Sprawozdanie z laboratorium: Bioinformatyka

Część II: Wyniki i analiza eksperymentu

4 kwietnia 2017

Prowadzący: prof. dr hab. inż. Marta Kasprzak

Autorzy:	**Damian Jurga**	inf122481	I2	damian.m.jurga@student.put.poznan.pl
	Grzegorz Miebs	inf122453	I2	grzegorz.miebs@student.put.poznan.pl

Zajęcia środowe, 11:45.

Oświadczamy, że niniejsze sprawozdanie zostało przygotowane wyłącznie przez powyższych autorów, a wszystkie elementy pochodzące z innych źródeł zostały odpowiednio zaznaczone i są cytowane w bibliografii.

1 Wstęp

Celem tego sprawozdania jest przedstawienie teoretycznego opracowania metody heurystycznej rozwiązującej problem sekwencjonowania łańcuchów DNA z błędami pozytywnymi oraz negatywnymi w czasie wielomianowym. Algorytm mając dany na wejściu zbiór oligonukleotydów (tj. ciągów nukleotydów: adeniny, tyminy, guaniny i cytozyny), długość sekwencji oryginalnej, powinien zwrócić jak najdłuższą sekwencję.

W tym celu zaproponowano następujący algorytm: sekwencja jest budowana za pomocą algorytmu wiązkowego, następnie poprawiana algorytmem wspinaczkowym i, jeżeli to możliwe, ponownie przedłużana algorytmem wiązkowym lub poprzez pełny przegląd. Przyjęto szerokość wiązki równą osiem, maksymalną liczbę iteracji algorytmu wspinaczkowego równą sto oraz granicę długości określającą, czy w trzeciej fazie zostanie zastosowany algorytm wiązkowy, czy pełnego przeglądu równą pięć.

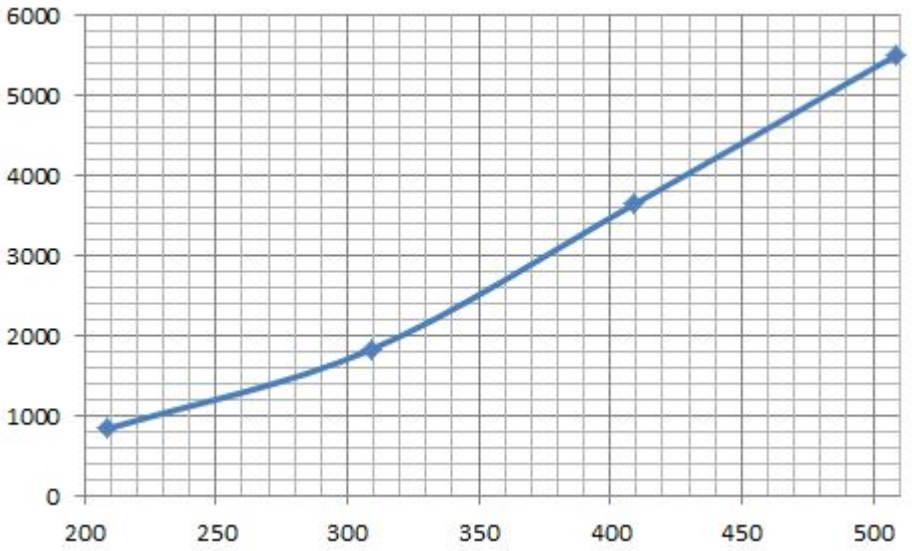
2 Wyniki

Pomiarów dokonano na Mierzono czas wykonania, długość sekwencji oraz liczbę wykorzystanych oligonukleotydów. Wyniki porównano według odległości euklidesowej długości sekwencji wynikowej n_w od długości sekwencji wejściowej n:

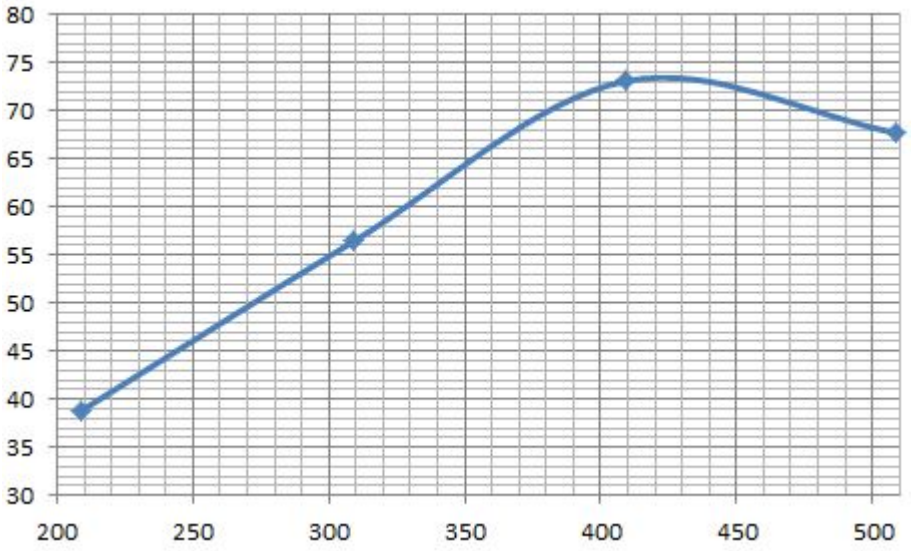
$$L_2 = \sqrt{(n_w - n)^2} = |n_w - n| \tag{1}$$

oraz średniego czasu wykonania. W obu wypadkach w funkcji długości sekwencji wejściowej i mocy zbioru oligonukleotydów.

2.1 Wszystkie instancje



Rysunek 1: *Na osi rzędnych odłożona jest długość sekwencji wejściowej, a na odciętych — średni czas wykonywania w milisekundach*



Rysunek 2: *Na osi rzędnych odłożona jest długość sekwencji wejściowej, a na odciętych — średnia odległość L_2*

Powyższe wykresy pokazują, że dla zadanego układu instancji testowych średni czas wykonania i średnia odległość L_2 jest co najwyżej wielomianem długości sekwencji wejściowej. Rozważanie wartości w funkcji mocy zbioru nie ma sensu, gdyż w większości przypadków wartość funkcji sprowadzałaby się do wartości funkcji obliczanych dla jednej z typów instancji.

2.2 Porównanie

Jak widać na rysunkach 3 i 4, najszybciej zbieżne do lokalnego optimum okazały się instancje z błędami negatywnymi, a najwolniej — z błędami pozytywnymi. Funkcje czasów wykonania wydają się być ograniczone od góry wielomianami.

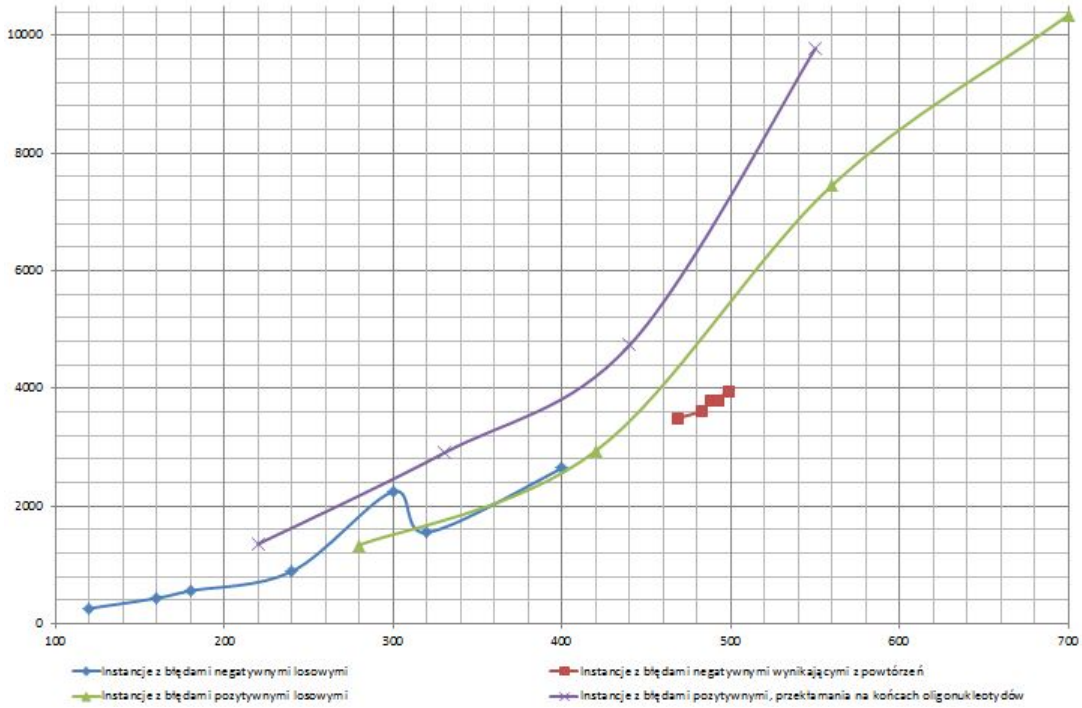
Na rysunkach 5 i 6 widać, że instancje z błędami negatywnymi powodowały utworzenie sekwencji wynikowej zdecydowanie bardziej odległej od wejściowej i bardziej zmiennej. Instancje z błędami pozytywnymi utrzymywały w przybliżeniu stałą odległość L_2.

