Raport

Struktura rozwiązania

Struktura plików: wejściowych, tworzonych pośrednio i wyników wygląda następnująco:

```
├── project.py
└── data
# pliki wejściowe
    ├── e_coli.fa
    ├── proms_e_coli.fa
    ├── query.fa
# podpunkt 1
    ├── db
    │   └── ...
    ├── blast_result.csv
# podpunkt 2
    ├── memeA
    │   └── ...
    ├── memeB
    │   └── ...
    ├── proms_e_coli_A.fa
    ├── proms_e_coli_B.fa
    ├── A.pfm
    ├── B.pfm
# podpunkt 3
    └── final_result.csv
```

Szczegóły implementacji

Rozwiązanie znajduje się w pliku `project.py`

Stuktura wywołań funkcji podąża za kolejnymi podpunktami zadania i wygląda następująco:

```
__main__
├── part01
│   ├── make_blast_database
│   └── perform_local_blast_search
├── part02
│   ├── filter_promoters
│   ├── perform_meme_search
│   └── process_meme_output
└── part03
    ├── parse_motifs
    ├── get_promoter_sequences
    └── compute_enrichments
        ├── count_trials
        └── count_hits
            ├── compute_log_odds
            └── all_windows
```

Podpunkt 1.

Z pomocą wrapperów dostępnych w bibliotece Biopython: - `NcbimakeblastdbCommandline` tworzy bazę danych - `NcbitblastnCommandline` wykonuje lokalne przeszukiwanie BLAST

Aby uwzględnić różnicę w alfabecie bazy danych i przeszukiwania (fragmenty protein są szukane w nukleotydowej bazie danych), zgodnie ze znalezioną dokumentacją wykorzystywany jest program `tblastn`, który przed przeszukiwaniem tłumaczy bazę danych na proteiny.

Następnie wyniki parsowane się do pliku `blast_result.csv` o kolumnach:

- identyfikator fragmentu proteiny
- identyfikator dopasowanego genu E-coli
- e-value przeszukiwania BLAST

(tabela dostępna także na końcu tego pliku).

Niemal wszystkie wartości e-value są rzędu od `10 ^ -100` do `10 ^ -80`, co świadczy o bardzo dobrym dopasowaniu fragmentów.

Podpunkt 2.

- `filter_promoters` dzieli wynikowe dopasowania (id e-coli) na dwa pliki odpowiadające promoterom genów dopasowanych do pragmentów protein z grupy A, i z grupy B.
- `perform_meme_search` wywołuje program MEME

Ponieważ biblioteka Biopython nie posiada wrappera do programu MEME, ręcznie generowana i wywoływana jest komenda:

```
meme {promoters_filename} -dna -oc {out_folder} -nostatus -nmotifs {num_motifs} -w {motif_length}
```

z parametrami `num_motifs = 10`, `motif_length = 15` zgodnie z treścią zadania.

Pełne wyniki działania programu MEME dostępne są pod linkami: grupa A, grupa B.

- `process_meme_output` parsuje, z pomocą modułu `Bio.motifs`, wyniki do plików w formacie JASPAR pfm. (Także dostępne na końcu tego pliku)

Podpunkt 3.

`compute_enrichments` dla każdego z (wszystkich 20) motywów, oblicza liczbę trafień oraz wynik *p-value* testu dwumianowego względem promotorów z grupy A i z grupy B.

Ostateczne wyniki prezentuje tabela:

motif (consensus)	number of hits (A)	number of hits (B)	enrichment p-value (A)	enrichment p-value (B)
GATCAAAATTTGACC	137	56	0.0005082759924008983	0.017236188076922256
TTTGATTTACATCAA	3	0	0.0	0.42276871150603257
CCGCTCCCCCCTTGC	2	0	0.0	0.6341699931406767
CGGGCTTGCAGCCTG	5	0	0.0	0.12519808020472165
GCCGAAATCCCTGGA	10	1	1.928028313707509e-05	0.0520029581906493
GAACACCGCCACGGC	2	0	0.0	0.6341699931406767
GGCTACCTCTGCCGC	2	0	0.0	0.6341699931406767
GTCACTCATCCAACC	2	0	0.0	0.6341699931406767
TGTCACTCTCCCGAA	2	0	0.0	0.6341699931406767
CGGAAACTCGTTGCC	2	0	0.0	0.6341699931406767
ATATTGCCGCAATAT	39	45	3.017622365693389e-07	1.1253063803974388e-05
GGGGCGCAAGGCCCG	3	8	0.000793974832879017	0.00033869213043908247
GGCGGATTTGCGGCC	0	4	0.0012374162921043501	0.0
GAACCAGGCAGACCG	0	2	0.05803233075219762	0.0
TTTTCTTAACCTGAA	0	8	9.932004704971322e-07	0.0
AGCGAAGCAACGAGA	0	2	0.05803233075219762	0.0
GAAAAAGCTTCACCC	0	5	0.0002948201493568265	0.0
TAAGTAAAGCGTGAA	0	2	0.05803233075219762	0.0
CCTCACGGAGAGGGT	0	2	0.05803233075219762	0.0
GCCTCTGAAGTTCAT	0	2	0.05803233075219762	0.0

Wyniki

Tablica `blast_result.csv`

id proteiny	id e-coli	e-value
groupA_0	queA	1.69948e-102
groupA_1	hupA	1.55016e-58
groupA_2	hupB	2.06824e-43
groupA_3	marR	5.18213e-97
groupA_4	nanA	8.19867e-92
groupA_5	acnB	1.0222e-79
groupA_6	proP	2.86771e-85
groupA_7	fadB	4.91179e-86
groupA_8	rplM	3.12687e-99
groupA_9	dmsA	2.11863e-95
groupA_10	narK	1.57537e-79
groupA_11	nirB	3.3309e-100
groupA_12	mazE	2.78059e-53
groupA_13	narG	6.04773e-94
groupA_14	deoC	2.93522e-103
groupA_15	aldB	7.7511e-96
groupA_16	mglA	2.29302e-85
groupA_17	pyrD	2.63261e-90
groupA_18	lpd	1.16991e-96
groupA_19	ndh	5.97922e-99
groupA_20	glnA	3.37715e-97
groupA_21	pflB	3.04718e-100
groupA_22	trg	5.0377e-67
groupA_23	fumB	1.7633e-100
groupA_24	nrfA	3.46468e-102
groupA_25	trmA	2.30632e-103
groupA_26	cbpA	6.79795e-98
groupA_27	nrdA	1.84118e-98
groupA_28	glnQ	3.4015e-92
groupA_29	katE	1.15457e-100
groupA_30	osmE	9.12435e-76
groupA_31	adhE	3.92265e-86

id proteiny	id e-coli	e-value
groupA_32	gyrA	9.03098e-95
groupA_33	ogt	2.39152e-105
groupA_34	gadA	9.20994e-108
groupA_35	msrA	1.25104e-105
groupA_36	hyaA	8.96046e-105
groupA_37	osmY	2.55716e-96
groupA_38	mtlA	2.62213e-90
groupA_39	malE	1.8629e-101
groupA_40	gyrB	1.28377e-98
groupA_41	ptsG	1.95485e-90
groupA_42	xylF	7.34111e-93
groupA_43	glpT	9.12496e-81
groupA_44	crp	1.17456e-97
groupA_45	sra	5.67859e-28
groupA_46	dps	5.70159e-98
groupA_47	tufB	2.52675e-83
groupA_48	pdxA	1.5488e-99
groupA_49	glcC	2.5071e-103
groupA_50	gltX	2.99595e-91
groupA_51	patA	3.79575e-100
groupA_52	dusB	7.40013e-106
groupA_53	acs	1.16718e-96
groupA_54	hns	9.12347e-64
groupA_55	nuoA	4.47428e-96
groupA_56	glpA	3.25389e-94
groupA_57	topA	4.41334e-93
groupA_58	ansB	5.70088e-100
groupA_59	carA	1.50343e-101
groupA_60	cycA	2.58564e-96
groupA_61	fixA	2.70041e-82
groupA_62	cspA	1.01319e-43
groupB_0	kduI	1.40312e-97

