Dokumentacja projektu “Parser Blasta”

Cel projektu **1**

Konfiguracja **1**

Działanie aplikacji **2**

Przykładowy fragment wyników BLAST w postaci xml 3

Podziały 3

Raport pdf 4

Raport excel 7

Cel projektu

Celem projektu było zaimplementowanie parsera wyników Blast. Została utworzona aplikacja okienkowa w języku python, która automatycznie generuje raporty ze sparsowanych wyników blasta (format xml). Aplikacja generuje raporty w dwóch formatach:

- Pdf
- Excel

Konfiguracja

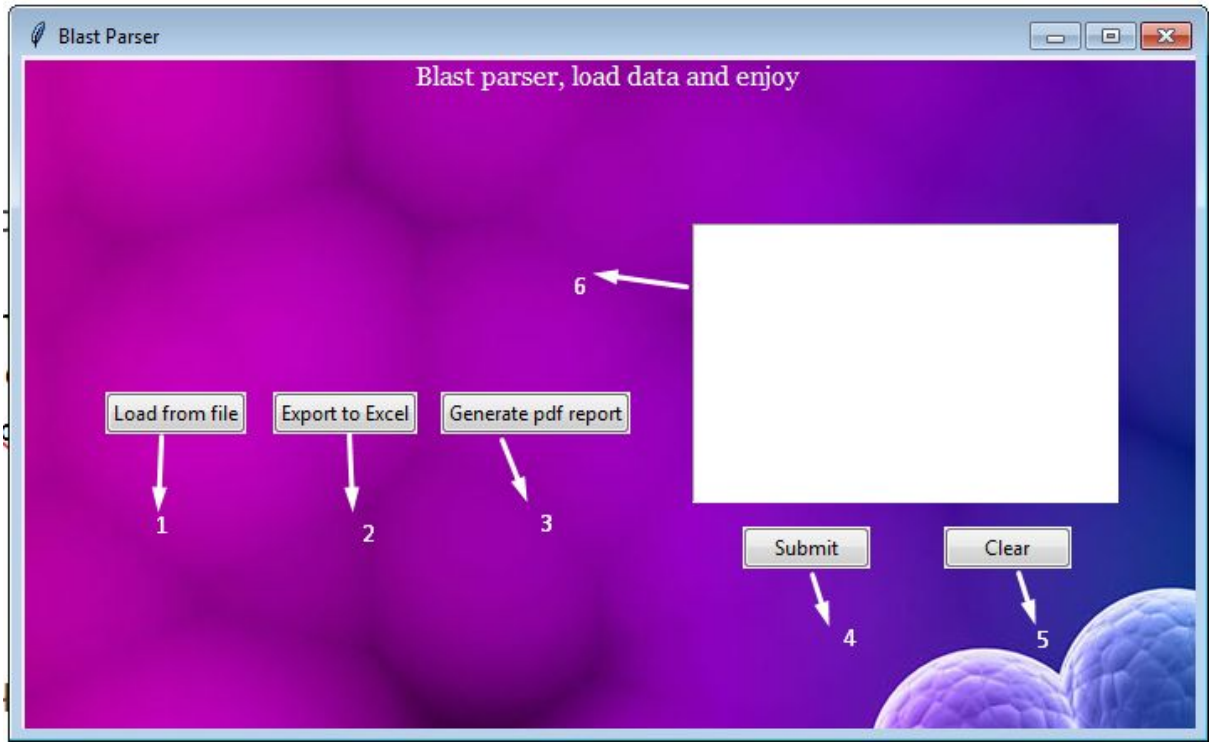
Aby uruchomić aplikację należy uruchomić plik main_frame.py. Przed uruchomieniem należy:

- zainstalować wszystkie paczki znajdujące się w pliku req.txt.
- Jeśli nie korzysta się z windowsa:
    - ściągnąć instalkę dla wkhtmltopdf i dodać plik wkhtmltopdf.exe do ścieżki systemowej i ustawić zmienną WKHTMLTOPDF na ścieżkę bezwzględną do pliku wykonywalnego wkhtmltopdf
    - Jesli krok wcześniejszy nie zadziała proszę zmienić w pliku generate_report.py konfiguracje dla wkhtmltopdf (linia 21) na ścieżkę do pliku wykonywalnego wkhtmltopdf
      (np. wkhtmltopdf=os.path.join("wkhtmltopdf","bin","wkhtmltopdf.exe"))

- Jeśli korzysta się z windowsa wszystko powinno zadziałać, w przeciwnym przypadku prosze wykonać krok poprzedni

Działanie aplikacji

Aplikacja posiada graficzny interfejs, który został przedstawiony poniżej.



