Poznań, 19.05.2022r.

Sekwencjonowanie łańcuchów DNA z błędami negatywnymi i pozytywnymi

Maurycy Oprus, 145207
Mikołaj Mrożewski, 145331

Politechnika Poznańska, bioinformatyka

1. Wstęp.

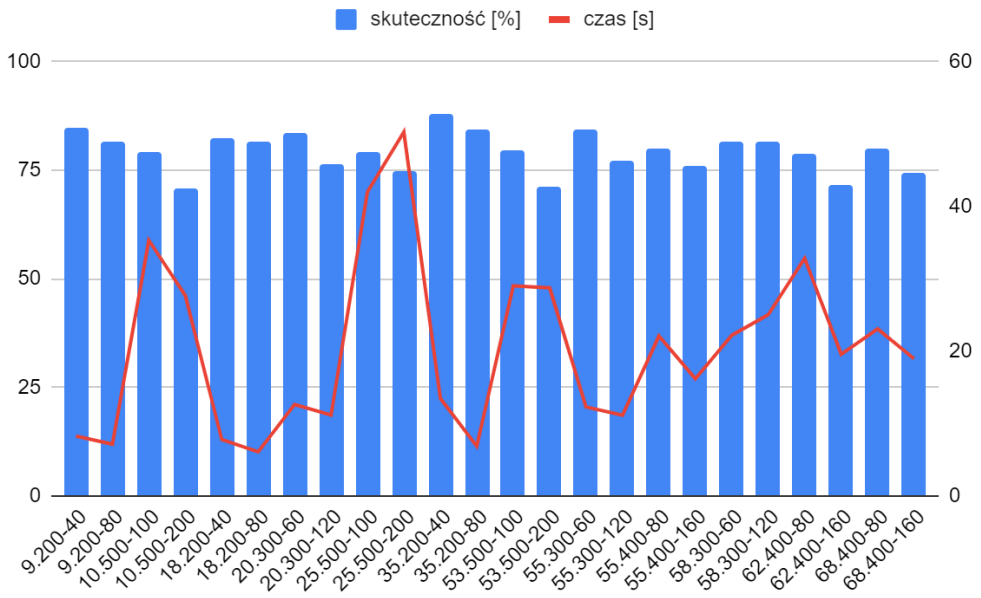
Problemem rozwiązywanym przez nas było jak najdokładniejsze odtworzenie sekwencji nukleotydów, łącząc oligonukletydy dane na wejściu. Podane zbiory danych wejściowych - spektra - zawierały błędy negatywne - brak danego oligonukleotydu w spektrum, mimo że występowało ono w oryginalnej sekwencji, oraz błędy pozytywne - dodanie oligonukleotydu oryginalnie nie występującego w sekwencji. Do sekwencjonowania użyliśmy algorytmu genetycznego. Czasy oraz jakość mierzone dla każdej instancji były średnią z dziesięciu pomiarów.

2. Opis metody.

- ***a.*** Łączenie oligonukleotydów w łańcuchy z maksymalnym dopasowaniem (w przypadku, gdy istnieje tylko jedno takie dopasowanie).
- ***b.*** Stworzenie ustalonej ilości osobników za pomocą permutacji wszystkich łańcuchów otrzymanych w pierwszym punkcie (rozmiar populacji = 200 osobników).
- ***c.*** Ocena osobników i wybranie najlepszych do dalszego krzyżowania - populacja początkowa. Kryterium oceny to najlepsze dopasowanie.
- ***d.*** Krzyżowanie początkowej populacji - wybór dwóch rodziców z populacji, losowanie czy dodać kolejny ciąg z pierwszego czy z drugiego rodzica. Dodanie wygenerowanego potomka do nowej populacji.
- ***e.*** Najgorsze osobniki z populacji zastąpione zostają najlepszymi potomkami (68% wymiany populacji w jednej iteracji)
- ***f.*** Mutacja (zamiana niektórych łańcuchów w 5% populacji) - w celu znajdowania potencjalnie lepszych rozwiązań.
- ***g.*** Krzyżowanie nowo powstałej populacji - 300 iteracji, przerywanych w przypadkach opisanych w pkt. h.
- ***h.*** Zakończenie algorytmu w momencie, gdy znaleziono najlepsze rozwiązanie/rozwiązanie nie polepsza się znacząco przez kilka iteracji.

