Warszawa, 17.04.2014

Marcin Kubik

Krzysztof Opasiak

Jacek Sosnowski

METODY BIOINFORMATYKI

**Implementacja algorytmu BLOSUM do obliczania macierzy substytucji**

Prowadzący: Marcin Lewandowski

1. **Treść projektu**

Tematem projektu jest implementacja algorytmu BLOSUM do obliczania macierzy substytucji. W ramach implementacji dostarczone będą macierze prezentujące kolejne etapy działania algorytmu oraz końcową macierz BLOSUM.

1. **Opis implementacji**

Implementacja algorytmu wyznaczania macierzy BLOSUM została wykonana w Języku C++ z wykorzystaniem elementów biblioteki Qt. Aby zapewnić uniwersalność przygotowanego rozwiązania zostało ono wykonane z wykorzystaniem mechanizmu szablonów. Szablon algorytmu jest parametryzowany:

1. Typem dla pojedynczego symbolu (np. char lub wchar)
2. Typem kontenera w którym znajduje się sekwencja
3. Typem stałopozycyjnym używanym do obliczeń (np. int lub long)
4. Typem zmiennopozycyjnym używanym do obliczeń (np. float lub double)

Podczas implementacji wykorzystano dość licznie klasy kontenerowe oparte o funkcje haszujące, co pozwoliło ograniczyć złożoność poszczególnych etapów do następujących wartości:

N – liczba sekwencji

L – długość sekwencji

M – liczba unikalnych symboli

Zliczanie par:

L\*(N + M2)

Obliczenie liczby wszystkich par:

1

Normalizacja liczby par:

M2

Obliczanie prawdopodobieństw:

M2

Estymacja prawdopodobieństw:

M2

Obliczenie logarytmów:

M2

Obliczenie macierzy BLOSUM:

M2

W przypadku białek M jest ograniczone i bardzo małe, przez co największy wkład w złożoność algorytmu ma krok pierwszy. Ogólna zatem złożoność wykonanej implementacji to w tym przypadku:

O(N\*L)

1. **Opis interfejsu graficznego**
2. **Wyniki przeprowadzonych testów**

Głównym parametrem, jaki można wykorzystywać przy pomiarach i analizie algorytmu prowadzącego do uzyskania macierzy BLOSUM jest długość sekwencji. Do przeprowadzenia testów wykorzystano zatem manipulację długością analizowanych sekwencji jak i ich ilością. Dane pobierane były z bazy danych <http://www.uniprot.org/uniprot/> - analizowane białka to:

* Albuminy
* Lizyny
* Keratyny
* Protaminy

Z każdej z nich do testów wybranych zostało około kilkudziesięciu białek.

Wszystkie pomiary (dla każdej długości sekwencji czy dla każdej ilości sekwencji) były wykonywane 10-krotnie a ich wyniki były uśredniane w celu uniknięcia nieścisłości wynikających z wahań czasów wykonania.

Sprzęt na jakim wykonywano testy:

Intel Core i7, Radeon HD 5800, 4 GB RAM, System – Linux Ubuntu 12.10

1. Wpływ długości sekwencji na czas obliczeń

Pierwszym z modyfikowanych parametrów jest długość sekwencji. Do celów testowych wykorzystano dane o łącznej liczbie 1400 sekwencji (im większa długość sekwencji tym danych było odpowiednio więcej). Wyniki w formie wykresu przedstawiono poniżej. Świadczą one o wyraźnym wpływie długości sekwencji na czas pomiarów. Największa różnica w wydajności ma swoje źródło w pierwszym kroku algorytmu, w którym następuje zliczenie ilości par aminokwasów. Ponieważ zliczanie to odbywa się kolumnami, im większa liczba kolumn (czyli większa liczba sekwencji) tym dłużej trwają obliczenia. Jest to krok algorytmu, który najlepiej nadaje się do możliwości zrównoleglenia obliczeń (na przykład z podziałem na procesory).

1. Wpływ ilości sekwencji na czas działania algorytmu

Drugim parametrem mającym wpływ na czas działania algorytmu jest ilość sekwencji. Podobnie jak w poprzednim przypadku jest to związane przede wszystkim z pierwszym krokiem algorytmu, w którym następuje zebranie informacji o częstości występowania par takich samych aminokwasów w poszczególnych kolumnach. Im dłuższa jest kolumna, w której trzeba dokonać obliczeń tym dłuższy jest czas działania całego algorytmu. Na poniższych diagramach przedstawiono wpływ ilości sekwencji na czas działania algorytmu. Analogicznie jak poprzednio wszystkie testy przeprowadzono 10-krotnie – ich wyniki zostały uśrednione.

