03.06.2019

Sekwencjonowanie łańcuchów DNA – sprawozdanie 2  
Mariusz Furmaniak 132222  
Kamil Górecki 142287

Przygotowany algorytm przygotowuje tablice pokyć słów, na podstawie których tworzy łańcuch, starając się zachować jak największą efektywność (najmniejszą liczbę słów nieużytych). W związku z tym jest to algorytm zachłanny, gwarantujący poprawne rozwiązanie – jednak niekoniecznie najbardziej optymalne. Jest to podstawowa wada – w przypadku, gdy potrzebne byłyby wyniki bardzo dokładne, zastosowana heurystyka się nie sprawdza. Jednak na potrzeby zadań, wymagających raczej ogólnego zarysu łańcucha, a nie jego idealnego odzwierciedlenia, algorytm spełnia swoją rolę i posiada sporo zalet – przede wszystkim jest prosty w implementacji i zrozumieniu, co w przypadku problemów niezwiązanych ściśle z dziedziną wiedzy autorów jest ważne. Pozostawia również szerokie możliwości dalszego rozwoju stosowanej heurystyki – dodając kolejne funkcje walidacyjne można próbować zwiększyć dokładność rozwiązania lub przyspieszyć średnie czasy otrzymywanych wyników.

Wyniki:

Tabela 1. Średnia dokładność otrzymywanych wyników na instancjach testowych

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Rozdzaj błędów | Minimalna dokładność [%] | Maksymalna dokładność [%] | Średnia dokładność [%] | Odchylenie standardowe [%] |
| Negatywne losowe | 90.7  (55.300-120) | 97.0  (10.500-200) | 94.2 | 3.16 |
| Negatywne wynikające z powtórzeń | 95.4  (34.500-32) | 98.8  (144.500-12) | 96.7 | 1.67 |
| Pozytywne losowe | 89.5  (68.400+160) | 98.2  (53.500+200) | 93.9 | 4.45 |
| Pozytywne z przekłamaniami na końcach | 93.5 (68.400+40) | 97.4  (25.500+50) | 95.8 | 2.26 |

gdzie dokładność oznacza:  
- w przypadku błędów negatywnych – procent słów użytych w rozwiązaniu finalnym względem słów posiadanych  
- w przypadku błędów pozytywnych – procent słów użytych w rozwiązaniu finalnym względem słów nie wynikających z błędów

Tabela 2. Średnia dokładność obliczeń w zależności od rozmiaru danych wejściowych

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Rodzaj błędów | Średnia dokładność dla n = 200 [%] | Średnia dokładność dla n = 300 [%] | Średnia dokładność dla n = 400 [%] | Średnia dokładność dla n = 500 [%] |
| Negatywne losowe | 93.8 | 93.2 | 94.6 | 95.0 |
| Negatywne wynikające z powtórzeń | - | - | - | 96.7 |
| Pozytywne losowe | 93.5 | 93.6 | 93.4 | 95.3 |
| Pozytywne z przekłamaniami na końcach | 93.9 | 96.4 | 96 | 96.9 |

Tabela 3. Średnie czasy wykonywania się obliczeń w zależności od rozmiaru danych wejściowych

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Rodzaj błędów | Czas wykonywania dla n = 200 [s] | Czas wykonywania dla n = 300 [s] | Czas wykonywania dla n = 400 [s] | Czas wykonywania dla n = 500 [s] |
| Negatywne losowe | 0.27 | 0.60 | 1.06 | 1.64 |
| Negatywne wynikające z powtórzeń | - | - | - | 3.14 |
| Pozytywne losowe | 1.01 | 2.25 | 3.99 | 6.19 |
| Pozytywne z przekłamaniami na końcach | 0.65 | 1.43 | 2.18 | 3.91 |

Na podstawie powyższych statystyk można zauważyć kilka zależności:

- dokładność rozwiązań jest zbliżona dla każdego rodzaju błędu

- dokładność rozwiązań jest niezależna od rozmiaru instancji – wyniki są bardzo zbliżone, a ilość instancji testowych zbyt mała, by uznać te wartości za różne. Jedynie w ramach ciekawostki można zauważyć, że największa dokładność (przy każdym rodzaju błędów) zawsze przypada na instancję o rozmiarze n = 500, co wydaje się być sprzeczne z intuicją.

- czasy są mocno zależne od rodzaju błędu występującego w instancji testowej – ze względu na zastosowaną metodę unikania wyboru fałszywego następnika gorzej wypadają instancję z błędami pozytywnymi – w szczególności w przypadkach przekłamań na końcach oligonukleotydów. Wynika to z charakterystyki tego typu błędów – zwiększają one dostępną pulę słów, a dodatkowo umożliwiają występowanie dużej liczby słów potencjalnie uznawanych za następne słowo w sekwencji – co wymaga dodatkowych obliczeń.