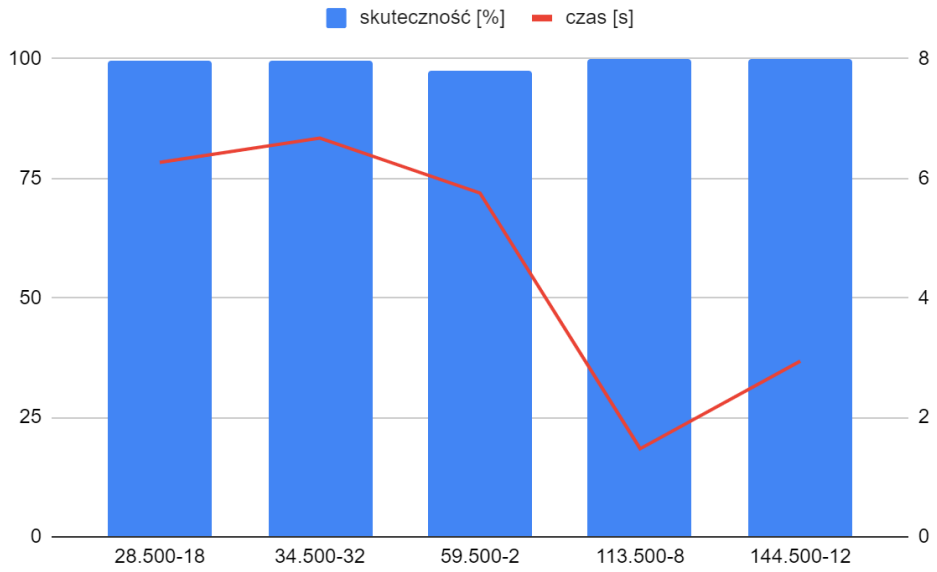
3. Wyniki

a. Instancje z błędami negatywnymi losowymi



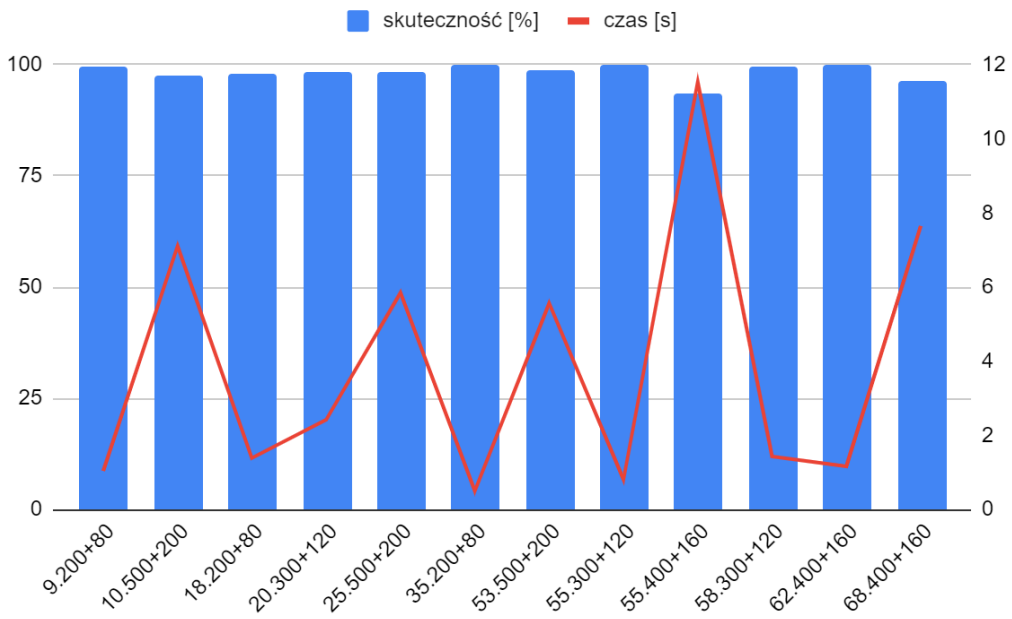
Średnia skuteczność: 79,33%
Średni czas: 20,40 s
Najlepszy wynik: 88,13%
Najgorszy wynik: 70,8%

