BIOINFORMATICS SESSION 7. PRACTICE

2023-10-16

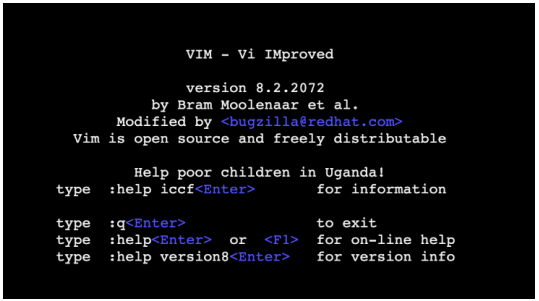
Evolution: What makes us human?

Basic Shell Commands

```
$ cd 2023123456_hyunwoo
$ mkdir week7_7
$ cd week7_7
```

Vim editor - Vim: text editor for Linux

```
$ ll
$ vi print_script.py
```



[In Vim]
i: insert

```
print(“Today is Wednesday!”)
```

[In Vim] “ESC”
:wq save and quit

Basic Shell Commands

```
$ ll
$ python print_script.py
```

[Output] Today is Wednesday!

```
$ less print_script.py
$ cat print_script.py
```

[Output] print “Today is Wednesday!”

Basic Shell Commands

```
$ mv print_script.py new_script.py
$ cp new_script.py new_script_copy.py
$ ll
```

mv : move file
cp : copy and paste

```
$ rm new_script_copy.py
$ ll
$ ln –s /path/to/file/filename
$ unlink filename
```

rm : remove
do not use “rm” for symbolic-linked files
please use “unlink”

Comparing FOXP2 in different animals

FOXP2_HUMAN	*Homo sapiens*
FOXP2_GORGO	*Gorilla gorilla*
FOXP2_MACMU	*Macaca mulatta (Rhesus macaque)*
FOXP2_PANTR	*Pan troglodytes (Chimpanzee)*
FOXP2_HYLLA	*Hylobates lar (Common gibbon)*
FOXP2_PONPY	*Pongo pygmaeus (Bornean orangutan)*
FOXP2_MOUSE	*Mus musculus (Mouse)*
FOXP2_XENLA	*Xenopus laevis (African clawed frog)*

```
$ ln -s /home/biguser/tutor/session6/swissprot.* .
```

Comparing FOXP2 in different animals

```
[biguser@R440 session7]$ cat seqids.txt
FOXP2_HUMAN
FOXP2_GORGO
FOXP2_MACMU
FOXP2_PANTR
FOXP2_HYLLA
FOXP2_PONPY
FOXP2_MOUSE
FOXP2_XENLA
```

Comparing FOXP2 in different animals

```
$ blastdbcmd -entry FOXP2_HYLLA -db swissprot -long_seqids  > FOXP2_HYLLA.fa
```

└ retrieving a single seqeucne with id containing FOXP2_HYLLA

```
$ blastdbcmd -entry_batch seqids.txt -db swissprot -long_seqids  > foxp2.fa
```

└ retrieving multiple seqeucnes with a list of identifiers

```
[biguser@R440 session7]$ blastdbcmd -entry_batch seqids.txt -db swissprot -long_seqids  > foxp2.fa
[biguser@R440 session7]$ ll
total 24
-rw-rw-r-- 1 biguser biguser 6518 Oct 10 16:18 foxp2.fa
-rw-r--r-- 1 biguser biguser 5278 Oct  9 18:43 foxp2.fasta
-rw-rw-r-- 1 biguser biguser  813 Oct  9 18:47 FOXP2_HYLLA.fa
-rw-r--r-- 1 biguser biguser   96 Oct  9 18:43 seqids.txt
lrwxrwxrwx 1 biguser biguser   21 Oct  9 18:47 swissprot -> ../session6/swissprot
lrwxrwxrwx 1 biguser biguser   25 Oct  9 18:47 swissprot.phr -> ../session6/swissprot.phr
lrwxrwxrwx 1 biguser biguser   25 Oct  9 18:47 swissprot.pin -> ../session6/swissprot.pin
lrwxrwxrwx 1 biguser biguser   25 Oct  9 18:47 swissprot.pog -> ../session6/swissprot.pog
lrwxrwxrwx 1 biguser biguser   25 Oct  9 18:47 swissprot.psd -> ../session6/swissprot.psd
lrwxrwxrwx 1 biguser biguser   25 Oct  9 18:47 swissprot.psi -> ../session6/swissprot.psi
lrwxrwxrwx 1 biguser biguser   25 Oct  9 18:47 swissprot.psq -> ../session6/swissprot.psq
```

