**Sprawozdanie 2**

**Temat:** Sekwencjonowanie łańcuchów DNA

Anastasiia Trubchaninova 143207

Ivan Bilunyk 141189

Algorytm przygotowuje tablice słów, na podstawie których tworzy ciąg, starając się być jak najbardziej wydajnym (jak najmniej niewykorzystanych słów). Dlatego jest to zachłanny algorytm, który gwarantuje rozwiązanie poprawne.

**Wyniki:**

Tabela 1. Średnia dokładność otrzymywanych wyników na instancjach testowych

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Rodzaj 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 ostatecznym rozwiązaniu w stosunku do dostępnych słów  
- w przypadku błędów pozytywnych –procent słów użytych w ostatecznym rozwiązaniu w porównaniu ze słowami niepowodowanymi błędami

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 danych podanych wyżej można zauważyć kilka zależności:

- dokładność rozwiązań nie zależy od wielkości instancji – wyniki są bardzo zbliżone, a liczba instancji testowych jest zbyt mała, aby uznać te wartości za różne.

- czasy są istotnie uzależnione od rodzaju błędu występującego w instancji testowej – ze względu na metodę zastosowaną w celu uniknięcia fałszywej selekcji następców, instancja z błędami dodatnimi wypada gorzej – szczególnie w przypadku przekłamań na końcach oligonukleotydów. Wynika to z charakterystyki tego typu błędów – zwiększają one dostępną pulę słów i dodatkowo pozwalają na występowanie dużej ilości słów potencjalnie uznawanych za kolejne słowo w sekwencji, co wymaga dodatkowych obliczeń.

- dokładność rozwiązań jest prawie identyczna dla każdego rodzaju błędu

**Wady:** rozwiązanie jest niekoniecznie najbardziej optymalne, w przypadku, gdyby były potrzebne bardzo dokładne wyniki. W takim przypadku zastosowana heurystyka się nie sprawdzi.

**Zalety:** dla zadań, które wymagają ogólnego zarysu łańcucha, a nie jego idealnego odzwierciedlenia, algorytm spełnia swoją rolę i ma wiele zalet – przede wszystkim jest łatwy do implementacji i zrozumienia, co jest ważne w przypadku zadań, które nie są ściśle związane z wiedzą autora. Daje to również szerokie możliwości dalszego rozwoju stosowanej heurystyki – dodając dodatkowe funkcje weryfikacyjne można spróbować poprawić dokładność rozwiązania lub przyspieszyć średni czas uzyskania wyników.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Negatywne losowe** | | **Negatywne wynikające z powtórzeń** | | **Pozytywne losowe** | | **Pozytywne z przekłamaniami na końcach** | |
| Instancja | % | Instancja | % | Instancja | % | Instancja | % |
| 9.200-40 | **96.5** | 59.500-2 | **97.6** | 9.200+80 | **95.5** | 9.200+20 | **93.5** |
| 9.200-80 | **93.5** | 113.500-8 | **95.6** | 18.200+80 | **93** | 18.200+20 | **94** |
| 18.200-40 | **94.2** | 144.500-12 | **98.8** | 35.200+80 | **92** | 35.200+20 | **94** |
| 18.200-80 | **92** | 28.500-18 | **96.2** | 20.300+120 | **94.3333** | 20.300+30 | **96.3333** |
| 35.200-40 | **95.5** | 34.500-32 | **95.4** | 55.300+120 | **93** | 55.300+30 | **96.3333** |
| 35.200-80 | **91.5** |  | | 58.300+120 | **93.3333** | 58.300+30 | **96.6667** |
| 20.300-60 | **93** | 55.400+160 | **95.25** | 55.400+40 | **97.25** |
| 20.300-120 | **92.6667** | 62.400+160 | **95.5** | 62.400+40 | **97.25** |
| 55.300-60 | **92** | 68.400+160 | **89.5** | 68.400+40 | **93.5** |
| 55.300-120 | **90.6667** | 10.500+200 | **95** | 10.500+50 | **97.2** |
| 58.300-60 | **95.6667** | 25.500+200 | **92.6** | 25.500+50 | **97.4** |
| 58.300-120 | **95.3333** | 53.500+200 | **98.2** | 53.500+50 | **96** |
| 55.400-80 | **95** |
| 55.400-160 | **92.75** |
| 62.400-80 | **95.75** |
| 62.400-160 | **92.75** |
| 68.400-80 | **96.25** |
| 68.400-160 | **95.25** |
| 10.500-100 | **96** |
| 10.500-200 | **97** |
| 25.500-100 | **94.8** |
| 25.500-200 | **93,8** |
| 53.500-100 | **95** |
| 53.500-200 | **93.4** |