Można zauważyć, że w przypadku instancji z błędami negatywnymi:

$$L_2 \approx 0.31(7) * n + 4,08(4) \tag{2}$$

3 Plusy

Niewątpliwym plusem tego rozwiązania jest wysoka odporność na złośliwe instancje. Odporność tak wynika z faktu iż algorytm ten składa się z kilku innych algorytmów, które w znacznym stopniu się uzupełniają i rekompensują swoje wady. Kolejnym plusem zaproponowanego rozwiązania jest jego duża elastyczność. Użytkownik może dowolnie sterować parametrami, które w jasny sposób wpływają na czas obliczeń oraz jakość końcowego rozwiązania. Jeśli zależy nam na szybkich obliczeniach możemy ustalić niewielkie wartości parametrów, jednakże jeśli bardziej cenimy jakość rozwiązań możemy zwiększyć ich wartość, należy jednak liczyć się z wydłużeniem czasu działania programu. Alogrytm ten można w znacznym stopniu



Rysunek 3: *Na osi rzędnych odłożona jest moc zbioru oligonukleotydów, a na odciętych — czas wykonywania w milisekundach*

zrównoleglić, co pozwoli wykonywać go w sposób efektywnych w systemach rozproszonych czy nawet na laptopie z wielordzeniowym procesorem. Rozwiązanie to jest podzielone na trzy etapy i w każdym z nich wykorzystywany jest inny algorytm. Ta modułowa budowa pozwala na stosunkowo łatwe wymienianie algorytmów stosowanych w poszczególnych etapach.