b. Instancje z błędami negatywnymi wynikającymi z powtórzeń



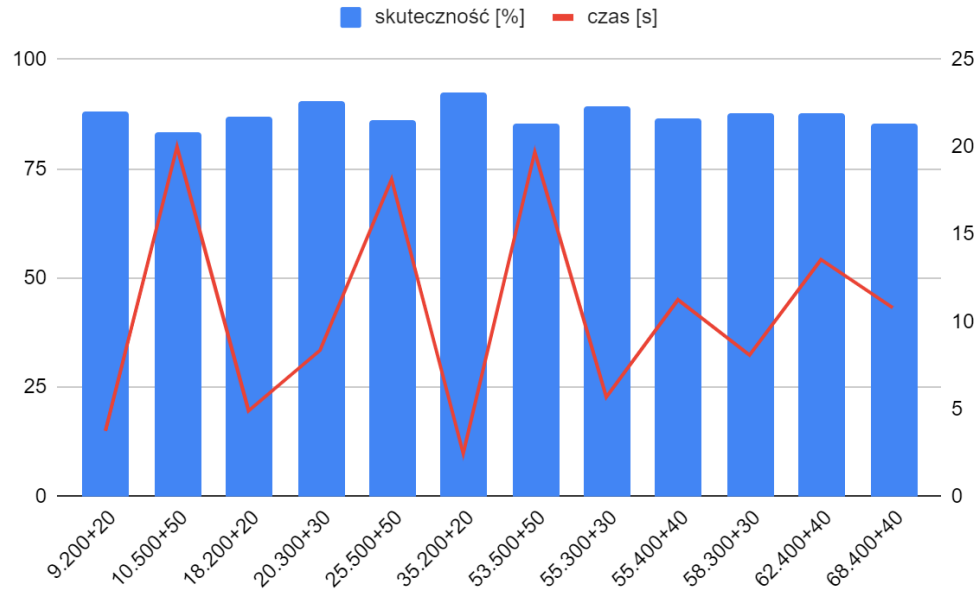
Średnia skuteczność: 99,31%
Średni czas: 4,62 s
Najlepszy wynik: 100%
Najgorszy wynik: 97,63%

c. Instancje z błędami pozytywnymi losowymi



Średnia skuteczność: 98,35%
Średni czas: 3,87 s
Najlepszy wynik: 100%
Najgorszy wynik: 93,52%

d. Instancje z błędami pozytywnymi, przekłamania na końcach oligonukleotydów



Średnia skuteczność: 87,41%
Średni czas: 10,56 s
Najlepszy wynik: 92,3%
Najgorszy wynik: 83,32%

4. Wnioski

a. *Wady*

- Może się zdarzyć, że na wczesnym etapie algorytm pójdzie ścieżką prowadzącą do rozwiązania kiepskiej jakości, i ciężko będzie mu wrócić do etapu, z którego może dojść do rozwiązania wysokiej jakości.
- Długi czas przeszukiwania dla instancji zawierających wiele łańcuchów wejściowych.

b. *Zalety*

- W stosunkowo dobrym czasie algorytm przeszukuje dużą ilość możliwych rozwiązań, i selekcjonuje najlepsze z nich na danym etapie.
- Łatwo można zmieniać ustawienia algorytmu, dostosowując jego działanie do swoich potrzeb - balans między oczekiwanym czasem działania a jakością rozwiązania.

c. *Porównanie działania na różnych instancjach*

- Na trudność instancji wpływają m. in. błędy. Jeśli oligonukleotyd wynikający z błędu pozytywnego jest dobrze dopasowany do jednego z nukleotydów, to algorytm może klasyfikować ten błędny oligonukleotyd jako pożądany w rozwiązaniu, co będzie wpływało na jego ostateczną jakość. Jeśli jednak błędny oligonukleotyd znacznie różni się od prawidłowych oligonukleotydów, to połączą się one w dłuższe łańcuchy, a błędne oligonukleotydy zostaną szybko oddzielone od tych prawidłowych. W ogólności im więcej błędów, tym trudniejsza jest instancja.
- Algorytm najgorzej radzi sobie z instancją zawierającą losowe błędy negatywne. Wynika to z tego, że każdy błąd negatywny powoduje, że nie jest możliwe utworzenie dłuższego łańcucha poprzez maksymalne dopasowanie, co wpływa na większą ilość możliwych permutacji wszystkich utworzonych łańcuchów, czyli możliwych rozwiązań.