Spis rzeczy

5 Uliniowienie wielu sekwencji **68**
5.7 Wielomianowy schemat aproksymacji 68

5 Uliniowienie wielu sekwencji

5.7 Wielomianowy schemat aproksymacji

Zajmiemy się ponownie problemem optymalnego uliniowienia z drzewem filogenetycznym. Przyjmiemy następujące założenia:

- Funkcja podobieństwa s jest odległością.
- Ustalamy stałą d, tak że wszystkie rozważane drzewa mają rząd nie większy niż d (tzn. rząd każdego wierzchołka jest nie większy od d).

Celem będzie skontruowanie wielomianowego schematu aproksymacji, który będzie znajdował uliniowienie z drzewem filogenetycznym T z zadaną dokładnością ε. Dla każdego ε (i dla każdego d) algorytm ma działać w czasie wielomianowm. Jednak zależność od ε i od d będzie wykładnicza.

Przyjmiemy następujące oznaczenia. Niech T będzi drzewem filogenetycznym o etykietach liści $S_1, \ldots, S_k$. Niech $T^{\#}$ będzie optymalnym uliniowieniem dla T (tzn. kształty $T^{\#}$ i T są identyczne oraz etykiety liści w obu drzewach są takie same). Każdy wierzchołek wewnętrzny v w $T^{\#}$ ma pewną etykietę $T^{\#}(v) \in \Sigma^*$. Waga drzewa $T^{\#}$ jest to

$$\overline{sim}(T^{\#}) = \sum_{(u,v)\in E} \overline{sim}(T^{\#}(u), T^{\#}(v)),$$

gdzie E jest relacją sąsiedztwa wierzchołków w $T^{\#}$.

Pojęcie wagi stosuje się również do fragmentów uliniowienia $T^{\#}$. Na przykład, jeśli $\Pi = v_1, \ldots, v_m$ jest drogą w $T^{\#}$, to

$$\overline{sim}(\Pi) = \sum_{i=1}^{m-1} \overline{sim}(T^{\#}(v_i), T^{\#}(v_{i+1})).$$

Dla wierzchołka v drzewa T, przez T_v będziemy oznaczać poddrzewo T zaczepione w v. Dla $t \in \mathbb{N}$, przez $T_{v,t}$ będziemy oznaczać *górną część* poddrzewa T_v zawierającą wszystkie wierzchołki oddalone od v o nie więcej niż

t krawędzi (czyli w należy do $T_{v,t}$, gdy istnieje droga $v_0, \ldots, v_m$, taka że $v_0 = v$, $v_m = w$ oraz $m \leq t$. Przez $L(T_{v,t})$ będziemy oznaczać zbiór wszystkich liści fragmentu $T_{v,t}$ (tzn. takie wierzchołki z $T_{v,t}$, których rząd w $T_{v,t}$ jest równy 1). Przez $I(T_{v,t})$ będziemy oznaczać zbiór wszystkich wierzchołków wewnętrznych $T_{v,t}$.

Niech teraz $T^{\#}$ będzie optymalnym uliniowieniem dla T. Dla każdego wierzchołka v, niech $\ell(v)$ będzie liściem o takiej własności, że dla każdego liścia w, jeśli Π_w jest drogą w $T^{\#}$ od v do w, to

$$\overline{sim}(\Pi_{\ell(v)} \leq \overline{sim}(\Pi_w).$$

Stosując podobne metody do tych z Sekcji 5.5 można łatwo pokazać, że jeśli T^{ℓ} jest uliniowieniem dla T, w którym wierzchołkowi v przypisana jest etykieta $T(\ell(v))$, to

$$\frac{\overline{sim}(T^{\ell})}{\overline{sim}(T^{\#})} \leq 2.$$

Niech v będzie wierzchołkiem drzewa T, różnym od korzenia i niech $t \in \mathbb{N}$. Rozważmy fragment $T_{v,t}$, w którym każdemu wierzchołkowi $w \in L(T_{v,t}) \cup \{v\}$ przyporządkowujemy etykietę $T(\ell(w))$. Etykiety pozostałych wierzchołków w $T_{v,t}$ dobieramy tak, aby w ten sposób otrzymane drzewo etykietowane $T'_{v,t}$ minimalizowało wagę $\overline{sim}(T'_{v,t})$ (użyjemy metody dynamicznego programowania).

