cffi_basic_avium

July 22, 2022

[ ]:
```
import cffi
import os
import numpy as np
import pandas as pd
from io import StringIO
from enum import Enum
from Bio import SeqIO
from pathlib import Path
from auxiliary import DATA_SEQ_DIR
from pytrsomix import TRScalculator, TRSanalyzer, AlignmentAnalyzer
```

1 Reading in the genomes and calculating the interiors accordinf the trs.txt file

[ ]:
```
trs_file = (DATA_SEQ_DIR/"avium"/"trs.txt").absolute().as_posix().encode()
trs1 = TRScalculator(sequence=(DATA_SEQ_DIR/"avium"/"avium subsp. avium strain␣
 ↪DSM 44156.fasta").absolute().as_posix().encode(), trs=trs_file, tmin=2000,␣
 ↪tmax=3000)
trs1.calculate()

trs2 = TRScalculator(sequence=(DATA_SEQ_DIR/"avium"/"avium subsp. hominissuis␣
 ↪strain H87.fasta").absolute().as_posix().encode(), trs=trs_file, tmin=2000,␣
 ↪tmax=3000)
trs2.calculate()

trs3 = TRScalculator(sequence=(DATA_SEQ_DIR/"avium"/"avium subsp.␣
 ↪paratuberculosis strain DSM 44135.fasta").absolute().as_posix().encode(),␣
 ↪trs=trs_file, tmin=2000, tmax=3000)
trs3.calculate()
```

```
name of genome file: /home/rafalb/molecules/TRS-omix/TRS-omix/data/avium/avium
subsp. avium strain DSM 44156.fasta
name of input file: /home/rafalb/molecules/TRS-omix/TRS-omix/data/avium/trs.txt
name of output file: interiors.txt
```

```
tmin: 2000
tmax: 3000
mode: 0

START

size of genome: 4956929
size of input: 9
status after LC_TRSPositionsFindAndSaveToVLt: 1
status after LC_InteriorsFindAndSaveToFile: 1
END
name of genome file: /home/rafalb/molecules/TRS-omix/TRS-omix/data/avium/avium
subsp. hominissuis strain H87.fasta
name of input file: /home/rafalb/molecules/TRS-omix/TRS-omix/data/avium/trs.txt
name of output file: interiors.txt
tmin: 2000
tmax: 3000
mode: 0

START

size of genome: 5626623
size of input: 9
status after LC_TRSPositionsFindAndSaveToVLt: 1
status after LC_InteriorsFindAndSaveToFile: 1
END
name of genome file: /home/rafalb/molecules/TRS-omix/TRS-omix/data/avium/avium
subsp. paratuberculosis strain DSM 44135.fasta
name of input file: /home/rafalb/molecules/TRS-omix/TRS-omix/data/avium/trs.txt
name of output file: interiors.txt
tmin: 2000
tmax: 3000
mode: 0

START

size of genome: 4839032
size of input: 9
status after LC_TRSPositionsFindAndSaveToVLt: 1
status after LC_InteriorsFindAndSaveToFile: 1
END
```

2 Instantiating the SequenceAnalyzer object

[ ]:
```
sa = TRSanalyzer.SeqAnalyzer([trs1.Result, trs2.Result, trs3.Result])
sa.Combined
```

[ ]:

	L-NoClass	L-No	LFS	Len(LFS)	L-POS(LFS)	R-POS(LFS)	\
0	7	20	CGACGACGA	9	3206	3214	
1	4	10	GGTGGTGGT	9	27822	27830	
2	8	22	CGTCGTCGT	9	35487	35495	
3	2	4	CGGCGGCGG	9	51835	51843	
4	3	8	CCACCACCA	9	57028	57036	
..	…	…	…	…	…	…	
926	2	4	CGGCGGCGGCGG	12	4745703	4745714	
927	3	8	CCACCACCA	9	4776478	4776486	
928	1	1	CCGCCGCCG	9	4788563	4788571	
929	2	6	GCGGCGGCG	9	4818519	4818527	
930	2	4	CGGCGGCGG	9	4827232	4827240	

	R-NoClass	R-No	RFS	Len(RFS)	L-POS(RFS)	R-POS(RFS)	\
0	13	38	AGAAGAAGA	9	5268	5276	
1	1	3	GCCGCCGCCGCC	12	29917	29928	
2	2	4	CGGCGGCGG	9	37498	37506	
3	2	4	CGGCGGCGG	9	54351	54359	
4	3	7	ACCACCACC	9	59422	59430	
..	…	…	…	…	…	…	
926	10	28	GCTGCTGCT	9	4748185	4748193	
927	7	20	CGACGACGA	9	4778776	4778784	
928	1	2	CGCCGCCGC	9	4790624	4790632	
929	18	52	GATGATGAT	9	4820807	4820815	
930	3	7	ACCACCACC	9	4829437	4829445	

