Sprawozdanie z zadania IV z algorytmów kombinatorycznych w bioinformatyce

Mateusz Roszczyk 151551

Opis algorytmu

Po krótkim czasie od momentu kompilacji program prosi użytkownika o podanie na wejście do konsoli nazwy pliku z instancją.

Następnie program otwiera plik z daną instancją i każdą odczytaną liczbę, oznaczającą długość odcinka DNA, zapisuje w wektorze „multiset".

Potem program tworzy mapę, a następnie oblicza każdą możliwą liczbę odcinków jaką może zawierać poprawny multizbiór. Zgodnie symbolem Newtona[1], istnieje multizbiór zawierający dokładnie jedną sumę wszystkich odcinków odpowiadających danej liczbie cięć. Na podstawie wyżej wymienionego wzoru przelicza się dopuszczalne wielkości multizbioru i porównuje się je z wektorem „multiset". Jeżeli wektor zawiera odpowiednią liczbę odcinków, program dalej przetwarza dane.

W dalszej kolejności program przygotowuje struktury danych, w których będą przechowywane: uzyskana mapa – wektor „map" i „final_map", odczytane odcinki przy budowie drzewa przeszukiwania – wektor „read_from_multiset" i pierwszy element stanowiący korzeń drzewa przeszukiwania, będący odcinkiem o długości różnicy ostatniego i przedostatniego najdłuższych odcinków w multizbiorze.

Następnie tworzone jest drzewo przeszukiwania, które wykonuje się rekurencyjnie dopóki nie znajdzie pełnej mapy dla danego multizbioru. Program wykonuje algorytm, który iteruje po każdym odcinku multizbioru, dodaje do mapy cięć odcinek multizboru, który nie znajduje się już w mapie cięć, a potem sprawdza, czy nowo powstały zbiór odcinków, powstały po pocięciu sekwencji według nowo powstałej mapy cięć, pokrywa się z podanym na wejściu multizbiorem odcinków. Jeżeli oba zbiory pokrywają się, to algorytm jest ponownie wywoływany i szukanie rozwiązania jest kontynuowane z mapą cięć uzupełnioną o znaleziony odcinek. Jeżeli oba te zbiory nie pokrywają się, znaleziony odcinek usuwany jest z mapy, a następnie program analogicznie jak na początku, dodaje następny odcinek iterując po całym multizbiorze.

W momencie zakończenia pracy algorytmu, program wyświetla informacje o znalezionej mapie albo o braku znalezienia takowej, wraz z informacją o czasie działania algorytmu w sekundach.

1 $|A| = \binom{k+2}{2}$

Wyniki testów

Zawartość multizbioru 5i1:

2 5 6 7 8 10 13 15 16 18 23 24 25 29 31

Mapa: **2 5 8 10 6**

Czas przetwarzania algorytmu: 0 s.

Zawartość multizbioru 5i2:

9 24 39 40 44 48 64 68 83 92 107 108 116 147 156

Mapa: **9 39 44 24 40**

Czas przetwarzania algorytmu: 0 s.

Zawartość multizbioru 8i1:

2 4 6 9 14 17 21 21 23 23 24 27 30 31 32 33 35 36 37 41 44 53 54 55 56 58 60 67 72 76 77 78 91 99 108 132

Mapa: **24 9 21 2 4 17 14 41**

Czas przetwarzania algorytmu: 0.002 s.

Zawartość multizbioru 8i2:

38 48 57 63 87 88 88 94 120 126 132 135 136 150 151 189 198 207 214 220 223 252 255 277 286 301 339 340 343 349 387 427 437 475 475 563

Mapa: **88 38 94 57 63 87 48 88**

Czas przetwarzania algorytmu: 0 s.

Zawartość multizbioru 11i1:

14 15 18 23 35 38 56 57 71 73 80 80 85 90 91 94 95 96 103 113 115 130 141 153 168 170 171 176 181 184 186 186 199 209 237 256 266 271 271 272 280 289 294 312 327 350 351 352 362 365 367 369 383 385 407 421 442 442 456 457 465 480 522 536 537 551

Mapa: **14 57 23 90 96 85 18 38 35 80 15**

Czas przetwarzania algorytmu: 0.002 s.

Zawartość multizbioru 11i2:

21 26 35 38 41 42 58 59 62 68 76 77 78 82 97 103 103 104 118 118 120 138 140 141 144 146 156 159 178 180 180 181 199 200 206 218 221 222 235 238 256 259 277 284 294 296 298 300 303 324 336 341 358 362 376 381 399 402 418 434 440 444 476 502 522 580

Mapa: **58 82 38 21 97 62 41 35 42 26 78**

Czas przetwarzania algorytmu: 0.011 s.

Zawartość multizbioru 14i1:

39 81 121 140 143 143 148 165 185 189 203 206 223 228 233 264 266 272 286 288 305 314 328 353 354 409 409 412 426 426 429 451 453 461 494 499 538 542 557 569 574 615 632 642 654 659 681 684 697 717 727 765 780 780 821 860 862 870 887 920 923 950 968 969 983 1008 1085 1093 1093 1109 1126 1148 1153 1174 1206 1241 1274 1296 1313 1322 1349 1359 1381 1395 1434 1462 1502 1507 1538 1546 1577 1627 1647 1650 1667 1748 1790 1810 1812 1891 1933 1955 2076 2076 2219

Mapa: **143 121 165 140 148 206 203 223 189 39 233 81 185 143**

Czas przetwarzania algorytmu: 0.003 s.

Zawartość multizbioru 14i2:

23 57 73 80 111 119 121 121 129 131 144 144 151 164 173 178 194 201 204 230 230 248 250 253 272 273 275 275 315 322 325 330 345 348 359 369 374 392 394 426 436 449 453 469 476 480 503 503 509 523 526 545 547 560 620 620 622 640 644 647 666 667 674 724 733 739 751 784 795 795 798 818 868 870 875 895 897 916 939 962 989 996 999 1012 1014 1048 1069 1073 1120 1143 1143 1169 1192 1200 1242 1264 1287 1321 1344 1373 1442 1465 1517 1522 1695

Mapa: **173 57 23 121 129 119 111 164 151 121 73 131 144 178**

Czas przetwarzania algorytmu: 0.003 s.

Zawartość multizbioru test1:

2 3 4 4 5 5 6 6 6 7 7 8 8 9 10 11 11 11 11 12 13 13 14 14 15 15 16 16 17 18 19 19 19 20 20 21 21 22 23 23 25 25 25 26 26 27 27 27 30 30 31 31 32 32 35 35 36 36 38 38 38 40 41 42 42 43 44 46 46 49 51 52 53 57 57 61 63 67

Mapa: **4 6 5 8 3 9 5 2 4 7 8 6**

Czas przetwarzania algorytmu: 18.713 s.

Zawartość multizbioru test2:

2 3 4 4 5 5 6 6 6 6 7 7 8 8 9 10 11 11 11 11 12 12 13 13 14 14 15 15 16 16 17 18 19 19 19 20 20 20 21 21 22 23 23 25 25 25 26 26 27 27 27 27 30 30 31 31 31 32 32 33 35 35 36 36 38 38 38 38 40 41 42 42 43 44 46 46 47 49 50 51 52 53 57 57 58 61 63 63 67 69 73

Mapa: **4 6 5 8 3 9 5 2 4 7 8 6 6**

Czas przetwarzania algorytmu: 71.46 s.

Zawartość multizbioru test3:

2 3 3 4 4 5 5 6 6 6 6 7 7 8 8 9 9 10 11 11 11 11 12 12 13 13 14 14 15 15 15 16 16 17 18 19 19 19 20 20 20 21 21 22 23 23 23 25 25 25 26 26 27 27 27 27 30 30 30 31 31 31 32 32 33 34 35 35 36 36 36 38 38 38 38 40 41 41 42 42 43 44 46 46 47 49 50 50 51 52 53 53 57 57 58 61 61 63 63 66 67 69 72 73 76

Mapa: **3 6 6 8 7 4 2 5 9 3 8 5 6 4**

Czas przetwarzania algorytmu: 222.131 s.