Comparing FOXP2 in different animals

```
$ clustalw2 foxp2.fa
```

```
[biguser@R440 session7]$ clustalw2 foxp2.fa


 CLUSTAL 2.1 Multiple Sequence Alignments


Sequence format is Pearson
Sequence 1: sp|O15409|FOXP2_HUMAN   715 aa
Sequence 2: sp|Q8MJ99|FOXP2_GORGO   713 aa
Sequence 3: sp|Q8MJ97|FOXP2_MACMU   714 aa
Sequence 4: sp|Q8MJA0|FOXP2_PANTR   716 aa
Sequence 5: sp|Q5QL03|FOXP2_HYLLA   713 aa
Sequence 6: sp|Q8MJ98|FOXP2_PONPY   713 aa
Sequence 7: sp|P58463|FOXP2_MOUSE   714 aa
Sequence 8: sp|Q4VYS1|FOXP2_XENLA   706 aa
```

```
Start of Pairwise alignments
Aligning...

Sequences (1:2) Aligned. Score:  99
Sequences (1:3) Aligned. Score:  99
Sequences (1:4) Aligned. Score:  99
Sequences (1:5) Aligned. Score:  99
Sequences (1:6) Aligned. Score:  99
Sequences (1:7) Aligned. Score:  99
Sequences (1:8) Aligned. Score:  95
Sequences (2:3) Aligned. Score:  100
Sequences (2:4) Aligned. Score:  100
Sequences (2:5) Aligned. Score:  98
Sequences (2:6) Aligned. Score:  98
Sequences (2:7) Aligned. Score:  99
Sequences (2:8) Aligned. Score:  95
Sequences (3:4) Aligned. Score:  100
Sequences (3:5) Aligned. Score:  100
Sequences (3:6) Aligned. Score:  99
Sequences (3:7) Aligned. Score:  98
Sequences (3:8) Aligned. Score:  95
Sequences (4:5) Aligned. Score:  100
Sequences (4:6) Aligned. Score:  99
Sequences (4:7) Aligned. Score:  99
Sequences (4:8) Aligned. Score:  95
Sequences (5:6) Aligned. Score:  99
Sequences (5:7) Aligned. Score:  99
Sequences (5:8) Aligned. Score:  95
Sequences (6:7) Aligned. Score:  99
Sequences (6:8) Aligned. Score:  95
Sequences (7:8) Aligned. Score:  95
Guide tree file created:   [foxp2.dnd]
```

```
There are 7 groups
Start of Multiple Alignment

Aligning...
Group 1: Sequences:   2      Score:15411
Group 2: Sequences:   3      Score:15416
Group 3: Sequences:   4      Score:15408
Group 4: Sequences:   2      Score:15421
Group 5: Sequences:   3      Score:15400
Group 6: Sequences:   7      Score:15379
Group 7: Sequences:   8      Score:14978
Alignment Score 118874

CLUSTAL-Alignment file created  [foxp2.aln]
```

```
-rw-rw-r-- 1 biguser biguser 10280 Oct 10 17:03 foxp2.aln
-rw-rw-r-- 1 biguser biguser   313 Oct 10 17:03 foxp2.dnd
```

