‘‘Algorytmy biologii molekularnej’’
(J. Tiuryn, w13.tex, 13 marca 2000)

Spis rzeczy

6 Mapy fizyczne genomów **81**
6.1 Mapy restrykcyjne 81
6.1.1 Algorytm budowania mapy restrykcyjnej 84

6 Mapy fizyczne genomów

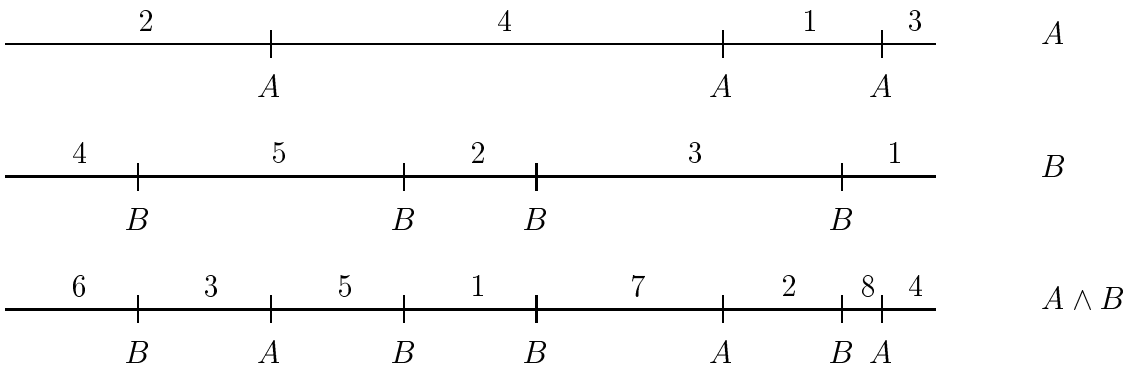
Mapa fizyczna genomu (lub jego fragmentu) przedstawia rozmieszczenie pewnych markerów wzdłóż tego fragmentu. Typowe zastosowanie map fizycznych polega na tym, że jeśli mamy zsekwencjonowany pewien fragment S, to mając mapę fizyczną chromosomu, z którego S pochodzi, możemy ustalić położenia markerów w S i w ten sposób dopasować miejsce położenia S na chromosomie. Omówimy problematykę związaną z *mapami restrykcyjnymi* i *mapami hybrydyzacyjnymi*.

6.1 Mapy restrykcyjne

Mapa restrykcyjna danego genomu polega na zaznaczeniu miejsc cięcia (czyli *trawienia*) tego genomu jednym lub kilkoma enzymami restrykcyjnymi. Miejsca te są nanoszone, stosując jako skalę liczbę par zasad. Są one charakterystycznymi punktami genomu, pozwalającymi umiejscowić geny (po ich zsekwencjonowaniu).

Rys.: mapa restrykcyjna plazmidu *E. coli* (pBR322) długości 4363 bp oraz mapa restrykcyjna λ faga, długości 48,502 bp.

Przykład 6.1.1 Rozważmy abstrakcyjny przykład trawienia pewnego genomu dwoma enzymami restrykcyjnymi A i B. Dowolnie numerujemy pocięte kawałki.



Miejsca trawienia enzymem A, enzymem B oraz enzymami A i B.

□

To co daje się zaobserwować eksperymentalnie, to długości pociętych kawałków (mierzone przy pomocy elekroforezy na żelu). Nie daje się natomiast ustalić kolejności tych kawałków. W pewnych sytuacjach (gdy odcinki w $A \wedge B$ są różnej długości) można stwierdzić, który odcinek z A (lub z B) zawiera w sobie dany odcinek z $A \wedge B$. Tzn. mamy macierz $I(A, A \wedge B)$ zdefiniowaną następująco. Niech $A_1, \ldots, A_m$ będzie numeracją odcinków powstałych z trawienia genomu enzymem A oraz niech $C_1, \ldots, C_k$ będzie numeracją odcinków powstałych z trawienia genomu enzymami A oraz B. Wówczas element macierzy $I(A, A \wedge B)$ o współrzędnych (i, j) wynosi 1, gdy A_i oraz C_j mają niepuste przecięcie; w przeciwnym przypadku element ten wynosi 0. Podobnie definiujemy macierz $I(B, A \wedge B)$ (pewna numeracja $B_1, \ldots, B_n$ oraz ta sama numeracja $C_1, \ldots, C_k$). Tak więc możemy przyjąć, że $I(A, A \wedge B)$ oraz $I(B, A \wedge B)$ są znane. Poszukujemy macierzy $I(A, B)$ (miejsca niepustych przecięć $A_1, \ldots, A_m$ oraz $B_1, \ldots, B_n$).