Niech teraz v^* będzie korzeniem T i niech $0 \leq i < t$. Rozważmy fragment $T_{v^*,i}$, w którym każdemu liściowi $w \in L(T_{v^*,i})$ przyporządkowujemy słowo $T(\ell(w))$ (czyli korzeniowi nie przyporządkowujemy żadnej etykiety). Ponownie, etykiety pozostałych wierzchołków w $T_{v^*,i}$ dobieramy tak, aby zminimalizować wagę. Niech $T'_{v^*,i}$ oznacza tak otrzymane drzewo.

Dla $0 \leq i < t$, uliniowienie $T^{(i)}$ drzewa T definiujemy jako następującą sumę

$$T^{(i)} = T'_{v^*,i} \cup \bigcup_{v \in G_i} T_{v,i},$$

gdzie G_i jest zbiorem wszystkich wierzchołków drzewa T, których odległość j od korzenia spełnia warunek $i \equiv j(\mod t)$ (czyli $j = i + mt$, dla pewnego $m \geq 0$). Pokażemy, że dla pewnego $i < t$, wartość $\overline{sim}(T^{(i)})$ dostatecznie dobrze aproksymuje $\overline{sim}(T^{\#})$.

Dla każdego wierzchołka v, niech $\rho(v) = \overline{sim}(\Pi_{\ell(v)})$ będzie wagą drogi Π w drzewie $T^{\#}$. Niech $h(v) = \sum_{i=1}^{d} \overline{sim}(T^{\#}(v), T^{\#}(v_i))$, gdzie $v_1, \ldots, v_d$ są wszystkimi synami v.

Lemat 5.7.1 *Niech v będzie wierzchołkiem różnym od korzenia i niech $v_1, \ldots, v_d$ będą wszystkimi synami w T wierzchołka v. Wówczas*

$$\overline{sim}(T'_{v,t}) \leq \sum_{u \in I(T_{v,t})} h(u) + \sum_{w \in L(T_{v,t})} \rho(w) + h(v) + \sum_{i=1}^{d} \rho(v_i).$$

Dowód: Rozważmy fragment $T_{v,t}$ i niech $T''_{v,t}$ będzie drzewem etykietowanym słowami w następujący sposób. Każdy wierzchołek $w \in L(T_{v,t}) \cup \{v\}$ ma etykietę $T(\ell(w))$, a każdy pozostały wierzchołek w w $T_{v,t}$ ma etykietę $T^{\#}(w)$. Niech $w \in L(T_{v,t})$ będzie dowolnym wierzchołkiem i niech u będzie ojcem w. Mamy wówczas

$$\begin{aligned}\overline{sim}(T^{\#}(u), T(\ell(w))) &\leq \overline{sim}(T^{\#}(u), T^{\#}(w)) + \overline{sim}(T^{\#}(w), T(\ell(w))) \\ &= \overline{sim}(T^{\#}(u), T^{\#}(w)) + \rho(w).\end{aligned}$$

Zatem sumując po wszystkich krawędziach w $T'_{v,t}$ dostajemy

$$\begin{aligned}\overline{sim}(T''_{v,t}) \leq \sum_{i=1}^{d} \overline{sim}(T(\ell(v)), T^{\#}(v_i)) + \sum_{u \in I(T_{v,t}) - \{v\}} h(u) \\ + \sum_{w \in L(T_{v,t})} \rho(w).\end{aligned} \tag{5.5.1}$$

Ponadto mamy

$$\begin{aligned}\sum_{i=1}^{d} \overline{sim}(T(\ell(v)), T^{\#}(v_i)) &\leq \sum_{i=1}^{d} [\overline{sim}(T^{\#}(v_i), T^{\#}(v)) + \overline{sim}(T^{\#}(v), T(\ell(v)))] \\ &\leq h(v) + d \cdot \rho(v).\end{aligned}$$

Zatem, z (5.5.1) dostajemy

$$\overline{sim}(T''_{v,t}) \leq \sum_{u \in I(T_{v,t})} h(u) + \sum_{w \in L(T_{v,t})} \rho(w) + d \cdot \rho(v).$$