Zawartosc multizbioru test4:

2 3 3 4 4 5 5 5 6 6 6 6 7 7 8 8 8 9 9 10 11 11 11 11 12 12 13 13 14 14 14 15 15 15 16 16 17 18 19 19 19 20 20 20 20 21 21 22 23 23 23 25 25 25 26 26 27 27 27 27 28 30 30 30 31 31 31 32 32 33 34 35 35 35 36 36 36 38 38 38 38 39 40 41 41 41 42 42 43 44 46 46 46 47 49 50 50 51 52 53 53 55 57 57 58 58 61 61 63 63 66 66 67 69 71 72 73 76 77 81

Mapa: **4 6 5 8 3 9 5 2 4 7 8 6 6 3 5**

Czas przetwarzania algorytmu: 1506.94 s.

Zawartosc multizbioru test5:

12 13 15 25 27 35 38 42 47 48 54 57 60 66 66 74 79 82 89 93 95 101 104 107 108 112 114 123 135 136 139 146 150 151 161 167 171 184 186 193 196 199 205 209 211 212 224 228 239 241 247 250 253 259 262 265 286 294 307 307 313 319 321 334 346 351 360 373 395 398 400 408 420 433 446 458 474 512

Mapa: **38 74 27 66 42 15 89 47 35 13 12 54**

Czas przetwarzania algorytmu: 0.131 s.

Zawartosc multizbioru test6:

512 474 458 446 433 420 408 400 398 395 373 360 351 346 334 321 319 313 307 307 294 286 265 262 259 253 250 247 241 239 228 224 212 211 209 205 199 196 193 186 184 171 167 161 151 150 146 139 136 135 123 114 112 108 107 104 101 95 93 89 82 79 74 66 66 60 57 54 48 47 42 38 35 27 25 15 13 12

Mapa: **38 74 27 66 42 15 89 47 35 13 12 54**

Czas przetwarzania algorytmu: 6.615 s.

Zawartosc multizbioru test7:

12 13 15 25 25 27 35 38 42 47 48 54 57 57 60 63 66 66 74 79 79 82 82 89 93 95 101 104 107 108 112 114 120 123 135 136 136 137 139 146 150 151 161 161 164 167 171 184 186 193 194 196 199 199 205 209 211 212 221 224 228 230 239 241 247 250 253 259 262 265 272 273 286 287 287 294 300 307 307 313 319 321 329 334 344 346 351 360 366 373 376 395 398 400 408 408 420 423 423 433 433 446 458 458 471 474 480 483 512 512 515 528 537 540 559 594 594 607 619 673

Mapa: **54 12 13 35 47 89 15 42 66 27 74 38 25 57 79**

Czas przetwarzania algorytmu: 0.005 s.

Zawartosc multizbioru test8:

673 619 607 594 594 559 540 537 528 515 512 512 483 480 474 471 458 458 446 433 433 423 423 420 408 408 400 398 395 376 373 366 360 351 346 344 334 329 321 319 313 307 307 300 294 287 287 286 273 272 265 262 259 253 250 247 241 239 230 228 224 221 212 211 209 205 199 199 196 194 193 186 184 171 167 164 161 161 151 150 146 139 137 136 136 135 123 120 114 112 108 107 104 101 95 93 89 82 82 79 79 74 66 66 63 60 57 57 54 48 47 42 38 35 27 25 25 15 13 12

Mapa: **54 12 13 35 47 89 15 42 66 27 74 38 25 57 79**

Czas przetwarzania algorytmu: 112.554 s.

Wnioski

Uporządkowanie długości odcinków we wczytywanym zbiorze znacząco wpływa na długość przetwarzania algorytmu. Dla zbiorów uporządkowanych malejąco przetwarza się w skrajnym przypadku prawie 22 500 razy dłużej[2] niż dla tego samego zbioru uporządkowanego rosnąco.

Dodanie do kodu flagi, która zapamiętuje daną długość odcinka, która została odrzucona podczas ostatniej iteracji w algorytmie przeszukiwania drzewa, redukuje czas pracy algorytmu ponad 6-krotnie[3] w przypadku posortowanych multizbiorów. Dzieje się tak dlatego, że algorytm w przypadku zbiorów posiadających wiele odcinków tej samej długości nie powtarza tej samej operacji sprawdzenia poprawności mapy dla odcinków o tej samej długości, co znacząco skraca czas przeszukiwania.

2 Czas wykonywania algorytmu dla instancji test7 wynosi 0,005 sekundy, a dla test8 wynosi 112,554 sekundy.

3 Dla instancji test8 czas działania algorytmu przed zastosowaniem flagi wynosił 741 sekund, a po zastosowaniu 112 sekund.