Przykład 6.1.2 Dla miejsc trawienia z Przykładu 6.1.1 mamy macierze.

$$I(A,B) = \begin{pmatrix} 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 \end{pmatrix}$$

$$I(A, A \wedge B) = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \end{pmatrix}$$

$$I(B, A \wedge B) = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}$$

□

Lemat 6.1.1

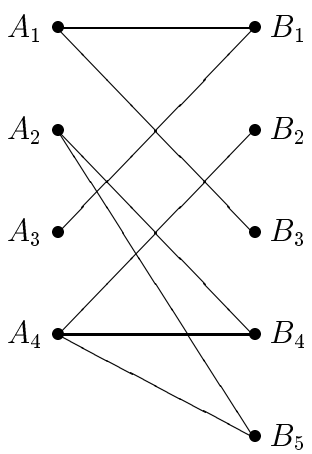
$$I(A, B) = I(A, A \wedge B) \cdot I(B, A \wedge B)^T,$$

gdzie I^T oznacza macierz transponowaną, tzn. $I^T(i, j) = I(j, i)$.

Dowód: $A_i \cap B_j \neq \emptyset$ zachodzi wtedy i tylko wtedy, gdy istnieje k takie, że $A_i \cap C_k \neq \emptyset$ oraz $B_j \cap C_k \neq \emptyset$ (takie k jest co najwyżej jedno). ■

Wygodnie jest przedstawić $I(A, B)$ jako *graf dwudzielny* $G(A, B)$. Wierzchołkami tego grafu są $A_1 \ldots, A_m$ (jeden poziom); drugi poziom tworzą wierzchołki $B_1, \ldots, B_n$. A_i oraz B_j łączy krawędź, gdy $A_i \cap B_j \neq \emptyset$. Oczywiście istnieje wzajemnie jednoznaczna odpowiedniość pomiędzy macierzami zero-jedynkowymi a grafami dwudzielnymi.

Przykład 6.1.3 Dla naszego przykładu mamy następujący graf.



□

Powiemy, że graf G jest *grafem przedziałowym*, gdy jest on izomorficzny z grafem, którego wierzchołkami są otwarte przedziały na prostej $\mathbb{R}$ oraz krawędź łączy dwa przedziały I oraz J, gdy $I \cap J \neq \emptyset$.

Macierz zero-jedynkową nazwiemy *schodkową*, gdy w każdym wierszu wszystkie jedynki stoją obok siebie oraz dla każdych dwóch sąsiednich wierszy, jeśli $i_1 < j_1$ są pozycjami lewego oraz prawego końca górnego wiersza i $i_2 < j_2$ są pozycjami końców niższego wiersza, to $i_1 \leq i_2$ oraz $j_1 \leq j_2$.

Twierdzenie 6.1.2 *Natępujące warunki są równoważne:*
(i) *Graf dwudzielny $G(A, B)$ jest grafem otrzymanym z pewnej mapy restrykcyjnej.*
(ii) *Graf $G(A, B)$ jest dwudzielnym przedziałowym grafem bez izolowanych wierzchołków.*
(iii) *$I(A, B)$ może być przekształcona poprzez permutacje wierszy i kolumn do macierzy schodkowej, nie zawierającej podmacierzy*

$$\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$$

Dowód: Równoważność (i) i (ii) jest oczywista. Permutacje wierszy i kolumn macierzy odpowiadają porządkowi odcinków, a schodkowa struktura odpowiada długościom i zazębieniom odcinków. ■

Przykład 6.1.4 Paermutacja macierzy (wierszy i kolumn) z Przykładu 6.1.2 wygląda następująco. Permutacje wierszy i kolumn są zaznaczone na brzegu macierzy.