Ponieważ $\rho(v) \leq \overline{sim}(T^{\#}(v), T^{\#}(v_i)) + \rho(v_i)$ zachodzi dla $i = 1, \ldots, k$, to

$$d \cdot \rho(v) \leq \sum_{i=1}^{d} [\overline{sim}(T^{\#}(v), T^{\#}(v_i)) + \rho(v_i)] = h(v) + \sum_{i=1}^{d} \rho(v_i).$$

Zatem

$$\overline{sim}(T''_{v,t}) \leq \sum_{u \in I(T_{v,t})} h(u) + \sum_{w \in L(T_{v,t})} \rho(w) + h(v) + \sum_{i=1}^{d} \rho(v_i).$$

Ponieważ $\overline{sim}(T'_{v,t}) \leq \overline{sim}(T''_{v,t})$, to dostajemy konkluzję lematu. ∎

Lemat 5.7.2 *Jeśli v_* jest korzeniem T, to dla każdego $i < t$, mamy*

$$\overline{sim}(T'_{v_*,i}) \leq \sum_{u \in I(T_{v_*,i})} h(u) + \sum_{w \in L(T_{v_*,i})} \rho(w).$$

Dowód: Podobnie jak w dowodzie Lematu 5.7.1 konstruujemy etykietowane drzewo $T''_{v_*,i}$, z tą różnicą, że korzeniowi v_* nie przypisujemy etykiety $T(\ell(v_*))$, ale $T^{\#}(v_*)$. Wówczas nierówność analogiczna do (5.5.1) wygląda następująco

$$\overline{sim}(T''_{v_*,i}) \leq \sum_{u \in I(T_{v_*,i})} h(u) + \sum_{w \in L(T_{v_*,i})} \rho(w).$$

Ponieważ $\overline{sim}(T'_{v_*,i}) \leq \overline{sim}(T''_{v_*,i})$, to dostajemy tezę lematu. ∎

Lemat 5.7.3 *Jeśli T nie zawiera wierzchołków rzędu 1, to*

$$\sum_{v \in T} \rho(v) \leq \overline{sim}(T^{\#}).$$

Dowód: Jeśli T nie zawiera wierzchołków rzędu 1 oraz k jest liczbą liści w T, to łatwo jest pokazać przez indukcję ze względu na k, że liczba wierzchołków wewnętrznych jest równa $k-1$. Co więcej, dla każdego wierzchołka wewnętrznego v istnieje droga Π_v w T od v do pewnego liścia, tak że dla $v \neq u$, drogi Π_v i Π_u są rozłączne. Zatem mamy

$$\sum_{v \in T} \rho(v) = \sum_{v \in I(T)} \rho(v) \leq \sum_{v \in I(T)} \overline{sim}(\Pi_v) \leq \overline{sim}(T^{\#}).$$

To kończy dowód. ∎

Lemat 5.7.4 *Jeśli T nie zawiera wierzchołków rzędu 1, to dla każdego $t \in \mathbb{N} - \{0\}$ istnieje $i < t$, takie że*

$$\overline{sim}(T^{(i)}) \leq (1 + 3/t) \cdot \overline{sim}(T^{\#}).$$

Dowód: Weźmy dowolne $t \in \mathbb{N} - \{0\}$. Wówczas

$$\begin{aligned}\sum_{i=0}^{t-1} \overline{sim}(T^{(i)}) &= \sum_{i=1}^{t-1} \overline{sim}(T'_{v_*,i}) + \sum_{v \neq v_*} \overline{sim}(T'_{v,t}) \\ &\leq (t+1) \cdot \sum_{v \in T} h(v) + 2 \cdot \sum_{v \in T} \rho(v) \\ &\leq (t+1) \cdot \overline{sim}(T^{\#}) + 2 \cdot \overline{sim}(T^{\#}) = (t+3) \cdot \overline{sim}(T^{\#}).\end{aligned}$$

Pierwsza z powyższych nierówności wynika z Lematu 5.7.1 oraz Lematu 5.7.2. Druga natomiast wynika z Lematu 5.7.3.