Comparing FOXP2 in different animals

foxp2_sequences.aln

```
CLUSTAL 2.1 multiple sequence alignment


sp|Q8MJ99|FOXP2_GORGO       MMQESATETISNSSMNQNGMSTLSSQLDAGSRDGRSSGDTSSEVSTVELL
sp|Q8MJ97|FOXP2_MACMU       MMQESATETISNSSMNQNGMSTLSSQLDAGSRDGRSSGDTSSEVSTVELL
sp|Q8MJA0|FOXP2_PANTR       MMQESATETISNSSMNQNGMSTLSSQLDAGSRDGRSSGDTSSEVSTVELL
sp|O15409|FOXP2_HUMAN       MMQESATETISNSSMNQNGMSTLSSQLDAGSRDGRSSGDTSSEVSTVELL
sp|Q5QL03|FOXP2_HYLLA       MMQESATETISNSSMNQNGMSTLSSQLDAGSRDGRSSGDTSSEVSTVELL
sp|Q8MJ98|FOXP2_PONPY       MMQESVTETISNSSMNQNGMSTLSSQLDAGSRDGRSSGDTSSEVSTVELL
sp|P58463|FOXP2_MOUSE       MMQESATETISNSSMNQNGMSTLSSQLDAGSRDGRSSGDTSSEVSTVELL
sp|Q4VYS1|FOXP2_XENLA       MMQESATETISNSSMNQNGMSTLSSQLDAGSRDGRSSSDTSSEVSTVELL
                            *****.*****************************.************

sp|Q8MJ99|FOXP2_GORGO       HLQQQQALQAARQLLLQQQTSGLKSPKSSDKQRPLQVPVSVAMMTPQVIT
sp|Q8MJ97|FOXP2_MACMU       HLQQQQALQAARQLLLQQQTSGLKSPKSSDKQRPLQVPVSVAMMTPQVIT
sp|Q8MJA0|FOXP2_PANTR       HLQQQQALQAARQLLLQQQTSGLKSPKSSDKQRPLQVPVSVAMMTPQVIT
sp|O15409|FOXP2_HUMAN       HLQQQQALQAARQLLLQQQTSGLKSPKSSDKQRPLQVPVSVAMMTPQVIT
sp|Q5QL03|FOXP2_HYLLA       HLQQQQALQAARQLLLQQQTSGLKSPKSSDKQRPLQVPVSVAMMTPQVIT
sp|Q8MJ98|FOXP2_PONPY       HLQQQQALQAARQLLLQQQTSGLKSPKSSDKQRPLQVPVSVAMMTPQVIT
sp|P58463|FOXP2_MOUSE       HLQQQQALQAARQLLLQQQTSGLKSPKSSEKQRPLQVPVSVAMMTPQVIT
sp|Q4VYS1|FOXP2_XENLA       HLQQQQALQAARQLLLQQQTSGLKSPKNNEKQRPLQVPVSMAMMTPQVIT
                            *************************..:**********:*********

sp|Q8MJ99|FOXP2_GORGO       PQQMQQILQQQVLSPQQLQALLQQQQAVMLQQQQLQEFYKKQQEQLHLQL
sp|Q8MJ97|FOXP2_MACMU       PQQMQQILQQQVLSPQQLQALLQQQQAVMLQQQQLQEFYKKQQEQLHLQL
sp|Q8MJA0|FOXP2_PANTR       PQQMQQILQQQVLSPQQLQALLQQQQAVMLQQQQLQEFYKKQQEQLHLQL
sp|O15409|FOXP2_HUMAN       PQQMQQILQQQVLSPQQLQALLQQQQAVMLQQQQLQEFYKKQQEQLHLQL
sp|Q5QL03|FOXP2_HYLLA       PQQMQQILQQQVLSPQQLQALLQQQQAVMLQQQQLQEFYKKQQEQLHLQL
sp|Q8MJ98|FOXP2_PONPY       PQQMQQILQQQVLSPQQLQALLQQQQAVMLQQQQLQEFYKKQQEQLHLQL
sp|P58463|FOXP2_MOUSE       PQQMQQILQQQVLSPQQLQALLQQQQAVMLQQQQLQEFYKKQQEQLHLQL
sp|Q4VYS1|FOXP2_XENLA       PQQMQQILQQQVLSPQQLQALLQQQQAVMLQQQQLQEFYKKQQEQLHLQL
                            **************************************************
```

