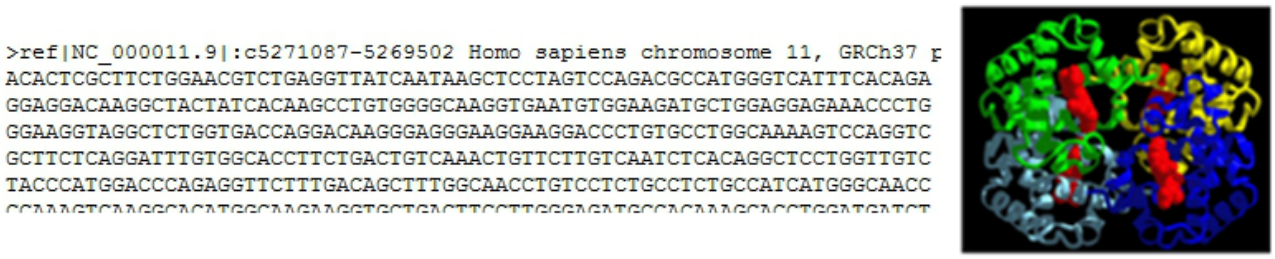
MATERIAŁY DO ZAJĘĆ
MŁODZIEŻOWYCH UNIWERSYTETÓW MATEMATYCZNYCH

Kraków, 2 marca 2013

BIOINFORMATYKA - NA STYKU INFORMATYKI, MATEMATYKI I BIOLOGII

Bioinformatyka jest nową, interdysciplinarną dziedziną, która powstała na skutek ogromnego postępu w dziedzinie technologii odczytujących informację biologiczną. Pojawiła się ona, jako odrębna dyscyplina, w latach 80-tych, ale na początku jej zadanie ograniczało się do projektowania i utrzymywania baz danych, tak aby ułatwić dostępność i analizę tych danych biologom, genetykom itd. Niedługo potem okazało się, że informatyka może zaproponować w tej dziedzinie dużo więcej i wraz z narzędziami matematycznymi zupełnie przeobraziła ramy nowej dyscypliny.

Obecnie bioinformatyka jest bardzo szeroką dziedziną, obejmującą między innymi analizę sekwencji DNA i RNA, struktry trójwymiarowej cząsteczek RNA, białek i kompleksów, analizę aktywności genów i ich sieci interakcji, symulacje dynamiki molekularnej i wiele innych [1, 2].



RYSUNEK 1. Fragment sekwencji DNA jednostki β hemoglobiny człowieka (z lewej), struktura trójwymiarowa hemoglobiny (z prawej)

Na wykładzie omówię kilka problemów z jakimi zmagają się bioinformatycy - opowiem o konstrukcji optymalnego dopasowania sekwencji z użyciem programowania dynamicznego jak również szybszego algorytmu heurystycznego o nazwie BLAST. Następnie przedstawię funkcjonał energii, który jest minimalizowany przy przewidywaniu struktury białka oraz symulowaniu dynamiki molekularnej. Na koniec wspomnę o modelach obliczeniowych konstruowanych na potrzeby analizy aktywności genów, gdzie wykorzystuje się zarówno statystykę jak i metody nauczania maszynowego.

Zagadnienie dopasowania sekwencji jest jednym z pierwszych problemów bioinformatyki. Wraz z rozwojem technologii naukowcy otrzymywali coraz więcej danych sekwencyjnych - sekwencji DNA, RNA czy sekwencji białek z różnych organizmów. Naturalnym, kolejnym krokiem było poszukiwanie podobieństwa wśrod tych sekwencji. Takie podobieństwo może świadczyć na przykład o powiązaniu ewolucyjnym organizmów lub też o podobnych funkcjach porównywanych sekwencji. Oczywiście przy dużej liczbie sekwencji sprawdzanie podobieństwa "ręcznie" byłoby bardzo nieefektywne, dlatego zdecydowano się na wykorzystanie komputerów. Algorytmy stworzone na celu porównania sekwencji musiały dodatkowo brać pod wagę różne typy zmian ewolucyjnych sekwencji np. możliwość zamiany jednego nukleotydu czy aminokwasu na inny (mutacja punktowa), wklejenie lub zniknięcie pewnego fragmentu sekwencji (insercja, delecja).