Niech i bedzie takie, że $\overline{sim}(T^{(i)}) \leq \overline{sim}(T^{(j)})$, dla $j < t$. Wówczas

$$t \cdot \overline{sim}(T^{(i)}) \leq \sum_{j=0}^{t-1} \overline{sim}(T^{(j)}),$$

co kończy dowód. ∎

Zauważmy, że jeśli T jest drzewem filogenetycznym zawierającym wierzchołki rzędu 1 oraz T_1 jest otrzymane z T przez usunięcie wszystkich wierzchołków rzędu 1 (przypadek, gdy wszystkie wierzchołki wewnętrzne T mają rząd 1 jest trywialny), to $\overline{sim}(T^{\#}) = \overline{sim}(T_1^{\#})$ (ponieważ jeśli rząd v w T jest 1 i u jest jedynym synem v, to $T^{\#}(v) = T^{\#}(u)$). Zatem założenie o rzędzie wierzchołków w Lemacie 5.7.4 daje się ominąć.

Zajmiemy się teraz problemem znajdowania drzew $T'_{v,t}$ oraz $T'_{v_*,i}$. Prowadzi to do następującego zagadnienia.

Znajdowanie optymalnego uliniowienia z drzewem filogenetycznym

Dane drzewo filogenetyczne T, chcemy znaleźć optymalne uliniowienie $T^{\#}$. Zastosujemy metodę programowania dynamicznego. Niech $S_1, \ldots, S_k$ będą etykietami liści $l_1, \ldots, l_k$ drzewa T i niech $|S_j| = n_j$, dla $j = 1, \ldots, k$. Dla $\vec{i} = i_1, \ldots, i_k$ (gdzie $0 \leq i_j \leq n_j$, dla $j = 1, \ldots, k$) niech $V(i_1, \ldots, i_k)$ będzie wartością optymalnego uliniowienia dla drzewa filogenetycznego $T_{\vec{i}}$, mającego ten sam kształt co T i takiego, że etykietą liścia l_j jest $S_j[1..i_j]$. Oczywiście $V(0, \ldots, 0) = 0$ (etykiety wszystkich wierzchołków w uliniowieniu są równe ε).

Aby ułożyć równanie rekurencyjne rozważmy optymalne uliniowienie $T^{\#}_{\vec{i}}$ dla drzewa $T_{\vec{i}}$, gdzie $\vec{i} \neq \vec{0}$. Przyjmijmy, że wierzchołek v w $T^{\#}_{\vec{i}}$ ma przypisaną

etykietę S_v. Niech v będzie dowolnym wierzchołkiem i niech $v_1, \ldots, v_d$ będą wszystkimi synami v. Rozważmy d rozszerzeń $S_v^{(1)\#}, \ldots, S_v^{(d)\#}$ słowa S_v, takich że $S_v^{(i)\#}$ wchodzi jako część optymalnego uliniowienia dla pary słów S_v, S_{v_i} (tzn. istnieje rozszerzenie $S_{v_i}^{\#}$ takie, że $s(S_{v_i}^{\#}, S_v^{(i)\#}) = \overline{sim}(S_{v_i}, S_v)$). Wierzchołek v nazwiemy *pełnym*, gdy wszystkie słowa $S_v^{(1)\#}, \ldots, S_v^{(d)\#}$ kończą się literą z alfabetu Σ (w tym przypadku jest to oczywiście ta sama litera), czyli spacja nie występuje na końcu żadnego z tych słów. Wierzchołek, który nie jest pełny nazwiemy *niepełnym*. Niech $S_{l_i}^{(0)\#}$ będzie rozszerzeniem słowa S_i wchodzącym jako część optymalnego uliniowienia z etykietą ojca liścia l_i.

Jeśli wszystkie wierzchołki wewnętrzne są pełne, to obcinając ostatnią literę we wszystkich rozszerzeniach w całym drzewie dostajemy optymalne uliniowienie dla drzewa $T_{\vec{i'}}$, gdzie $\vec{i'} = i'_1, \ldots, i'_k$ jest takie, że dla przynajmniej jednego j mamy $i'_j = i_j - 1$ (każde takie j, że $S_{l_j}^{(0)\#}$ kończy się na literę; gdyby żadne takie słowo nie kończyło się na literę, to uliniowienie by nie było optymalne).