id proteiny	id e-coli	e-value
groupB_1	rsmF	3.767e-97
groupB_2	queD	2.93753e-85
groupB_3	ugd	2.80451e-98
groupB_4	gdhA	4.27892e-98
groupB_5	ivy	6.88973e-98
groupB_6	nac	1.08745e-88
groupB_7	codB	2.29695e-89
groupB_8	ligB	4.39904e-100
groupB_9	dld	4.64929e-102
groupB_10	rhtA	1.5413e-88
groupB_11	pphB	6.33574e-110
groupB_12	feaR	5.70929e-103
groupB_13	rfaH	1.40791e-105
groupB_14	ackA	5.93861e-105
groupB_15	prfC	2.7041e-100
groupB_16	rlmF	1.12559e-102
groupB_17	kdgR	2.18738e-106
groupB_18	eamA	3.33807e-87
groupB_19	mdtG	2.70142e-73
groupB_20	greB	1.59626e-105
groupB_21	serA	1.44541e-96
groupB_22	ddlA	1.95111e-96
groupB_23	gabD	9.33046e-100
groupB_24	uspC	1.82623e-99
groupB_25	asnC	5.96953e-105
groupB_26	dcp	1.04928e-100
groupB_27	mtlD	5.64412e-103
groupB_28	wecH	3.55718e-93
groupB_29	abgR	5.67501e-110
groupB_30	fklB	6.78231e-105
groupB_31	sdiA	1.55165e-106
groupB_32	tdcR	5.86086e-83

id proteiny	id e-coli	e-value
groupB_33	rimM	7.11364e-109
groupB_34	ispB	2.06592e-92