	4	5	2	3	1
2	1	1	0	0	0
4	0	1	1	1	0
1	0	0	0	1	1
3	0	0	0	0	1

□

6.1.1 Algorytm budowania mapy restrykcyjnej

Będziemy przyjmowali, że mamy dany graf dwudzielny. Celem będzie sprawdzenie czy graf ten pochodzi od pewnej mapy restrykcyjnej i ewentualnie

rekonstrukcja tej mapy. Oczywiście map mających ten sam graf dwudzielny może być dużo. Bez zmniejesznia ogólności możemy przyjąć, że graf jest spójny — w przeciwnym przypadku rozważamy każdą spójną składową oddzielnie.[1]

Niech $\deg(v)$ oznacza stopień wierzchołka v. Niech $G = (V, E)$ oraz

$$V_2 = \{v \mid \deg(v) \geq 2\},$$

$$V_1 = \{v \mid \deg(v) = 1\}.$$

Wówczas $V = V_1 \cup V_2$

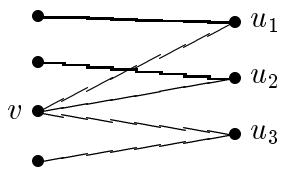
Koroną w grafie G nazwiemy ciąg wierzchołków $v_1, \ldots, v_n, u_1, \ldots, u_n \in V$ taki, że dla $i = 1, \ldots, n$ mamy $(v_i, u_i) \in E$ oraz $(u_i, v_{i+1}) \in E$, przy czym przyjmujemy, że $v_{n+1} = v_1$.

Lemat 6.1.3 *Jeśli G pochodzi od mapy restrykcyjnej, to*
(i) *Dla każdego $v \in V_2$*

$$|\{u \in V_2 \mid (u, v) \in E\}| \leq 2.$$

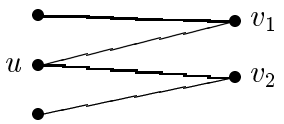
(ii) *G nie zawiera korony.*

Dowód: (i) Jeśli mamy taką sytuację



to odcinek v musi przecinać odcinki u_1, u_2, u_3 (parami rozłączne). Zatem jeden z nich musi być całkowicie zawarty w v, czyli nie może on przecinać innego odcinka rozłącznego z v.

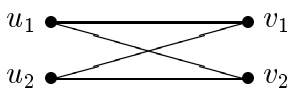
Dla dowodu (ii) zauważmy, że jeśli mamy sytuację



[1] Niespójność grafu pojawia się, gdy A i B tną w tym samym miejscu. Wówczas liczba składowych jest równa 1 plus liczba wspólnych miejsc cięcia.

to v_1 i v_2 muszą sąsiadować ze sobą (z dokładnością do ewentualnych odcinków całkowicie zawartych w u, oddzielających v_1 i v_2 od siebie). Dla $n > 2$ mamy więc sytuację taką, że odcinki $v_1, \ldots, v_n, v_1$ muszą ze sobą sąsiadować, co jest niemożliwe.

Dla $n = 2$ mamy



Czyli u_1, u_2 sąsiadują ze sobą, v_1, v_2 sąsiadują ze sobą oraz każde u_i przecina każde u_j. To jest niemożliwe. ■

Mapy restrykcyjne będziemy przedstawiać w postaci ciągu zbiorów dwuelementowych $\{u, v\}$, gdzie wierzchołki u oraz v są różnych poziomów grafu. Umieszczenie u oraz v w jednym zbiorze oznacza, że odcinki te przecinają się. Kolejność $\{u, v\}, \{u', v'\}$ oznacza, że odcinek u znajduje się na lewo od odcinka u' i podobnie odcinek v znajduje się na lewo od odcinka v'.[2]

Kroki algorytmu

1. *Znajdowanie końca:* znajdź $v \in V_2$ takie, że

$$|\{u \in V_2 \mid (u, v) \in E\}| = 1. \tag{6.6.1}$$

Jeśli $V_2 = \emptyset$, to G jest jednowierzchołkowy (wówczas G nie pochodzi od mapy restrykcyjnej) lub G zawiera dokładnie dwa wierzchołki u, v i są one połączone krawędzią. W tym przypadku poszukiwaną mapą jest $\{u, v\}$, czyli żaden z enzymów nie przeciął DNA.

