Spis treści

5 Progresywne uliniowienie

Metoda ta została zaproponowana przez Fenga i Doolitla w 1987r. Stanowi podstawę do dalszych ulepszeń, które zaowocowały m.in. dwoma popularnymi heurystycznymi programami używanymi w praktyce do uliniowienia wielu sekwencji. Programy te to CLUSTAL W oraz T-Coffee. Zostaną one nieco omówione w tym rozdziale. Zainteresowanego czytelnika odsyłam do publikacji na ten temat (pdfy znajdują się na stronie tego wykładu).

Pomysł progresywnego uliniowienia polega na sukcesywnym budowaniu uliniowienia, zaczynając od pojedynczych sekwencji, następnie budując uliniowienia uliniowień, itd. Kolejność budowy uliniowień wyznaczona jest przez pewne drzewo filogenetyczne zbudowane na podstawie danych sekwencji. Więcej szczegółów poniżej.

5.1 CLUSTAL W

Jest to program powszechnie używany do multiuliniowień. Powstał w 1994r. (Thompson, Higgins, Gibson). Budowa uliniowienia dla sekwencji $S_1, \ldots, S_k$ odbywa się w trzech krokach:

1. *Uliniowienia parami:* prównujemy każdą parę sekwencji S_i z S_j (dla $i \neq j$). Jako wartość porównania przyjmujemy liczbę dopasowanych m-słów w najlepszym uliniowieniu (parametr m jest równy 1-2 dla białek oraz 2-4 dla kwasów DNA lub RNA) pomniejszoną o pewną wartość kary za wprowadzone przerwy. Dopuszcza się dwie metody znajdowania wartości takiego porównywania: przybliżoną (ale za to szybką) oraz dokładną (dynamiczne programowanie, afiniczna funkcja kary za przerwy). Otrzymaną wartość dzieli się przez liczbę pozycji porównywanych (z wyłączeniem pozycji zawierających spacje) oraz odejmuje od wartości 1. W ten sposób dostajemy średnią liczbę różnic na jedną pozycję.

2. *Drzewo filogenetyczne:* na podstawie tablicy porównań otrzymanej w poprzednim kroku budujemy drzewo filogenetyczne (metoda najbliższego sąsiada). Drzewo to jest bez korzenia, ale każda krawędź ma przypisaną długość odpowiadającą odległości ewolucyjnej. Korzeń w takim drzewie umieszczamy tak, aby średnie odległości do liści po obu stronach korzenia były równe. W ten sposób otrzymujemy ukorzenione drzewo filogenetyczne z długościami przypisanymi krawędziom. Przy pomocy tego drzewa przypisujemy każdej sekwencji pewną wagę: niech S będzie sekwencją, która jest etykietą liścia v i niech $v_0, \ldots, v_n$ będzie drogą w drzewie od korzenia v_0 do liścia $v_n = v$. Niech d_i, dla $0 \leq i < n$, będzie długością krawędzi $\langle v_i, v_{i+1} \rangle$ oraz niech L_i będzie liczbą liści widocznych z v_{i+1}. Wówczas waga S jest równa $w_S = \sum_{i=0}^{n-1} d_i/L_i$. W ten sposób sekwencje odległe ewolucyjnie od reszty dostają największą wagę.

3. *Progresywne uliniowienie:* kolejność uliniawiania jest następująca. Dla każedgo wierzchołka wewnętrznego v w drzewie filogenetycznym będziemy budować uliniowienie $S_{i_1}, \ldots, S_{i_p}$, gdzie sekwencje $S_{i_1}, \ldots, S_{i_p}$ są wszystkimi sekwencjami widocznymi z v. Ciężar wierzchołka v jest to suma długości dróg z v do wszystkich liści. Konstruujemy uliniowienia odpowiadające wierzchołkom w kolejności wielkości ich wag (od najmniejszej do największej). Za każdym razem konstruujemy uliniowienie uliniowień. Wartość takiego uliniowienia jest obliczana według następujących zasad: uliniowienia kolumny symboli $x_1, \ldots x_{n_1}$ z kolumną symboli $y_1, \ldots, y_{n_2}$ daje wartość