Najbardziej znanym algorytmem wyszukującym optymalne dopasowanie sekwencji jest algorytm Needlemana–Wunscha, który opiera się na tzw. programowaniu dynamicznym. Programowanie dynamiczne można zastosować do pewnej klasy problemów, które charakteryzuje tzw. własność optymalnej podstruktury tzn. że końcowe rozwiązanie problemu można skonstruować z jego mniejszych podproblemów. Algorytm Needlemana–Wunscha konstruuje optymalne dopasowanie dwóch sekwencji $s1$ i $s2$ budując tablicę, gdzie odpowiednio na przecięciu wiersza i oraz kolumny j (Tablica[i][j]) mamy zakodowane optymalne dopasowanie dla prefiksów $s1$ i $s2$ o długościach i oraz j. W nieco łatwiejszej wersji tego algorytmu możemy przyjąć, że jeśli znaki na pozycji i sekwencji $s1$ oraz pozycji j sekwencji $s2$ się zgadzają, to nagradzamy takie rozwiązanie dodatkowo jednym punktem, czyli jako wynik dopasowania w polu na przecięciu wiersza i oraz kolumny j wpisujemy Tablica[i-1][j-1]+1. W przeciwnym wypadku z dwóch sekwencji zakodowanych w pozycjach Tablica[i][j-1], Tablica[i-1][j] wybieramy tą lepiej punktowaną, a w drugiej wstawiamy przerwę. Ostatecznie, rozwiązanie problemu, czyli optymalne dopasowanie dla obu sekwencji jest zawarte w prawym, dolnym rogu tablicy [1, 2].

Algorytm Needlemana–Wunscha znajduje optymalne dopasowanie sekwecji, jednakże potrzebuje on stosunkowo dużo czasu i pamięci, przez co staje się dość niepraktyczny jeśli musimy porównać wiele, bardzo długich sekwencji. Czasami, jeśli bardzo zależy nam na szybkości i wydajności warto użyć tzw. algorytmów heurystycznych. Algorytmy te nie dają gwarancji optymalności rozwiązania, zwracają one rozwiązanie w pewnym sensie przybliżone, jednak ich siłą jest szybkość i wydajność. Przykładem takiego rozwiązania dla problemu dopasowania sekwencji jest algorytm BLAST. Algorytm ten służy do znalezienia w zbiorze sekwencji takiej sekwencji lub jej fragmentu, aby była ona możliwie najbardziej podobna do sekwencji wyszukiwanej. Zakładając, że fragmentów podobnych będzie niewiele możemy je szybko przefiltrować wykorzystując znaną z matematyki dyskretnej zasadę szufladkową Dirichleta. Załóżmy, że mamy 9-literowe fragmenty dwóch sekwencji, które różnią się dwoma pozycjami. Z zasady szufladkowej, wiemy, że jeśli jedna szufladka oznaczałaby trzyliterowy fragment, to istnienieje co najmniej jedna szufladka, gdzie wszystkie pozycje będą zgodne. Bazując na tym spostrzeżeniu BLAST szuka takich krótkich, dokładnych dopasowań. Jeśli jest ich odpowiednio wiele, to stanowią one zaczątek dopasowania, które jest następnie poszerzane w sposób dynamiczny. Dzięki takiemu heurystycznemu, szybkiemu wyszukiwaniu BLAST jest w stanie przeszukać całą bazę dostępnych sekwencji w najwyżej kilka minut. Webserwis umożliwiający użycie tego algorytmu jest dostępny dla każdego na stronie blast.ncbi.nlm.nih.gov [3].