Plik `A.pfm`

```
 >None GATCAAAATTTGACC
A [31746.00 682540.00 111111.00 111111.00 809524.00 698413.00 809524.00 746032.00 15873.00 111111.00 63492.00 47619.00 793651.00 317
C [79365.00 111111.00 15873.00 698413.00 31746.00 63492.00 79365.00 47619.00 158730.00 111111.00 111111.00 63492.00 15873.00 825397.
G [825397.00 63492.00 95238.00 95238.00 63492.00 63492.00 47619.00 63492.00 111111.00 95238.00 63492.00 857143.00 142857.00 79365.00
T [63492.00 142857.00 777778.00 95238.00 95238.00 174603.00 63492.00 142857.00 714286.00 682540.00 761905.00 31746.00 47619.00 63492
>None TTTGATTTWCATSAA
A [  0.00   0.00   0.00   0.00 1000000.00   0.00   0.00   0.00 666667.00   0.00 1000000.00   0.00   0.00 1000000.00 1000000.00]
C [  0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00 1000000.00   0.00   0.00 666667.00   0.00   0.00]
G [  0.00   0.00   0.00 1000000.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00 333333.00   0.00   0.00]
T [1000000.00 1000000.00 1000000.00   0.00   0.00 1000000.00 1000000.00 1000000.00 333333.00   0.00   0.00 1000000.00   0.00   0.00
>None SCKCTCCCCYYTTGC
A [  0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00]
C [500000.00 1000000.00   0.00 1000000.00   0.00 1000000.00 1000000.00 1000000.00 1000000.00 500000.00 500000.00   0.00   0.00   0.0
G [500000.00   0.00 500000.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00 1000000.00   0.00]
T [  0.00   0.00 500000.00   0.00 1000000.00   0.00   0.00   0.00   0.00 500000.00 500000.00 1000000.00 1000000.00   0.00   0.00]
>None CGSGNNTSCDGCCKG
A [  0.00   0.00   0.00   0.00 200000.00 200000.00   0.00   0.00   0.00 400000.00   0.00   0.00   0.00   0.00   0.00]
C [1000000.00 200000.00 400000.00 200000.00 400000.00 200000.00 200000.00 400000.00 800000.00   0.00   0.00 1000000.00 1000000.00
G [  0.00 800000.00 600000.00 800000.00 200000.00 200000.00   0.00 600000.00   0.00 200000.00 1000000.00   0.00   0.00 400000.00 100
T [  0.00   0.00   0.00   0.00 200000.00 400000.00 800000.00   0.00 200000.00 400000.00   0.00   0.00   0.00 600000.00   0.00]
>None GVVKAAAHBCCTGGA
A [  0.00 200000.00 300000.00   0.00 500000.00 700000.00 1000000.00 300000.00 100000.00   0.00   0.00 200000.00   0.00   0.00 100000
C [300000.00 400000.00 400000.00 200000.00 200000.00 200000.00   0.00 300000.00 400000.00 1000000.00 1000000.00 100000.00   0.00 100
G [700000.00 400000.00 300000.00 500000.00 200000.00   0.00   0.00   0.00 200000.00   0.00   0.00   0.00 1000000.00 800000.00   0.00
T [  0.00   0.00   0.00 300000.00 100000.00 100000.00   0.00 400000.00 300000.00   0.00   0.00 700000.00   0.00 100000.00   0.00]
>None GRASASCGCCMCGGC
A [  0.00 500000.00 1000000.00   0.00 1000000.00   0.00   0.00   0.00   0.00   0.00 500000.00   0.00   0.00   0.00   0.00]
C [  0.00   0.00   0.00 500000.00   0.00 500000.00 1000000.00   0.00 1000000.00 1000000.00 500000.00 1000000.00   0.00   0.00 100000
G [1000000.00 500000.00   0.00 500000.00   0.00 500000.00   0.00 1000000.00   0.00   0.00   0.00   0.00 1000000.00 1000000.00   0.00
T [  0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00]
>None GGCTAYSTSTKCCGC
A [  0.00   0.00   0.00   0.00 1000000.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00]
C [  0.00   0.00 1000000.00   0.00   0.00 500000.00 500000.00   0.00 500000.00   0.00   0.00 1000000.00 1000000.00   0.00 1000000.00
G [1000000.00 1000000.00   0.00   0.00   0.00   0.00 500000.00   0.00 500000.00   0.00 500000.00   0.00   0.00 1000000.00   0.00]
T [  0.00   0.00   0.00 1000000.00   0.00 500000.00   0.00 1000000.00   0.00 1000000.00 500000.00   0.00   0.00   0.00   0.00]
>None GTCRCTCATCYRMCC
A [  0.00   0.00   0.00 500000.00   0.00   0.00   0.00 1000000.00   0.00   0.00   0.00 500000.00 500000.00   0.00   0.00]
C [  0.00   0.00 1000000.00   0.00 1000000.00   0.00 1000000.00   0.00   0.00 1000000.00 500000.00   0.00 500000.00 1000000.00 10000
G [1000000.00   0.00   0.00 500000.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00 500000.00   0.00   0.00   0.00]
T [  0.00 1000000.00   0.00   0.00   0.00 1000000.00   0.00   0.00 1000000.00   0.00 500000.00   0.00   0.00   0.00   0.00]
>None TGTCRSTCTCCYGAA
A [  0.00   0.00   0.00   0.00 500000.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00 1000000.00 1000000.00]
C [  0.00   0.00   0.00 1000000.00   0.00 500000.00   0.00 1000000.00   0.00 1000000.00 1000000.00 500000.00   0.00   0.00   0.00]
G [  0.00 1000000.00   0.00   0.00 500000.00 500000.00   0.00   0.00   0.00   0.00   0.00   0.00 1000000.00   0.00   0.00]
T [1000000.00   0.00 1000000.00   0.00   0.00   0.00 1000000.00   0.00 1000000.00   0.00   0.00 500000.00   0.00   0.00   0.00]
>None CGKWAAYTCGTTGCC
A [  0.00   0.00   0.00 500000.00 1000000.00 1000000.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00]
C [1000000.00   0.00   0.00   0.00   0.00   0.00 500000.00   0.00 1000000.00   0.00   0.00   0.00   0.00 1000000.00 1000000.00]
G [  0.00 1000000.00 500000.00   0.00   0.00   0.00   0.00   0.00   0.00 1000000.00   0.00   0.00 1000000.00   0.00   0.00]
T [  0.00   0.00 500000.00 500000.00   0.00   0.00 500000.00 1000000.00   0.00   0.00 1000000.00 1000000.00   0.00   0.00   0.00]
```