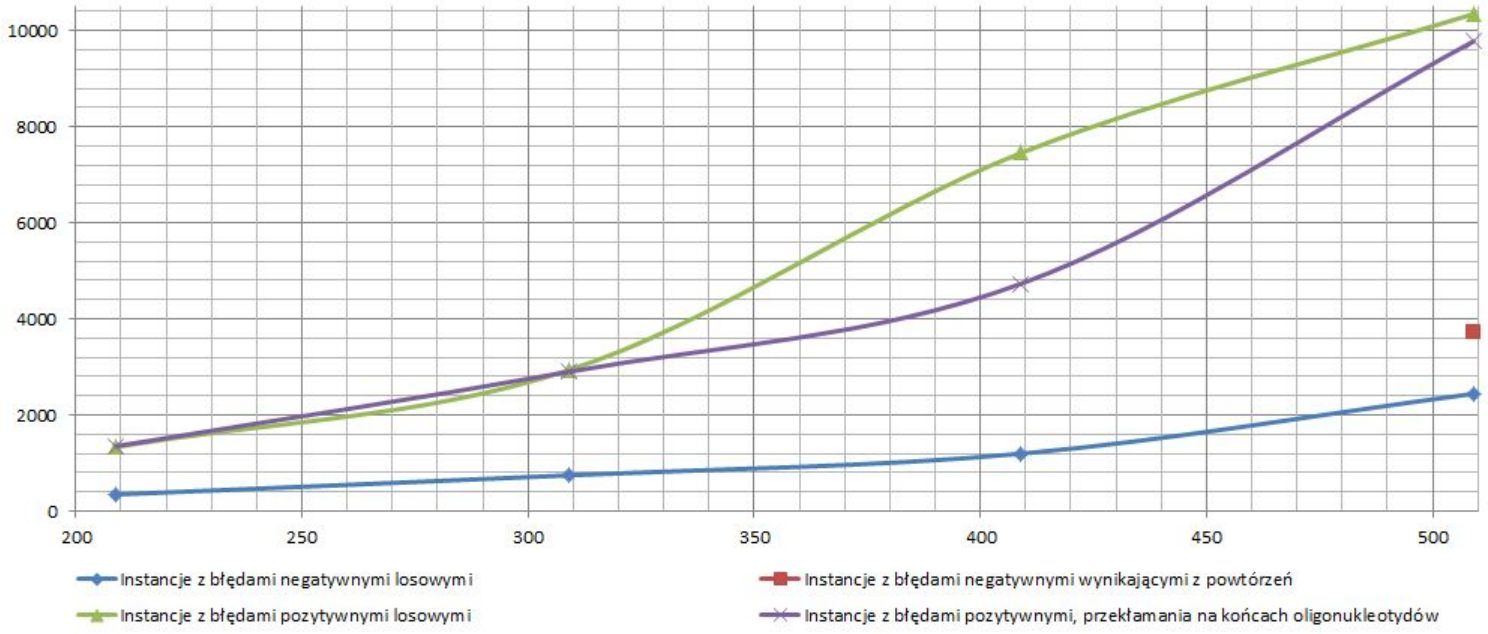
4 Minusy

Algroytm składa się z kilku etapów i w każdym z nich stosowany jest inny algorytm, dlatego, mimo determinizmu tego rozwiązania, niemalże niemożliwe jest przewidzenie bez dokładnej analizy jak zachowa się algorytm dla konkretnej instancji. Ciężko jest także wskazać klasy instancji dla których algorytm poradzi sobie dobrze, bądź źle.

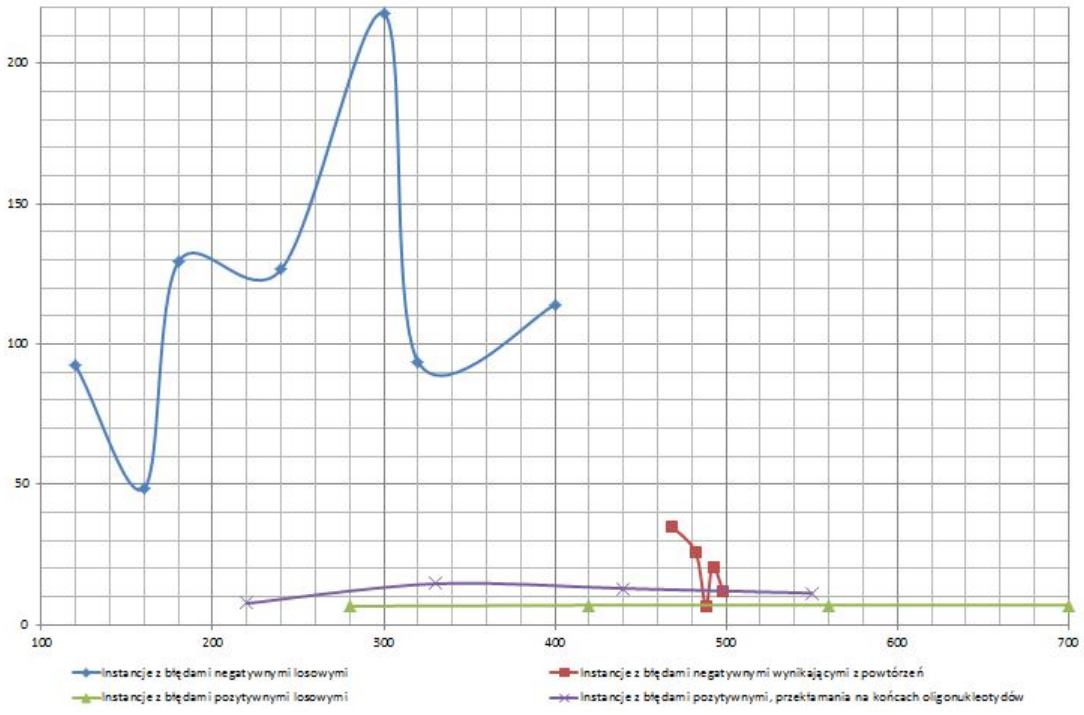
5 Dalsze eksperymenty

Wyniki sugerują, że dałoby się zaproponować takie parametry algorytmu, by zoptymalizować czas wykonywania i odległość od sekwencji wejściowej. Jest to problem optymalizacji wielokryterialnej wymagający zbadania wpływu szerokości wiązki, maksymalnej liczby iteracji algorytmu wspinaczkowego, wartości kary oraz granicy długości określającej, czy w trzeciej fazie zostanie zastosowany algorytm wiązkowy, czy pełnego przeglądu, na oba kryteria.

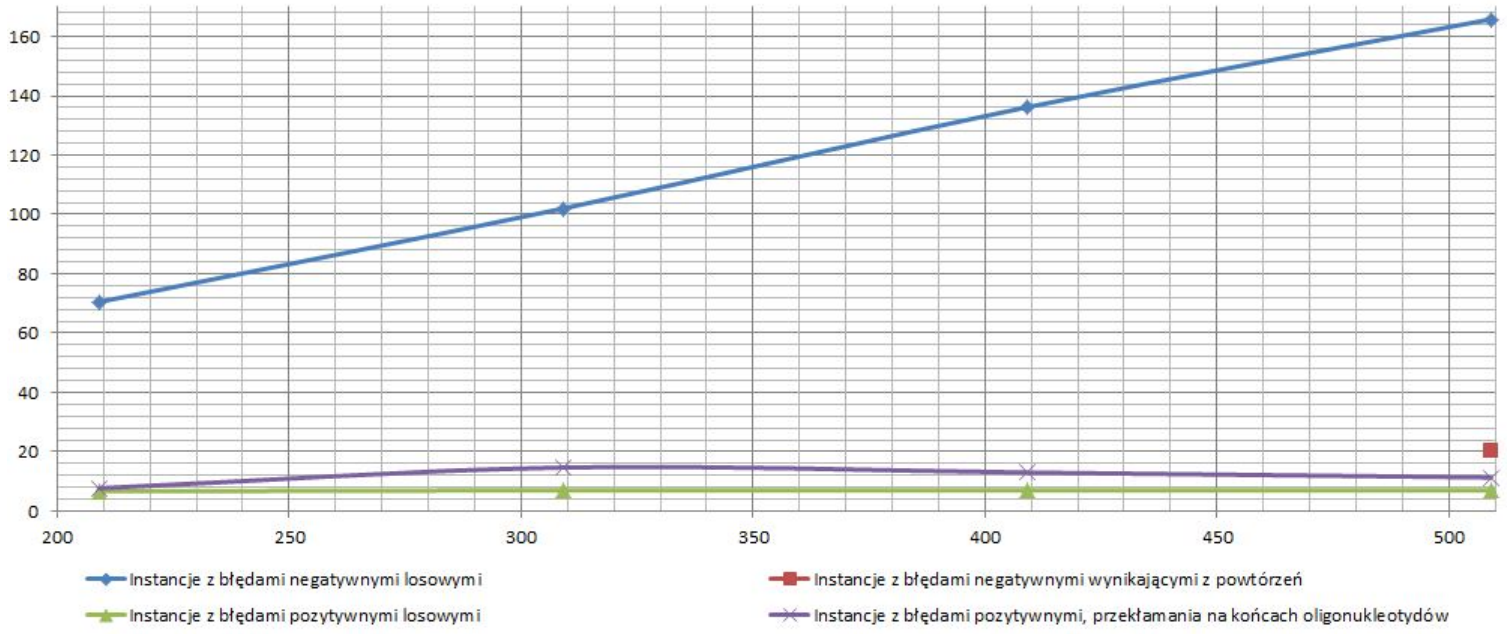
Najprostszym lecz czasochłonnym sposobem wydaje się zastosowanie algorytmu pełnego przeglądu lub wspinaczkowego na charakterystycznych wartościach parametrów wejściowych.



Rysunek 4: *Na osi rzędnych odłożona jest długość sekwencji wejściowej, a na odciętych — czas wykonywania w milisekundach*



Rysunek 5: *Na osi rzędnych odłożona jest moc zbioru oligonukleotydów, a na odciętych — odległości euklidesowej długości sekwencji wynikowej n_w od długości sekwencji wejściowej n*



Rysunek 6: *Na osi rzędnych odłożona jest długość sekwencji wejściowej, a na odciętych — odległości euklidesowej długości sekwencji wynikowej n_w od długości sekwencji wejściowej n*