Załóżmy teraz, że pewien wewnętrzny wierzchołek v jest niepełny i niech v będzie możliwie najdalej od korzenia położonym takim wierzchołkiem (tzn. wszystkie wierzchołki wewnętrzne osiągalne z v i różne od v są pełne). Niech i będzie takie, że $S_v^{(i)\#}$ kończy się spacją. Niech v_i będzie i-tym synem v. Jeśli w $S_v^{(i)\#}$ obetniemy ostatnią spację oraz każdej etykiecie wierzchołków wewnętrznych poddrzewa zaczepionego w v_i obetniemy ostatnią literę, a w etykiecie S_j obetniemy ostatnią literę, gdy $S_j^{\#}$ kończy się literą, to tak otrzymane drzewo będzie optymalnym uliniowieniem dla $T_{\vec{i'}}$, gdzie $\vec{i'} = i'_1, \ldots, i'_k$ ma taką własność, że dla pewnego j, $i'_j = i_j - 1$.

W poniższym równaniu rekurencyjnym będziemy używać funkcji $f : I(T) \longrightarrow \Sigma$, która reprezentuje stan uliniowienia taki, że ostatnią literą etykiety wierzchołka v jest $f(v)$. Dla $\vec{\varepsilon} = \varepsilon_1, \ldots, \varepsilon_k \in \{0,1\}^k$ przyjmijmy następujące oznaczenie: dla dowolnego liścia l_j definiujemy

$$f_{\vec{i}}^{\vec{\varepsilon}}(l_j) = \begin{cases} S_j(i_j), & \text{gdy } \varepsilon_j = 1 \\ -, & \text{gdy } \varepsilon_j = 0 \end{cases}$$

Dla wewnętrznego wierzchołka v przyjmujemy $f_{\vec{i}}^{\vec{\varepsilon}}(v) = f(v)$.

W sytuacji, gdy wszystkie wierzchołki są pełne mamy następujące równanie.

$$V_1(i_1, \ldots, i_k) = \min_{\vec{\varepsilon} \in A_1} \min_{f : I(T) \longrightarrow \Sigma} V(i_1 - \varepsilon_1, \ldots, i_k - \varepsilon_k) + \sum_{v \in I(T)} \sum_{j=1}^{d_v} s(f(v), f_{\vec{i}}^{\vec{\varepsilon}}(v_j)),$$

gdzie

$$A_1 = \{\vec{\varepsilon} \in \{0,1\}^k - \{(0, \ldots, 0)\} \mid \text{ dla } j = 1, \ldots, k : i_j - \varepsilon_j \geq 0\}.$$

Natomiast, gdy v jest minimalnym wewnętrznym wierzchołkiem niepełnym, to mamy równanie:

$$V_2(i_1,\ldots,i_k) = \min_{\vec{\varepsilon}\in A_2}\min_{1\le j\le d_v}\min_{f:I(T)\longrightarrow\Sigma} V(i_1-\varepsilon_1,\ldots,i_k-\varepsilon_k) + s(-,f(v_j)) + \sum_{u\in I(T_{v_j})}\sum_{m=1}^{d_u} s(f(u),f^{\vec{\varepsilon}}_{\vec{i}}(u_m)),$$

gdzie

$$A_2 = \{\vec{\varepsilon}\in\{0,1\}^k - \{(0,\ldots,0)\} \mid \text{ dla } j=1,\ldots,k : i_j-\varepsilon_j\ge 0 \text{ oraz jeśli } \varepsilon_j=1, \text{ to } l_j\in T_v\}.$$

Ostatecznie przyjmujemy

$$V(i_1,\ldots,i_k) = \min[V_1(i_1,\ldots,i_k), V_2(i_1,\ldots,i_k)].$$

Jeśli będziemy systematycznie budować tablicę $V(\vec{i})$ (jej rozmiar jest $O(n^k)$), to czas potrzebny na obliczenie $V_1(\vec{i})$ i $V_2(\vec{i})$, zgodnie z równaniami, wynosi $O(2^k\cdot(|\Sigma|+1)^{|T|}\cdot|T|) = O(c^{|T|})$, dla pewnej stałej c. Zatem czas obliczenia całej tablicy wynosi $O((cn)^{|T|})$.

Tak więc udowodniliśmy następujący wynik.

Twierdzenie 5.7.5 *Istnieje stała $c>1$, taka że problem znajdowania optymalnego uliniowienia dla drzewa filogenetycznego T daje się rozwiązać w czasie $O((cn)^{|T|})$, gdzie n jest ograniczeniem górnym na długość etykiet liści w T.*

Drugi problem polegający na znalezieniu optymalnego uliniowienia dla danego drzewa filogenetycznego, w którym ustalona jest etykieta korzenia rozwiązuje się całkowicie analogicznie do wyżej opisanego zagadnienia. Pozostawiamy to jako zadanie.