#dendrogram
foxp2 _sequences.dnd

```
(
(
(
sp|O15409|FOXP2_HUMAN:0.00122,
(
(
sp|Q8MJ99|FOXP2_GORGO:0.00117,
sp|Q8MJ97|FOXP2_MACMU:-0.00117)
:0.00196,
sp|Q8MJA0|FOXP2_PANTR:-0.00196)
:0.00157)
:0.00059,
sp|Q4VYS1|FOXP2_XENLA:0.04051)
:0.00083,
(
sp|Q5QL03|FOXP2_HYLLA:0.00000,
sp|Q8MJ98|FOXP2_PONPY:0.00140)
:0.00066,
sp|P58463|FOXP2_MOUSE:0.00074);
```

```
*  -- all residues or nucleotides in that column are identical
:  -- conserved substitutions have been observed
.  -- semi-conserved substitutions have been observed
   -- no match.
```

Identification of mutation specific to human

```
$ clustalw2 foxp2.fa –output=fasta
$ less foxp2.fasta
```

```
>sp|Q8MJ99|FOXP2_GORGO
MMQESATETISNSSMNQNGMSTLSSQLDAGSRDGRSSGDTSSEVSTVELL
HLQQQQALQAARQLLLQQQTSGLKSPKSSDKQRPLQVPVSVAMMTPQVIT
PQQMQQILQQQVLSPQQLQALLQQQQAVMLQQQQLQEFYKKQQEQLHLQL
LQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQ---HPGKQAKE
QQQQQQQQQQLAAQQLVFQQQLLQMQQLQQQQHLLSLQRQGLISIPPGQA
ALPVQSLPQAGLSPAEIQQLWKEVTGVHSMEDNGIKHGGLDLTTNNSSST
TSSTTSKASPPITHHSIVNGQSSVLNARRDSSSHEETGASHTLYGHGVCK
WPGCESICEDFGQFLKHLNNEHALDDRSTAQCRVQMQVVQQLEIQLSKER
ERLQAMMTHLHMRPSEPKPSPKPLNLVSSVTMSKNMLETSPQSLPQTPTT
PTAPVTPITQGPSVITPASVPNVGAIRRRHSDKYNIPMSSEIAPNYEFYK
NADVRPPFTYATLIRQAIMESSDRQLTLNEIYSWFTRTFAYFRRNAATWK
NAVRHNLSLHKCFVRVENVKGAVWTVDEVEYQKRRSQKITGSPTLVKNIP
TSLGYGAALNASLQAALAESSLPLLSNPGLINNASSGLLQAVHEDLNGSL
DHIDSNGNSSPGCSPQPHIHSIHVKEEPVIAEDEDCPMSLVTTANHSPEL
EDDREIEEEPLSEDLE
```

만약에 linux 상에서 작업할 수 없거나 결과를 만들지 못했으면

```
ln –s /home/biguser/tutor/session7/foxp2.fasta .
```

foxp2.fasta 파일 링크가 형성됐는지 확인하세요!

Identification of mutation specific to human

- foxp2.ipynb

```
import re

inFile = open('/home/biguser/your_directory/session7/foxp2.fasta','r')

nonhuman = dict()

id = ''
seq = ''

for line in inFile.readlines():
    line = line.strip()
    #print line
    if re.search('^>', line):
        if id != '':
            if not "HUMAN" in id:
                nonhuman[id[1:]] = seq
            else:
                id_human = id[1:]
                seq_human = seq
            id = line
            seq = ''
        else:
            id = line
    else:
        seq += line

if not "HUMAN" in id:
    nonhuman[id[1:]] = seq
else:
    id_human = id[1:]
    seq_human = seq

inFile.close()
```

	POSITION 1	POSITION 2
HUMAN	A	T
GORGO	A	A
MACMU	A	C
PANTR	A	C
HYLLA	A	A
PONPY	A	G

```
for i in range(0,len(seq_human)):
    unique = True
    for id in nonhuman.keys():
        human_seq_posi= seq_human[i]
        nonhuman_seq_posi= nonhuman[id][i]
        if human_seq_posi == nonhuman_seq_posi:
            unique = False
    if unique:
        pos = i + 1
        print("At position ", pos)
        aa = human_seq_posi
        print(id_human, 'Wt', aa)
        for ID in nonhuman.keys():
            print(ID,'Wt', nonhuman[ID][i])
```