1. Wpływ rodzaju podziału danych na czas działania algorytmu

Algorytm do tworzenia macierzy BLOSUM operuje przede wszystkim na ciągach aminokwasów. Można zatem ciąg wielu sekwencji traktować jako jedną długą sekwencję, którą można w dowolny sposób dzielić. Dzięki temu można rozważać, w jaki sposób podzielić ciąg aminokwasów na sekwencje aby uzyskać jak najbardziej optymalne wyniki (pod względem wydajności). Oczywiście z biologicznego punktu widzenia badanie takie nie ma znaczenia (pozycja aminokwasu ma znaczenie w kontekście rozważania sekwencji białkowych), z punktu widzenia informatycznego testy takie mogą być przydatne dla osób, które rozważają wydajność algorytmu, czy poszukują dróg jego optymalizacji. Decyzja, czy wprowadzać sekwencje dłuższe, ale mniej czy dużą liczbę krótkich sekwencji może mieć także znaczenie przy decydowaniu, jakie dane posłużą za dane testowe do tworzenia macierzy BLOSUM. Na poniższym diagramie przedstawiono wyniki działania dla sekwencji o łącznej liczbie około 40 tys. aminokwasów. Dzielone były one na coraz dłuższe sekwencje (co za tym idzie było ich coraz mniej).

Na przedstawionym diagramie nie można zauważyć żadnej korelacji między podziałem na sekwencje określonej długości a czasem wykonania algorytmu. Wszystkie wartości mieszczą się w zakresie 15-20 ms (za wyjątkiem pierwszej, granicznej wartości nieznacznie wykraczającej ponad 20 ms). Świadczy to o małym znaczeniu podziału dostarczonych danych wejściowych. Dzieje się tak, ponieważ pierwszy krok algorytmu wymaga iterowania zarówno po wszystkich elementach danej kolumny jak i po wszystkich kolumnach przynajmniej raz (w zaimplementowanej na potrzeby niniejszego projektu wersji). Oznaczając *m*  jako długość sekwencji i *n*  jako liczbę tych sekwencji:

*m* × *n* = const

Dalsze kroki algorytmu nie biorą już pod uwagę samych sekwencji a symbole w nich występujące. Powoduje to, że nie są one czułe na zmiany długości sekwencji czy zmianę ich ilości.

1. Wnioski

Algorytm do tworzenia macierzy BLOSUM wykorzystywany jest zazwyczaj jednokrotnie dla danego zestawu danych. Dane te są następnie wykorzystywane dla dalszych badań. Wzrost czasu działania algorytmu w zależności od ilości wprowadzonych danych powoduje, że w metodach bioinformatycznych (często operujących na milionach danych, na przykład sekwencji aminokwasów) używanie tego algorytmu wielokrotnie mogłoby być uciążliwe i czasochłonne. Krokiem najbardziej wpływającym na czas działania algorytmu jest krok pierwszy (zebranie informacji o częstości występowania poszczególnych par aminokwasów), wymaga on bowiem przeglądania wszystkich dostarczonych sekwencji. Dalsze kroki algorytmu w praktyce w ogóle nie zależą od danych wejściowych, są zatem dużo szybsze. Czynnikiem, który w największym stopniu wpływa na wydajność dalszych kroków algorytmu jest nie ilość dostarczonych sekwencji aminokwasów ale ich podobieństwo. Im mniej różnych symboli obecnych jest w danych wejściowych, tym mniej jest par aminokwasów, jakie należy rozważać a co za tym idzie zmniejsza się także macierz, na której algorytm operuje. Duża ilość różnych aminokwasów powoduje znaczący spadek wydajności. W czasie testowania algorytmu nie udało się niestety znaleźć danych reprezentatywnych dla różnej ilości, wnioski dotyczące wpływu różnorodności symboli zostały zatem wysnute na podstawie analizy schematu algorytmu do tworzenia macierzy BLOSUM.

Algorytm BLOSUM jest bardzo przydatnym narzędziem do badania podobieństw białek. Blokowość operacji może powodować konieczność wstępnej obróbki dostarczanych danych, z drugiej jednak strony pozwala na stworzenie łatwych, powtarzalnych kroków pozwalających na otrzymanie oczekiwanych rezultatów. Zalety macierzy BLOSUM powodują, że są one szeroko wykorzystywane w różnego rodzaju badaniach białek.