	>SEQ	Len(SEQ)	\
0	>CAGCCCGCCGAGCGGCAGCGGGCCGTTCAGCGCGCTGCCCACCGA…	2053	
1	>GCAGCTCGTGTGCGGTGTGCACCGCGACCCAGCGCAGCGAGCGCT…	2086	
2	>CGCCGAGAACGGTCCACGACTCACGCAACGAGACCGGCGAGATAA…	2002	
3	>CCGCCACCCGATCCAGCTCGGCGCGCAGCTCGGGCTCGGCCAGCA…	2507	
4	>GGTTGGTGACCAGGTAGATCAGCACCAGCACCGTCACGATGGACA…	2385	
..	…	…	
926	>CAAACTTGAGTTCACCCTCATTGGTGACGCCGTCAACGTTGCGGC…	2470	
927	>CCGCCGAAGGCCCGGTGCGCCCGGTGAGTTCGTCCAGCGTCCAGC…	2289	
928	>CGGGTCCGGTAAACGTCGGCGCCGGCGGCCGGCGGACCCGGCATC…	2052	
929	>GTCAGCAGCTCGGCATCCTGGGCGGCGTCAGGCACGTTCGATCAG…	2279	
930	>CCTCAACAAGGACGAGCTGGGCCTCAATGGGCCCTCGTCGTCCAC…	2196	

	GENOME
0	NZ_CP046507.1
1	NZ_CP046507.1

```
2    NZ_CP046507.1
3    NZ_CP046507.1
4    NZ_CP046507.1
..             …
926  NZ_CP053068.1
927  NZ_CP053068.1
928  NZ_CP053068.1
929  NZ_CP053068.1
930  NZ_CP053068.1

[931 rows x 15 columns]
```

2.1 Unique gnomes in the table

[ ]:
```
sa.Combined["GENOME"].unique()
```

[ ]:
```
array(['NZ_CP046507.1', 'NZ_CP018363.1', 'NZ_CP053068.1'], dtype=object)
```

2.2 Calculating the Needleman-Wunsch alignment scores with respect to chosen sequence

[ ]:
```
algns = sa.calculate_all_alignments(0)
```

2.3 10 most similar scores

- the first one in the similarity to itself (the highest possible score here…)

[ ]:
```
aa = AlignmentAnalyzer(algns)
most_similar = aa.get_sorted_scores().sort_values("score", ascending=False)[:10]
most_similar
```

[ ]:
```
       score
index
0      13501
281    13450
707    10137
813    10116
78     10104
740    10101
489    10073
74     10072
152    10052
699    10036
```

2.4 10 most similar sequences

[ ]:
```
sa.Combined.loc[most_similar.index, :]
```

[ ]:

index	L-NoClass	L-No	LFS	Len(LFS)	L-POS(LFS)	R-POS(LFS)
0	7	20	CGACGACGA	9	3206	3214
281	7	20	CGACGACGA	9	3205	3213
707	18	52	GATGATGAT	9	1074679	1074687
813	1	1	CCGCCGCCG	9	2854176	2854184
78	2	4	CGGCGGCGG	9	1416328	1416336
740	4	10	GGTGGTGGT	9	1621620	1621628
489	9	27	CAGCAGCAG	9	3072968	3072976
74	4	12	TGGTGGTGG	9	1314096	1314104
152	9	27	CAGCAGCAG	9	2533720	2533728
699	9	25	AGCAGCAGC	9	993356	993364

index	R-NoClass	R-No	RFS	Len(RFS)	L-POS(RFS)	R-POS(RFS)
0	13	38	AGAAGAAGA	9	5268	5276
281	13	38	AGAAGAAGA	9	5267	5275
707	8	22	CGTCGTCGT	9	1077673	1077681
813	2	4	CGGCGGCGG	9	2857172	2857180
78	2	6	GCGGCGGCG	9	1419312	1419320
740	12	34	GTTGTTGTT	9	1624606	1624614
489	7	20	CGACGACGA	9	3075943	3075951
74	1	2	CGCCGCCGC	9	1316923	1316931
152	7	20	CGACGACGA	9	2536695	2536703
699	1	3	GCCGCCGCC	9	996319	996327

index	>SEQ	Len(SEQ)
0	>CAGCCCGCCGAGCGGCAGCGGGCCGTTCAGCGCGCTGCCCACCGA…	2053
281	>CAGCCCGCCGAGCGGCAGCGGGCCGTTCAGCGCGCTGCCCACCGA…	2053
707	>GTTGGGCGGGTCGCCGACCAACGTCGCCGCGCCGCCGACGTTCGA…	2985
813	>CGCACCCCCAGGCCGTGCGCCTCGTCGACGATCAGCAGGGCCCGG…	2987
78	>AGCTATCGGTGTGGCCGCCGGCGGATGCCGAGGCGGTCGACGTGG…	2975
740	>GCCGTCCAACCCGGCCGACGCGTACTGGCTGCTGCGCCACGCCAT…	2977
489	>GTCCCGGGACCGGCGGCCGTCGGCCACCGCCCGCAACGCGCTCAC…	2966
74	>CCAGCCCGCCGCAGGCCCAGTTCGACGGGCTGCCGCATCCGACCA…	2818
152	>GTCCCGGGACCGGCGGCCGTCGGCCACCGCCCGCAACGCGCTCAC…	2966
699	>TCCGCGTCGGCCCAGACCTGTTCGGCGGTGTCGGCCAGTTGGGCG…	2954

index	GENOME
0	NZ_CP046507.1
281	NZ_CP018363.1

```
707    NZ_CP053068.1
813    NZ_CP053068.1
78     NZ_CP046507.1
740    NZ_CP053068.1
489    NZ_CP018363.1
74     NZ_CP046507.1
152    NZ_CP046507.1
699    NZ_CP053068.1
```

[ ]: