MBI DOKUMENTACJA WSTĘPNA

Adam Niziński Michał Przyłuski

9 grudnia 2009

Projekt ma na celu implementację algorytmu badającego podobieństwo dwu sekwencji o liniowej złożoności pamięciowej.

1 Algorytm

Badanie podobieństwa (globalnego) dwu sekwencji jest częstym i ważnym zadaniem, przed którym stoi biologia obliczeniowa, zwana także, bioinformatyką. Niestety, klasyczny algorytm Needlemana-Wunscha, oparty o ideę programowania dynamicznego, charakteryzuje się złożonością pamięciową rzędu $O(m \star n)$, gdzie m i n są odpowiednio długościami rozpatrywanych sekwencji. Algorytm ten został szczegółowo opisany w [1].

Warto jednak zauważyć, że w danym przebiegu pętli algorytmu Needlemana-Wunscha wykorzystywane są dane tylko z bezpośrednio poprzedzającej kolumny. Wynika to z faktu, iż ona właśnie zawiera najlepsze dopasowanie, które udało się znaleźć do bieżącej pozycji. Co więcej, algorytm ten nie wymaga „patrzenia w przyszłość", gdyż do wyznaczenia wartości dopasowania na danej pozycji, potrzebne są tylko informacje z jej najbliższego otoczenia. Ta prosta obserwacja, pozwoliła stworzyć algorytm badający podobieństwo dwu sekwencji o liniowej złożoności pamięciowej, który omówiono w [2]. Algorytm ten wykorzystuje fakt, iż kolumna, która już nie będzie wykorzystana zostaje zwolniona, a kolumny, które nie są jeszcze potrzebne, nie zostały jeszcze zaalokowane w pamięci. Takie podejście pozwoliło znacznie zmniejszyć wymagania pamięciowe algorytmu.

2 Funkcjonalność

Program będzie wczytywał dane wejściowe w formacie *FASTA*. Wybór taki został podyktowany prostotą formatu, która ułatwia wczytywanie danych. Nie potrzebujemy bowiem wszystkich dodatkowych (poza sekwencją) informacji zawartych w formacie *GenBank*.

3 Testowanie

Program zostanie przetestowany na pewnych dostępnych sekwencjach z bazy NCBI [3]. Planujemy testować program zarówno na sekwencjach „podobnych", jak i na losowych. Sekwencje podobne zostaną wybrane jako silnie konserwowane fragmenty genomu u blisko spokrewnionych organizmów lub homologiczne sekwencje (zbliżone u wielu organizmów) takie, jak gen COX1 kodujący oksydazę cytochromu c [4].

4 Interfejs

Interakcja użytkownika z programem będzie ograniczona do wskazania sekwencji do prównania. Użytkownik wywoła program podając jako jego dwa argumenty linii poleceń nazwy plików, w których znajdują się sekwencje w formacie *FASTA*. Następnie program wyznaczy wartość podobieństwa tych sekwencji, i wyświetli użytkownikowi wynik.

W dalszym etapie, po uzyskaniu podstawowej funkcjonalości, interfejs użytkownika zostanie rozszerzony, poprzez wykorzystanie django oraz pozostałych elementów z sugerowanego środowiska, aby umożliwić działanie aplikacji jako aplikacji internetowej.

Literatura

[1] J. Tiuryn: *Wstęp do biologii obliczeniowej*, notatki do wykładu, http://www.mimuw.edu.pl/~tiuryn

[2] Neil C. Jones, Pavel A. Pevzner: *Introduction to Bioinformatics Algorithms*, The MIT Press, 2004

[3] http://www.ncbi.nlm.nih.gov/

[4] http://www.ncbi.nlm.nih.gov/sites/homologene/5016