEMGSGDYDSIKEPCFREENAHFNRIFLPTIYSIIFLTGIVGNGLVILVMGYQKKLRSMT
DKYRLHLSVADLLFVITLPFWAVDAVANWYFGNFLCKAVHVIYTVNLYSSVLILAFISLD
RYLAIVHATNSQRPRKLLAEKVVYVGVWIPALLLTIPHFIFASVSEADDRYICDRFYPND
LWVVVFQFQHIMVGLILPGIVILSCYCIIISKLSHSKGHQKRKALKTTVILILAFFACWL
PYYIGISIDSFILLEIIKQGCEFENTVHKWISITEALAFFHCCLNPILYAFLGAKFKTSA
QHALTSVSRGSSLKILSKGKRGGHSSVSTESESS
>ABX55952.1 chemokine receptor, partial [Oryctolagus cuniculus]
TSDNYTEELGSGDYDSIKEPCFREENAHFNRIFLPTIYSIIFLTGIVGNGLVILVMGYQK
KQRSMTDKYRLHLSVADLLFVITLPFWAVDAVANWYFGKFLCKAVHVIYTVNLYSSVLIL
AFISLDRYLAIVHATNSQKPRKLLAEKVVYVGVWIPALLLTIPDFIFANVREAEGRYICD
RFYPSDLWVVVFQFQHIMVGLILPGIVILSCYCIIISKLSHSKGHQKRKALKTTVILILA
FFACWLPYYIGISIDSFILLEIIKQGCEFENTVHKWISITEALAFFHCCLNPILYAFLGA
KFKTSAQHALTSVSRGSSLKILSKGKRGGHSSVSTESES
>AAF37288.1 CXCR4 receptor, partial [Saimiri boliviensis]
EEMGSGDYDSMKEPCFREENAHFNRIFLPTIYSIIFLTGIVGNGLVILVMGYQKKLRSMT
DKYRLHLSVADLLFVITLPFWAVDAVANWYFGKFLCKAVHVIYTVNLYSSVLILAFISLD
RYLAIVHATNSQRPRKLLAEKVVYVGVWIPALLLTIPDFIFANVSEADDRYICDRFYPND
LWVVVFQFQHIMVGLILPGIVILSCYCIIISKLSHSKGHQKRKALKTTVILILAFFACWL
PYYIGISIDSFILLEIIRQGCEFENTVHKWISITEALAFFHCCLNPILYAFLGAKFKTSA
QHALTSVSRRSNLKILSKGKRGGHSSVSTESESS
>EFB23364.1 hypothetical protein PANDA_017529, partial [Ailuropoda melanoleuca]
MSIPLPLLQIYPSDNYTEDDLGSGDYDSMKEPCFREENAHFNRIFLPTVYSIIFLTGIVG
NGLVILVMGYQKKLRSMTDKYRLHLSVADLLFVLTLPFWAVDAVANWYFGKFLCKAVHVI
YTVNLYSSVLILAFISLDRYLAIVHATNSQRPRKLLAEKVVYVGVWIPALLLTIPDFIFA
NVREADGRYICDRFYPNDSWLVVFQFQHIMVGLILPGIVILSCYCIIISKLSHSKGYQKR
KALKTTVILILAFFACWLPYYIGISIDSFILLEIIKQGCEFESTVHKWISITEALAFFHC
CLNPILYAFLGAKFKTSAQHALTSVSRGSSLKILSKGKRGGHSSVSTESESSSFHSS
```

Problem 3:

Use the FASTA format of the sequences from problem 2 and make a multiple sequence alignment and a phylogenetic tree with the Clustal Omega web interface: http://www.ebi.ac.uk/Tools/msa/clustalo/ (http://www.ebi.ac.uk/Tools/msa/clustalo/).