$$\sum_{i=1}^{n_1} \sum_{j=1}^{n_2} s(x_i, y_j) \cdot w_{S_{x_i}} \cdot w_{S_{y_j}},$$

gdzie S_{x_i} jest sekwencją, z której pochodzi litera x_i. Podobnie S_{y_j}. Wartość $s(a, b)$ pochodzi z tablicy substytucyjnej przeskalowanej tak aby $s(a, b) > 0$ dla a, b będących literami, oraz $s(a, b) = 0$ jeśli a lub b jest spacją. Wprowadzone przerwy w uliniowieniach są zachowywane na dalszych etapach konstrukcji. Kary za otwarcie przerwy i kontynuację są zmieniane w zależności od położenia przerwy (mniejsze tam gdzie już jest przerwa lub np. w ciągu aminokwasów hydrofilowych, większe gdy jest otwierana nowa przerwa blisko już istniejącej innej przerwy). Reguły karania za przerwy są skomplikowane. Tablice substytucyjne zmienia się w zależności od tego czy mamy sekwencje zbliżone ewolucyjnie (początek konstrukcji) czy też są odległe. Szczegóły w pracy na stronie wykładu.

5.2 T-Coffee

Pomysł programu polega na dalszym ulepszeniu metody progresywnego uliniowienia. Powstał w 2000r. (Notredame, Higgins, Heringa). Dane sekwencje $S_1, \ldots, S_k$. Główne kroki programu są następujące.

1. Każda para S_i, S_j jest globalnie uliniowiona (przy użyciu CLUSTAL W) oraz lokalnie uliniowiona (przy użyciu programu Lalign z pakietu FASTA). Z lokalnego uliniowienia każdej pary S_i, S_j bierze się 10 najlepszych nieprzecinających się lokalnych uliniowień. Powstają w ten sposób dwie biblioteki uliniowień.

2. Każde uliniowienie dostaje wagę równą liczbie dopasowań (czyli par identycznych symboli) pomnożoną przez 100 i podzieloną przez liczbę par w uliniowieniu nie zawierających spacji.

3. Następnie obie biblioteki są scalane w jedną. Powtarzające się pary uliniowień pochodzące z obu bibliotek są łączone z wagą będącą sumą wag. Tak otrzymaną bibliotekę nazywamy główną.

4. Zasadniczy pomysł podejścia zaproponowanego w T-Coffee polega na tworzeniu tzw. rozszerzonej biblioteki. W ustalonym uliniowieniu pochodzącym z biblioteki głównej każda para odpowiadających sobie aminokwasów ma tę samą wagę (jest to waga całego uliniowienia). W bibliotece rozszerzonej dla ustalonej pary sekwencji A, B mamy do czynienia z sytuacją, gdy jeden aminokwas (identyfikowany przez pozycję) z sekwencji A jest 'parowany' z różnymi aminokwasami (tzn. pozycjami) z sekwencji B o różnych wagach. W tym celu budujemy uliniowienia wszystkich trójek A, C, B, gdzie C przebiega wszystkie sekwencje zadane na wejściu (za wyjątkiem A oraz B, oczywiście). Dla takiej trójki przyjmujemy wagę uliniowienia równą minimum wag w_1 oraz w_2, gdzie w_1 jest wagą dla A, C, a w_2 jest wagą dla C, B. Bierzemy pod uwagę tylko te pozycje (aminokwasy) z A i B, które są wspólnie sparowane z jakimś symbolem z C. Waga dla takiej pary aminokwasów to suma wag: jedna pochodząca z A, B i druga pochodząca z trójki A, C, B. W ten sposób dla danej pary sekwencji A, B dostajemy tablicę substytucyjną (dla par aminokwasów, które nie pojawiły się dajemy wagę 0). Przy użyciu tej tablicy możemy poprawić uliniowienie dla A i B (metodą dynamicznego programowania).

5. Dalej stosujemy metodę progresywnego uliniowienia w sposób standardowy, tak jak to było opisane dla CLUSTAL W. Pozostałe szczegóły można znaleźć w pracy, która znajduje się na stronie wykładu.