Zadanie 5.7.1 Dane drzewo T z etykietami pewnych wierzchołków (nie tylko liści). Dobrać etykiety pozostałych wierzchołków tak, aby otrzymać drzewo o minimalnej wadze. Czas powinien być $O((cn)^{|T|})$, gdzie n jest ograniczeniem długości wszystkich etykiet w T.

Wielomianowy schemat aproksymacji

Rozwżania tej Sekcji prowadzą do następującego wielomianowego schematu aproksymacji. Niech T będzie drzewem filogenetycznym i niech $\varepsilon>0$.

- Wybieramy $t \in \mathbb{N}$ tak, aby $3/t < \varepsilon$.

- Niech v^* będzie korzeniem T. Dla każdego $0 \leq i < t$ rozważamy fragment $T_{v^*,i}$. Jest on rozmiaru $O(d^t)$ i ma $O(d^t)$ liści. Ponieważ nie znamy wartości funkcji $\ell(v)$ dla $v \in I(T)$, to musimy wypróbować wszystkie możliwe przyporządkowania etykiet liści w T, liściom w $T_{v^*,i}$ (oczywiście dla liścia $w \in L(T_{v^*,i})$ wystarczy tylko rozważać etykiety znajdujące się w T_w). Dla każdego takiego przyporządkowania znajdujemy optymalne uliniowienie $T'_{v^*,i}$ (stosujemy tu Twierdzenie 5.7.5). Ostatecznie wybieramy takie przyporządkowanie etykiet, dla którego waga $T'_{v^*,i}$ jest najmniejsza. Ponieważ takich przyporządkowań jest co najwyżej $\binom{k}{d^t}$, to ten krok zajmuje $O(\binom{k}{d^t} \cdot (cn)^{d^t})$ czasu. Ponieważ d oraz t są ustalone, to czas zużyty w tym kroku algorytmu zależy wielomianowo od k i od n.

- dla $v \in G_i$ podobnie postępujemy znajdując drzewa $T'_{v,t}$. Ten krok też zużywa czas wielomianowo zależny od k i od n.

- Sumujemy wagi optymalnych uliniowień dla $T'_{v^*,i}$ oraz $\{T'_{v,t} \mid v \in G_i\}$. Zauważmy, że $|G_i| = O(|T|/t)$. W ten sposób obliczamy wagę drzewa $T^{(i)}$.

- Wybieramy drzewo $T^{(i)}$ o minimalnej wadze. Uliniowienie skonstruowane dla tego drzewa jest poszukiwaną aproksymacją optymalnego uliniowienia. Z Lematu 5.7.4 wynika, że błąd tego uliniowienia jest nie większy niż $3/t < \varepsilon$.

Mamy więc następujące twierdzenie.

Twierdzenie 5.7.6 (Wang, Jiang, Lawler'96)
Jeśli s jest odległością, to istnieje wielomianowy schemat aproksymacji, który dla dowolnego $\varepsilon > 0$ i dla dowolnego drzewa filogenetycznego T o rzędzie wierzchołków nie większym od d, znajduje uliniowienie T' dla T takie, że jeśli $T^\#$ jets optymalnym uliniowieniem dla T, to

$$\overline{sim}(T') \leq (1+\varepsilon) \cdot \overline{sim}(T^\#).$$

Ponadto czas działania tego algorytmu wynosi $O(\binom{k}{d^t} \cdot (cn)^{d^t} \cdot |T|)$, a zatem jest ograniczony przez wielomian od k, n oraz $|T|$.

Jak wynika z Twierdzenia 5.6.8, jeśli nie założymy ograniczoności rzędu drzewa filogenetycznego T oraz dopuścimy funcje s nie spełniające warunek trójkąta, to nie da się znaleźć wielomianowego schematu aproksymacji

dla problemu uliniowienia z drzewem T (chyba, że P=NP). Ciekawe zatem wydaje się pytanie o istnienie wielomianowego schematu aproksymacji dla dowolnych drzew filogenetycznych, przy założeniu, że s jest odległością.