Identification of mutation specific to human

- results

```
At position  304
sp|O15409|FOXP2_HUMAN    N
sp|Q8MJ99|FOXP2_GORGO    T
sp|Q8MJ97|FOXP2_MACMU    T
sp|Q8MJA0|FOXP2_PANTR    T
sp|Q5QL03|FOXP2_HYLLA    T
sp|Q8MJ98|FOXP2_PONPY    T
sp|P58463|FOXP2_MOUSE    T
sp|Q4VYS1|FOXP2_XENLA    T
```

Human having Asparagine(N)
whereas all non-human sequences have threonine(T)

not equal because every sequence has different length
thus, gaps are inserted

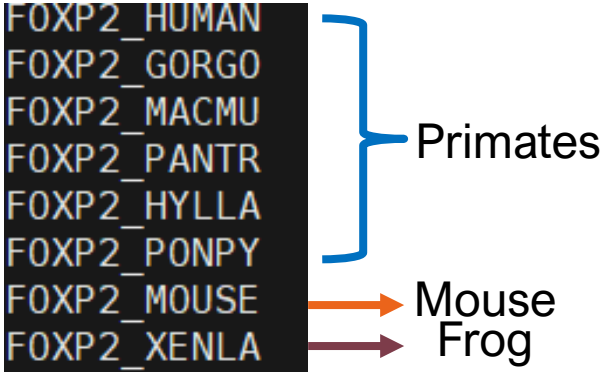
Human-specific T303N

- T303N and N325S are positively selected for language and speech
- No N325S in our res. due to *Xenopus*

Exercise

- Modify foxp2.ipynb such that you may identify

1. positions where the human sequence is different from all other primates and mouse;
2. positions where the human and primate sequences are identical but different to both mouse and frog.



Exercise -1

```
At position  304
sp|O15409|FOXP2_HUMAN    N
sp|Q8MJ99|FOXP2_GORGO    T
sp|Q8MJ97|FOXP2_MACMU    T
sp|Q8MJA0|FOXP2_PANTR    T
sp|Q5QL03|FOXP2_HYLLA    T
sp|Q8MJ98|FOXP2_PONPY    T
sp|P58463|FOXP2_MOUSE    T
At position  326
sp|O15409|FOXP2_HUMAN    S
sp|Q8MJ99|FOXP2_GORGO    N
sp|Q8MJ97|FOXP2_MACMU    N
sp|Q8MJA0|FOXP2_PANTR    N
sp|Q5QL03|FOXP2_HYLLA    N
sp|Q8MJ98|FOXP2_PONPY    N
sp|P58463|FOXP2_MOUSE    N
```

the first part of the code is same as foxp2.ipynb

```
for i in range(0,len(seq_human)):
    unique = True
    for id in nonhuman.keys():
        if not "XENLA" in id:
            human_seq_posi= seq_human[i]
            nonhuman_seq_posi= nonhuman[id][i]
            if human_seq_posi == nonhuman_seq_posi:
                unique = False
    if unique:
        pos = i + 1
        print("At position ", pos)
        aa = human_seq_posi
        print(id_human, '₩t', aa)
        for ID in nonhuman.keys():
            if not "XENLA" in ID:
                print(ID,'₩t', nonhuman[ID][i])
```

Exercise-2

```
At position  80
sp|O15409|FOXP2_HUMAN    D
sp|Q8MJ99|FOXP2_GORGO    D
sp|Q8MJ97|FOXP2_MACMU    D
sp|Q8MJA0|FOXP2_PANTR    D
sp|Q5QL03|FOXP2_HYLLA    D
sp|Q8MJ98|FOXP2_PONPY    D
sp|P58463|FOXP2_MOUSE    E
sp|Q4VYS1|FOXP2_XENLA    E
```

```python
import re


for i in range(0,len(seq_human)):
    unique = True
    for id in nonhuman.keys():
        human_seq_posi= seq_human[i]
        if "XENLA" in id or "MOUSE" in id:
            if human_seq_posi == nonhuman[id][i]:
                unique=False
        else:
            if human_seq_posi != nonhuman[id][i]:
                unique=False
    if unique:
        pos = i + 1
        print("At position ", pos)
        aa = human_seq_posi
        print(id_human, '₩t', aa)
        for ID in nonhuman.keys():
            print(ID,'₩t', nonhuman[ID][i])
```

