PRZEDMIOT OBIERALNY: BIOINFORMATYKA, LABORATORIUM

Wersja pliku: v 1.1, 13 kwietnia 2021
Autor: dr inż. Marcin Radom

Dzisiejszy materiał będzie jest krótki, ale też niezmiernie ważny: porównywanie dwóch DNA oraz co to właściwie jest DNA (takie prawdziwe, a nie wersja, którą sobie generujemy w programach).

PORÓWNYWANIE ŁAŃCUCHÓW ZNAKÓW

Dość popularnym podejście jest użycie miary odległości Levenshteina, która podaje nam, ile operacji trzeba wykonać, aby zmienić jeden łańcuch znaków w drugi – czyli im mniej tym lepiej.

Oczywiście nie trzeba wyważać otwartych drzwi i silić się na samodzielną implementację (choć można). Na przykład na pytanie o algorytm w C#, znajdziemy pod linkiem

https://stackoverflow.com/questions/6944056/c-sharp-compare-string-similarity

Znajdziemy kod z algorytmem. Oczywiście od razu z komentarzem, że da się lepiej, a to „lepiej" też jest pod linkiem dostępne (przynajmniej w C#-owej wersji. Ponieważ jednak algorytm w zasadzie jedyne obliczenia przeprowadza na tablicach dwuwymiarowych, konwersja kodu do każdego innego języka programowania nie powinna nikomu sprawić problemu.

Miar takich jest oczywiście więcej, zainteresowanych odsyłam tutaj:

https://itnext.io/string-similarity-the-basic-know-your-algorithms-guide-3de3d7346227

Istotne w tym momencie stają się jednak dwie sprawy. Po pierwsze, dysponując algorytmem miary podobieństwa, można w ramach testów sensownie oceniać jakość otrzymanych rekonstrukcji DNA. Druga sprawa jest taka, że miary są różne oraz mają różne punktacje. Należy więc zdecydować się na jedną konkretną miarę, umieścić informację o niej w sprawozdaniu i dokładnie wyjaśnić co oznaczają wartości z tej miary w kontekście: podobne/niepodobne.

DNA

Podejrzewam, że na ten moment Państwa algorytmy generowania DNA sprowadzają się do losowego wybierania jednej z czterech literek. Ok, tak też można, i testy wykonywane na tego typu „sekwencjach" mogą stanowić główną bazę testów w sprawozdaniu. Należy sobie jednak zadać pytanie, czy podobne wyniki jak na tego typu „konstrukcie" osiąga się na prawdziwych sekwencjach? Skoro zadaję tutaj to pytanie, to pewnie nie, jak się można domyślić.

Warto byłoby więc rozważyć pewne testy na sekwencjach mniej sztucznych. Źródeł jest bez liku. Na przykład na politechnicznym bioserwerze pod adresem

http://bio.cs.put.poznan.pl/research_fields/5193ee2f9dfb89b38b000001

Znajdują się archiwa sekwencji oraz spektr. Spod tego linku oczywiście należy zignorować spektra izotermiczne, one są do innego algorytmu i metodologii SBH. W tabelce 1 znajdują się tam spektra z błędami, warto

sprawdzić, jak algorytmy które skonstruujecie radzą sobie z takim materiałem (chodzi o spektra, choć same sekwencje do ściągnięcia też są interesujące).

Innym bardzo ciekawym materiałem jest najpopularniejsza ostatnio zaraza, czyli nasz ulubiony koronawirus, który niedawno obchodził roczek w Polsce. Na rozluźnienie rysunek okolicznościowy dla naszego ukochanego solenizanta, i niech lepiej nikt nie ośmieli się śpiewać „Sto lat":

Kod genetyczny SARC-CoV-2 dostępny tutaj:

https://www.ncbi.nlm.nih.gov/nuccore/NC_045512

oraz w wersji FASTA (sam kod bez podziałów na spacje i inne):

https://www.ncbi.nlm.nih.gov/nuccore/NC_045512.2?report=fasta

Wystarczy ściągnąć te 30 tysięcy szczęścia ukryte w czterech nukleotydach, losowo poszatkować na mniejsze fragmenty i potestować algorytmy. Czemu ma to znaczenie, tj. testy prawdziwego wirusowego DNA[1], czy w ogóle normalnych sekwencji. Na przykład tutaj:

https://www.ncbi.nlm.nih.gov/genome/guide/human/

mamy nasz kod genetyczny – wszystkie chromosomy, sekwencje DNA można sobie ściągnąć (tylko będzie liczona w gigabajtach, bo my jako ludzie DNA od długości trochę ponad 3 miliardów nukleotydów. W każdym razie, znowu spłycając temat: w DNA mamy geny, które tworzą białka. To pewnie większość wie / słyszała / pamięta. Ale to, że geny stanowią marny, jednocyfrowy procent DNA już pewnie wie mniej osób. O ile tak

[1] Koronawirus jest RNA-wirusem, ale różnica *na nasze potrzeby* sprowadzać się może tylko do tego, że RNA ma uracyl (U) zamiast tyniny (T) w DNA. Oczywiście DNA i RNA to zupełnie osobne wszechświaty, ale na potrzeby tego przedmiotu i testów nie musimy w to wnikać.

zwane CDS – sekwencje kodujące które kodują białka (każde 3 nukleotydy to jakiś aminokwas), i tam rozkład literek jest bardziej równomierny (zazwyczaj), to większość kodu genetycznego już taka fajna nie jest. Są tam liczne powtarzające się małe i duże fragmenty i cala masa innych nie-statystycznych utrudnień dla idei sekwencjonowania. Dlatego nie należy się dziwić, że prawdziwe DNA plus jeszcze błędy w spektrum powodują, że algorytmy rekonstrukcji mają naprawdę ciężki orzech do zgryzienia.

Dlatego też polecam sprawdzenie działania Państwa algorytmów także na prawdziwych sekwencjach. Oczywiście nie należy ich mieszać, właśnie dlatego, że o ile sztuczne DNA z równym rozkładem nukleotydów można traktować jako *easy-mode*, to w prawdziwym DNA można trafić na fragmenty, gdzie np. liczba błędów wynikających z powtórzeń może momentami iść tak ostro w górę, że przysłowiowo wybija w swojej wielkości dziury w suficie i sprawia, że algorytm sekwencjonowania, jaki by nie był, naprawdę będzie się męczyć. A i wyniki będą odpowiednio „interesujące”.