1) Użytkownik może załadować plik zawierający wyniki BLAST (format xml) z pliku
2) Użytkownik może wyeksportować dane do excela
3) Uzytkownik może wygenerować raport pdf
4) Użytkownik po wpisaniu sekwencji ręcznie może zatwierdzić dane
5) Użytkownik może wyczyścić pole tekstowe
6) Użytkownik może wpisać dane w pole tekstowe

Scenariusze użycia zostaną przedstawione w osobnym pliku (Scenerios.pdf)

Przykładowy fragment wyników BLAST w postaci xml

W ramach projektu zostało zaimplementowane parsowanie plików xml zawierających wyniki BLAST, a następnie została stworzona struktura przechowywująca dane. Parsowanie odbywa się w pliku parser_blast.py. Głównym kontenerem jest klasa MainAlignment, która przechowuje cały Hit. Zawiera ona dodatkowo listę dopasowań, ponieważ jeden hit może mieć kilka dopasowań w obszarze sekwencji. Klasa Alignment przechowuje wszystkie informacje dotyczące dopasowań tj. Score, length, gaps, title, identities a dodatkowo obliczane jest procentowa jakość dopasowania.

Podziały

Aby raport na temat dopasowań był sensowny sekwencje zostały podzielone na odpowiednie podgrupy:

- **Sekwencje zwyczajne**
  Np. Homo sapiens BRCA1, DNA repair associated (BRCA1), transcript variant 1, mRNA

- **Sekwencje predicted**
  Np. PREDICTED: Pan troglodytes BRCA1, DNA repair associated (BRCA1), transcript variant X2, mRNA

- **Sekwencje synthetic**
  Np. Synthetic construct DNA, clone: pF1KB5593, Homo sapiens BRCA1 gene for breast cancer type 1 susceptibility protein, complete cds, without stop codon, in Flexi System

- **Sekwencje inne** (czasami zdarza się, że sekwencja nie pasuje do żadnej z powyższych)
  Np. hIRS-1=rat insulin receptor substrate-1 homolog [human, cell line FOCUS, Genomic, 6152 nt]

Dodatkowo w obrębie sekwencji normalnych i predicted dostał dokonany podział na gatunki.

Raport pdf

W raporcie pdf umieszczone są wyniki dla podziałów na grupy, a także w obrębie tych grup podziały na gatunki oraz lista występujących gatunków wraz z ilością dopasowań należących do danego gatunku. Dla każdej grupy jak również dla wszystkich dopasowań wyliczone zostały średnie punktów, identycznych , długości, przerw. Zostały zamieszone też wykresy dla wszystkich dopasowań jak i udział poszczególnych grup. Część przykładowego raportu znajduje się poniżej.

Divided by species in normal alignments

	0
Cercopithecus wolfi	1
Colobus guereza	1
Homo sapiens	15
Hylobates agilis	1
Hylobates lar	1
Hylobates pileatus	1
Lophocebus albigena	1
Macaca fascicularis	1
Macaca mulatta	1
Miopithecus talapoin	1
Nomascus gabriellae	1

Synthetic

Means for synthetic alignments

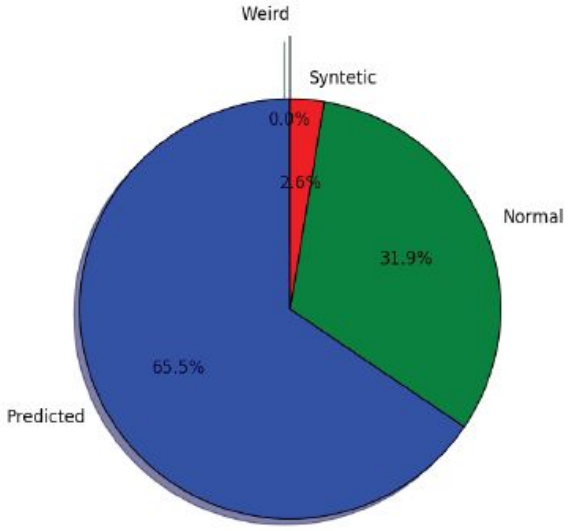
Mean length	Mean identities	Mean score	Mean percent
1494.0	1494.0	2760.01	100.0