Plik `B.pfm`

```
 >None ATATTGCCGCAATAT
A [848485.00 151515.00 909091.00 60606.00 60606.00 90909.00 30303.00   0.00 60606.00 30303.00 636364.00 848485.00 60606.00 787879.00
C [  0.00 60606.00   0.00 60606.00   0.00 60606.00 818182.00 818182.00 60606.00 636364.00 60606.00 90909.00   0.00 30303.00 121212.0
G [90909.00 30303.00 90909.00 60606.00 181818.00 818182.00 90909.00 121212.00 878788.00 272727.00 212121.00 60606.00 30303.00 90909.
T [60606.00 757576.00   0.00 818182.00 757576.00 30303.00 60606.00 60606.00   0.00 60606.00 90909.00   0.00 909091.00 90909.00 75757
>None GGBGCGCNWKGCVCG
A [  0.00   0.00   0.00   0.00 125000.00   0.00 125000.00 250000.00 500000.00 125000.00   0.00   0.00 250000.00   0.00   0.00]
C [125000.00 250000.00 250000.00   0.00 625000.00   0.00 625000.00 250000.00 125000.00 125000.00 125000.00 875000.00 375000.00 75000
G [875000.00 625000.00 375000.00 1000000.00   0.00 875000.00   0.00 250000.00   0.00 500000.00 875000.00 125000.00 375000.00 250000.
T [  0.00 125000.00 375000.00   0.00 250000.00 125000.00 250000.00 250000.00 375000.00 250000.00   0.00   0.00   0.00   0.00   0.00]
>None GGCBGATDTGCBGCC
A [  0.00   0.00   0.00   0.00   0.00 1000000.00 250000.00 250000.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00]
C [  0.00   0.00 750000.00 250000.00 250000.00   0.00   0.00   0.00 250000.00   0.00 1000000.00 250000.00   0.00 1000000.00 1000000.
G [750000.00 1000000.00 250000.00 500000.00 750000.00   0.00   0.00 250000.00   0.00 1000000.00   0.00 500000.00 750000.00   0.00
T [250000.00   0.00   0.00 250000.00   0.00   0.00 750000.00 500000.00 750000.00   0.00   0.00 250000.00 250000.00   0.00   0.00]
>None GMACCAGGSWGRYCG
A [  0.00 500000.00 1000000.00   0.00   0.00 1000000.00   0.00   0.00   0.00 500000.00   0.00 500000.00   0.00   0.00   0.00]
C [  0.00 500000.00   0.00 1000000.00 1000000.00   0.00   0.00   0.00 500000.00   0.00   0.00   0.00 500000.00 1000000.00   0.00]
G [1000000.00   0.00   0.00   0.00   0.00   0.00 1000000.00 1000000.00 500000.00   0.00 1000000.00 500000.00   0.00   0.00 1000000.0
T [  0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00 500000.00   0.00   0.00 500000.00   0.00   0.00]
>None TTTTCKYMWCCWGAW
A [  0.00   0.00 250000.00 125000.00 125000.00   0.00   0.00 500000.00 500000.00   0.00 250000.00 375000.00   0.00 1000000.00 500000
C [  0.00   0.00   0.00   0.00 625000.00   0.00 375000.00 500000.00   0.00 750000.00 750000.00   0.00 125000.00   0.00 125000.00]
G [  0.00   0.00   0.00 125000.00   0.00 375000.00   0.00   0.00   0.00 250000.00   0.00   0.00 875000.00   0.00   0.00]
T [1000000.00 1000000.00 750000.00 750000.00 250000.00 625000.00 625000.00   0.00 500000.00   0.00   0.00 625000.00   0.00   0.00 37