Jedną z największych, do tej pory nierozwiązanych zagadek bioinformatyki, jest przewidywanie trójwymiarowej struktury białka. Pomimo ponad 50 lat pracy wielu badaczy, postęp w tej dziedzinie jest niewielki [3]. Algorytm wykorzystywany przez naturę do tworzenia trójwymiarowych struktur odpowiedzialnych za funkcję białka wciąż pozostaje nieodkryty. Zrozumienie tego mechanizmu pozwoliłoby na ogromny rozwój szczególne w medycynie, ponieważ w konsekwencji prowadziłby on do projektowania białek o określonych funkcjach np. leku na daną chorobę. Opracowane do tej pory metody bazują na pewnej funkcji energii, która opisuje daną trójwymiarową strukturę. Bierze ona pod uwagę długości wiązań oraz kąty pomiędzy atomami, oddziaływania elektrostatyczne oraz oddziaływania dipol-dipol (nazywane oddziaływaniami van der Waals'a). Następnie w sposób stochastyczny poszukiwane jest minimum globalne tej funkcji, ponieważ przyjmuje się, ze w naturze procesy spontanicznie dążą do stanu minimum energetycznego. Dla tak skomplikowanych funkcji znalezienie minimum globalnego jest oczywiście bardzo trudne, więc w praktyce znalezione zostaje pewne minimum lokalne. Okazuje się jednak, że struktura opisywana przez znalezione minimum lokalne rzadko przypomina strukturę danego białka w naturze. Metoda ta sprawdza się relatywnie dobrze tylko dla małych białek - rzędu 100-150 aminokwasów. Co dwa lata organizowany jest konkurs o nazwie CASP (Critical Assessment of protein Structure Prediction), który ocenia postępy w tej dziedzinie. Biorą w nim oczywiście udział naukowcy zajmujący się tym problemem, ale konkurs jest otwarty i każdy może wziąć w nim udział.

W ostatnich latach nastąpił wyjątkowo szybki postęp w dziedzinie technologii, które mierzą aktywność genów w komórce. Przykładem takiej technologii są mikromacierze, które umożliwiają zmierzenie aktywności kilkuset tysięcy genów pacjenta w jednym badaniu. Wyzwaniem dla bioinformatyki jest analiza tych danych, która może służyć na przykład do diagnozowania stanu zdrowia, doboru odpowiedniego schematu leczenia czy oceniania postępów terapii. Jednym z pierwszych podejść jest oczywiście analiza czysto statystyczna. Załóżmy, że chcielibyśmy odnaleźć geny, które są powiązane z daną chorobą. Możemy to zrobić używając np. testu statycznego T-Studenta dla dwóch grup pacjentów: grupy chorych i zdrowych, wśród których zmierzono poziom aktywności genów. Ten test statystyczny pozwoli nam wychwycić te geny, dla których istnieje statystyczne uzasadnienie by sądzić, ze ich średnie aktywności w grupie chorych i zdrowych są różne.

Poza klasycznymi metodami statystycznymi duży wkład w analizę danych aktywności genów w ostatnim czasie mają metody nauczania maszynowego. Przy ich użyciu można np. automatycznie pogrupować geny ze względu na podobieństwo ich profilu aktywności oraz otrzymać pewną hierarchiczną strukturę tych grup. Buduje się również modele nazywane klasyfikatorami, które, po wcześniejszym nauczeniu, są w stanie automatycznie przyporządkować klasę danemu pacjentowi (może to być etykieta - zdrowy lub chory, bądź stadium choroby). Wielu badaczy próbuje wykorzystać dane o aktywności genów dla stworzenia modeli sieci interakcji genów - tworzenie takich modeli wyodrębniło poddziedzinę bioinformatyki o nazwie biologia systemów.

Podsumowując bioinformatyka jest niezwykle ciekawą, interdyscyplinarną nauką, która łaczy biologię z matematyką, informatyką oraz innymi naukami przyrodniczymi. Już teraz wpływ jej wyników na inne dziedziny takie jak genetyka, medycyna czy farmakologia wydaje się być ogromny, a to dopiero początek. Jej rozwój może przynieść wymierne korzyści zarówno dla nauki, jak i środowiska czy społeczeństwa.

Literatura

[1] Higgs P., Attwood T., *Bioinformatyka i ewolucja molekularna*, PWN 2012.

[2] Baxevanis A.D., Ouellette B.F., *Bioinformatyka*, PWN 2005.

[3] National Center for Biotechnology Information: www.ncbi.nlm.nih.gov - publicznie dostępna baza informacji biologicznej.

[4] Dill K., MacCallum J., *The Protein-Folding Problem, 50 Years On*, Science 23 November 2012: 338 (6110), 1042-1046.

[5] Alterovitz G., Kellis M., Ramoni M., *Bioinformatics and Proteomics*, Massachusetts Institute of Technology: MIT OpenCourseWare, 2005, http://ocw.mit.edu.

Ewa Matczyńska