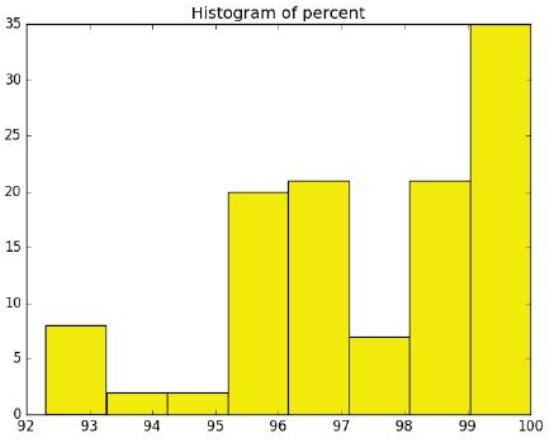
Details

Title	Percent	Score	Length	Gap	Identities
Synthetic construct DNA, clone: pF1KB5593, Homo sapiens BRCA1 gene for breast cancer type 1 susceptibility protein, complete cds, without stop codon, in Flexi system	100.0	2760.01	1494	0	1494
Synthetic construct Homo sapiens clone FLH147954.01L breast cancer 1 early onset (BRCA1) mRNA, partial cds	100.0	2760.01	1494	0	1494
Synthetic construct Homo sapiens clone FLH147909.01X breast cancer 1 early onset (BRCA1) mRNA, complete cds	100.0	2760.01	1494	0	1494

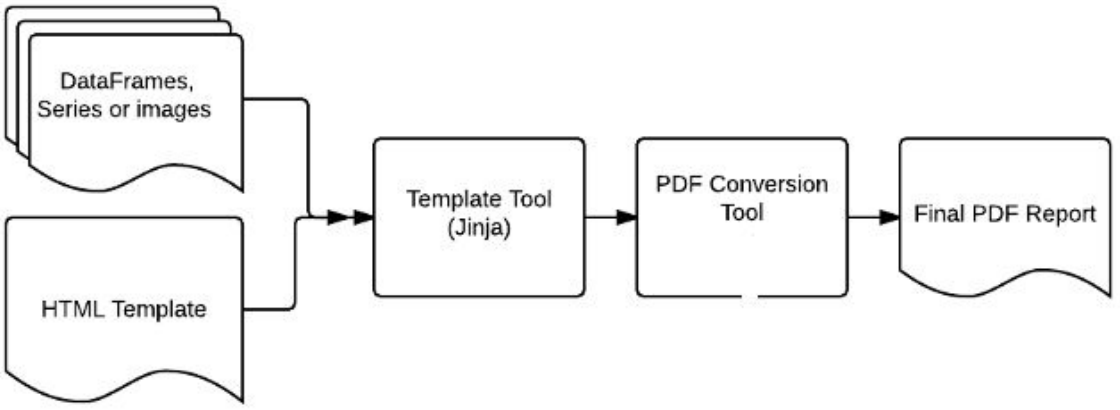
Summary

Number of all alignments	Number of predicted	Number of normal alignments	Number of syntetic	Number of weird	Number of species	Number of species in predicted
116	76	37	3	0	18	10





Do generowania automatycznego raportów do plików pdf użyto biblioteki pandas, a także templatów html (jinja) i paczki pdfkit. Poniżej został przedstawiony sposób działania tego procesu:



Raport excel

W pliku excel zamieszczono dane z podziałem na grupy. Każda grupa znajduje się na osobnej zakładce. Dodatkowo w zakładce “Summary” zostało umieszczone zbiorcze podsumowanie dopasowań. Przykład poniżej:

	A	B	C	D	E	F
1	Title	Percent	Score	Length	Gap	Identities
2	Cercopithecus wolfi breast cancer type 1 susceptibility protein (BRCA1) mRNA, complete cds	96,32	2451.62	1494	3	1439
3	Colobus guereza breast cancer type 1 susceptibility protein (BRCA1) mRNA, complete cds	96,45	2466.4	1494	0	1441
4	Homo sapiens BRCA1, DNA repair associated (BRCA1), transcript variant 1, mRNA	100	3059.17	1656	0	1656
5	Homo sapiens BRCA1, DNA repair associated (BRCA1), transcript variant 2, mRNA	100	3059.17	1656	0	1656
6	Homo sapiens BRCA1, DNA repair associated (BRCA1), transcript variant 3, mRNA	100	2512.56	1360	0	1360
7	Homo sapiens BRCA1, DNA repair associated (BRCA1), transcript variant 3, mRNA	97,52	409.23	242	6	236
8	Homo sapiens IRIS mRNA, complete cds; alternatively spliced	99,42	2828.34	1558	0	1549
9	Homo sapiens breast and ovarian cancer susceptibility (BRCA1) mRNA, complete cds	99,69	2950.22	1613	1	1608
10	Homo sapiens breast and ovarian cancer susceptibility protein (BRCA1) mRNA, BRCA1-2201T/2430C/2731T/3232G/36	100	2760.01	1494	0	1494
11	Homo sapiens breast and ovarian cancer sususceptibility protein 1 (BRCA1) mRNA, complete cds	100	2760.01	1494	0	1494
12	Homo sapiens breast cancer 1, early onset, mRNA (cDNA clone IMAGE:3996658), with apparent retained intron	99,94	2974.23	1613	0	1612
13	Homo sapiens breast cancer 1, early onset, mRNA (cDNA clone IMAGE:40017569), partial cds	100	2712	1468	0	1468
14	Homo sapiens breast cancer 1, early onset, mRNA (cDNA clone IMAGE:40017573)	100	2374.07	1285	0	1285
15	Homo sapiens breast cancer 1, early onset, mRNA (cDNA clone IMAGE:40017573)	100	339.057	183	0	183
16	Homo sapiens breast cancer 1, early onset, mRNA (cDNA clone IMAGE:4804551), partial cds	100	2979.77	1613	0	1613
17	Homo sapiens breast cancer 1, early onset, mRNA (cDNA clone MGC:131629 IMAGE:7961446), complete cds	99,94	2985.31	1619	0	1618
18	Homo sapiens cDNA clone IMAGE:40017575, containing frame-shift errors	99,52	2667.68	1469	7	1462
19	Homo sapiens cDNA clone IMAGE:7961445 >gi|146147644|gb|BC115038.1| Homo sapiens cDNA clone IMAGE:796144	100	2512.56	1360	0	1360
20	Homo sapiens cDNA clone IMAGE:7961445 >gi|146147644|gb|BC115038.1| Homo sapiens cDNA clone IMAGE:796144	97,03	335.364	202	6	196
21	Homo sapiens cDNA, FLJ98032	99,93	2791.41	1514	0	1513
22	Homo sapiens cDNA, FLJ98032	100	98.9927	53	0	53
23	Hylobates agilis breast cancer type 1 susceptibility protein (BRCA1) mRNA, complete cds	98,13	2599.36	1494	6	1466
24	Hylobates lar breast cancer type 1 susceptibility protein (BRCA1) mRNA, complete cds	98,06	2593.82	1494	6	1465
25	Hylobates pileatus breast cancer type 1 susceptibility protein (BRCA1) mRNA, complete cds	98,19	2604.9	1494	6	1467
26	Lophocebus albigena breast cancer type 1 susceptibility protein (BRCA1) mRNA, complete cds	96,06	2433.16	1497	6	1438
27	Macaca fascicularis BRCA1, DNA repair associated (BRCA1), mRNA >gi|667713702|gb|KM017624.1| Macaca fascicula	96,45	2462.7	1494	3	1441
28	Macaca mulatta BRCA1, DNA repair associated (BRCA1), mRNA	96,45	2462.7	1494	3	1441
29	Miopithecus talapoin breast cancer type 1 susceptibility protein (BRCA1) mRNA, complete cds	96,45	2462.7	1494	3	1441

All data | Normal | Synethic | Summary