Jeśli natomiast $|V_2| \geq 2$, to v spełniające (6.6.1) musi istnieć, gdyż w przeciwnym razie G musiałby zawierać koronę, więc na mocy Lematu 6.1.3(ii) G nie pochodziłby od żadnej mapy restrykcyjnej. Załóżmy więc, że takie v zostało znalezione.

2. *Zygzak V_2:* zaczynając od v znaj"dź wierzchołki $v_1, \ldots, v_k$ w zbiorze V_2 takie, że $v_1 = v$ oraz $(v_i, v_{i+1}) \in E$ dla $i = 1, \ldots, k-1$, o tej własności, że droga $v_1, \ldots, v_k$ jest maksymalnej długości. Z lematu 6.1.3(i) wynika, że droga ta jest wyznaczona jednoznacznie oraz v_k jest też końcem, tzn. $|\{u \in V_2 \mid (u, v_k) \in E\}| = 1$. Co więcej, musi być tak, że

$$V_2 = \{v_1, \ldots, v_k\},$$

[2]Zakładamy tu, że u, u' pochodzą z jednego poziomu grafu oraz v, v' pochodzą z drugiego poziomu.

bo inaczej G byłby niespójny. Tworzymy listę odcinków (jako częściową mapę):

$$\{v_1, v_2\}, \{v_2, v_3\}, \ldots, \{v_{k-1}, v_k\}.$$

3. *Dołączenie V_1:* Dla każdego $u \in V_1$ niech $v_i \in V_2$ będzie jedynym wierzchołkiem takim, że $(u, v_i) \in E$. Jeśli v_i jest jednym z końców, to dołączamy $\{u, v_i\}$ do końca listy (z lewej lub prawej, w zależności od tego, którym końcem jest v_i). Jeśli natomiast v_i nie jest końcem, to $\{u, v_i\}$ wstawiamy do listy pomiędzy $\{v_{i-1}, v_i\}$ oraz $\{v_i, v_{i+1}\}$. Po wyczerpaniu wszystkich wierzchołków z V_1 otrzymana lista par odcinków tworzy mapę, dla której G jest grafem dwudzielnym.

Przykład 6.1.5 Zilustrujemy powyższy algorytm na grafie G z Przykładu 6.1.3. Mamy

$$V_2 = \{A_1, A_2, A_4, B_1, B_3, B_5\},$$

$$V_1 = \{A_3, B_2, B_4\}.$$

B_1 jest końcem znalezionym w kroku 1. Zygzakiem znalezionym w kroku 2 algorytmu jest

$$B_1, A_1, B_3, A_4, B_5, A_2.$$

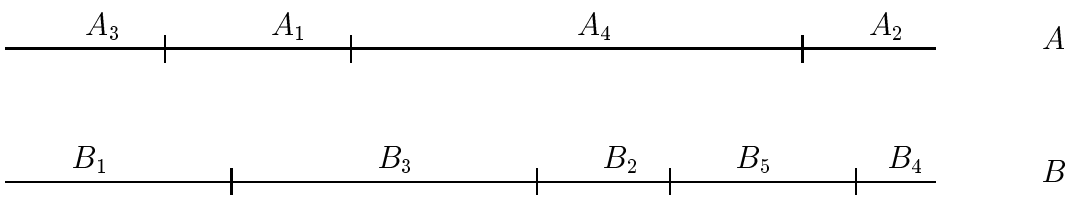
Tak więc na tym etapie mamy listę

$$\{B_1, A_1\}, \{A_1, B_3\}, \{B_3, A_4\}, \{A_4, B_5\}, \{B_5, A_2\}.$$

Po dołączeniu V_1 (krok 3) mamy listę

$$\{A_3, B_1\}, \{B_1, A_1\}, \{A_1, B_3\}, \{B_3, A_4\}, \{B_2, A_4\}, \{A_4, B_5\}, \{B_5, A_2\}, \{A_2, B_4\}.$$

To daje następującą mapę



Jest to odbicie symetryczne mapy z Przykładu 6.1.1. □

Udowodniliśmy przy okazji następujące twierdzenie.

Twierdzenie 6.1.4 *Spójny dwudzielny graf G pochodzi od mapy restrykcyjnej wtedy i tylko wtedy, gdy G ma więcej niż jeden wierzchołek oraz spełnia warunki (i), (ii) Lematu 6.1.3.*