Assignment

- In the default output from the ClustalW program (a file named as foxp2.aln) there are asterisks (*) that indicate positions where the sequence is same in all sequences. Make a Python script to count the total length and number of such positions in all alignment.
- Output
- clustalw의 결과 중 foxp2.aln을 이용해서 algin된 총 길이와 seqeunce들이 모두 동일함을 뜻하는 * 표시가 총 몇 개인지 아래와 같이 print.

```
Total length :
Number of asterisk (match) :
```

- 과제 제출 기한: 10/22 Sunday 23:59 @ LMS
- 작성한 코드와 해당 코드의 결과를 캡처한 뒤 워드에 첨부(코드만 긁어와서 붙여넣지 말기), 코드에 대한 설명 간략히 작성
  워드 파일명은 n주차_학번_이름 형식으로 제출(e.g. 7주차_2023123456_김현우)

Assignment

`ln -s/home/biguser/tutor/Week7/foxp2.aln .`
를 linux환경에서 여러분의 directory에서 실행시키고 해당 파일 link가 형성 됐는지 확인. 해당 파일을 읽어서 작업할 것

```
CLUSTAL 2.1 multiple sequence alignment


gi|51701430|sp|Q8MJ99.1|FOXP2_      MMQESATETISNSSMNQNGMSTLSSQLDAGSRDGRSSGDTSSEVSTVELL
gi|51701429|sp|Q8MJ97.1|FOXP2_      MMQESATETISNSSMNQNGMSTLSSQLDAGSRDGRSSGDTSSEVSTVELL
gi|38503046|sp|Q8MJA0.1|FOXP2_      MMQESATETISNSSMNQNGMSTLSSQLDAGSRDGRSSGDTSSEVSTVELL
gi|17432967|sp|O15409.2|FOXP2_      MMQESATETISNSSMNQNGMSTLSSQLDAGSRDGRSSGDTSSEVSTVELL
gi|62286912|sp|Q5QL03.1|FOXP2_      MMQESATETISNSSMNQNGMSTLSSQLDAGSRDGRSSGDTSSEVSTVELL
gi|146345420|sp|Q8MJ98.3|FOXP2      MMQESVTETISNSSMNQNGMSTLSSQLDAGSRDGRSSGDTSSEVSTVELL
gi|51704301|sp|P58463.2|FOXP2_      MMQESATETISNSSMNQNGMSTLSSQLDAGSRDGRSSGDTSSEVSTVELL
gi|82227296|sp|Q4VYS1.1|FOXP2_      MMQESATETISNSSMNQNGMSTLSSQLDAGSRDGRSSSDTSSEVSTVELL
                                    *****.*******************************.************


gi|51701430|sp|Q8MJ99.1|FOXP2_      HLQQQQALQAARQLLLQQQTSGLKSPKSSDKQRPLQVPVSVAMMTPQVIT
gi|51701429|sp|Q8MJ97.1|FOXP2_      HLQQQQALQAARQLLLQQQTSGLKSPKSSDKQRPLQVPVSVAMMTPQVIT
gi|38503046|sp|Q8MJA0.1|FOXP2_      HLQQQQALQAARQLLLQQQTSGLKSPKSSDKQRPLQVPVSVAMMTPQVIT
gi|17432967|sp|O15409.2|FOXP2_      HLQQQQALQAARQLLLQQQTSGLKSPKSSDKQRPLQVPVSVAMMTPQVIT
gi|62286912|sp|Q5QL03.1|FOXP2_      HLQQQQALQAARQLLLQQQTSGLKSPKSSDKQRPLQVPVSVAMMTPQVIT
gi|146345420|sp|Q8MJ98.3|FOXP2      HLQQQQALQAARQLLLQQQTSGLKSPKSSDKQRPLQVPVSVAMMTPQVIT
gi|51704301|sp|P58463.2|FOXP2_      HLQQQQALQAARQLLLQQQTSGLKSPKSSEKQRPLQVPVSVAMMTPQVIT
gi|82227296|sp|Q4VYS1.1|FOXP2_      HLQQQQALQAARQLLLQQQTSGLKSPKNNEKQRPLQVPVSMAMMTPQVIT
                                    **************************..:**********;*********
```