>None AGSGMAKCAACGMGA
A [1000000.00   0.00   0.00   0.00 500000.00 1000000.00   0.00   0.00 1000000.00 1000000.00   0.00   0.00 500000.00   0.00 1000000.0
C [  0.00   0.00 500000.00   0.00 500000.00   0.00   0.00 1000000.00   0.00   0.00 1000000.00   0.00 500000.00   0.00   0.00]
G [  0.00 1000000.00 500000.00 1000000.00   0.00   0.00 500000.00   0.00   0.00   0.00   0.00 1000000.00   0.00 1000000.00   0.00]
T [  0.00   0.00   0.00   0.00   0.00   0.00 500000.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00]
>None GAMAAAGHTKYACCC
A [  0.00 600000.00 600000.00 1000000.00 600000.00 1000000.00   0.00 200000.00   0.00   0.00   0.00 800000.00   0.00   0.00   0.00]
C [  0.00 200000.00 400000.00   0.00 200000.00   0.00   0.00 400000.00 200000.00   0.00 600000.00   0.00 800000.00 800000.00 1000000
G [1000000.00 200000.00   0.00   0.00   0.00   0.00 1000000.00   0.00   0.00 400000.00   0.00 200000.00 200000.00 200000.00   0.00]
T [  0.00   0.00   0.00   0.00 200000.00   0.00   0.00 400000.00 800000.00 600000.00 400000.00   0.00   0.00   0.00   0.00]
>None TMAGTRAAGCGTGRA
A [  0.00 500000.00 1000000.00   0.00   0.00 500000.00 1000000.00 1000000.00   0.00   0.00   0.00   0.00   0.00 500000.00 1000000.00
C [  0.00 500000.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00 1000000.00   0.00   0.00   0.00   0.00   0.00]
G [  0.00   0.00   0.00 1000000.00   0.00 500000.00   0.00   0.00 1000000.00   0.00 1000000.00   0.00 1000000.00 500000.00   0.00]
T [1000000.00   0.00   0.00   0.00 1000000.00   0.00   0.00   0.00   0.00   0.00   0.00 1000000.00   0.00   0.00   0.00]
>None CCTCAYKGMGAGKGT
A [  0.00   0.00   0.00   0.00 1000000.00   0.00   0.00   0.00 500000.00   0.00 1000000.00   0.00   0.00   0.00   0.00]
C [1000000.00 1000000.00   0.00 1000000.00   0.00 500000.00   0.00   0.00 500000.00   0.00   0.00   0.00   0.00   0.00   0.00]
G [  0.00   0.00   0.00   0.00   0.00   0.00 500000.00 1000000.00   0.00 1000000.00   0.00 1000000.00 500000.00 1000000.00   0.00]
T [  0.00   0.00 1000000.00   0.00   0.00 500000.00 500000.00   0.00   0.00   0.00   0.00   0.00 500000.00   0.00 1000000.00]
>None GCSTCTGAAKTTCAT
A [  0.00   0.00   0.00   0.00   0.00   0.00   0.00 1000000.00 1000000.00   0.00   0.00   0.00   0.00 1000000.00   0.00]
C [  0.00 1000000.00 500000.00   0.00 1000000.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00 1000000.00   0.00   0.00]
G [1000000.00   0.00 500000.00   0.00   0.00   0.00 1000000.00   0.00   0.00 500000.00   0.00   0.00   0.00   0.00   0.00]
T [  0.00   0.00   0.00 1000000.00   0.00 1000000.00   0.00   0.00   0.00 500000.00 1000000.00 1000000.00   0.00   0.00 1000000.00]
```

Autor

Dawid Borys

nr indeksu: 394094