

Bioinformatyka

wykład 7: mapowanie

prof. dr hab. inż. **Marta Kasprzak**
Instytut Informatyki, Politechnika Poznańska

Mapowanie

Pojęcie „mapowanie” odnosi się do różnych problemów

- Trzeci etap składania sekwencji genomowej *de novo*
- Przypisywanie genów do odpowiednich miejsc chromosomów na podstawie efektów rekombinacji
- Dopasowywanie odczytów z sekwenatora do genomu referencyjnego, tzw. resekwencjonowanie
- Porównywanie wyników eksperymentalnych dla fragmentów dwóch genomów, tzw. *fingerprinting*
- Dopasowywanie struktur przestrzennych białek i RNA
- Inne

Mapowanie

- W trzecim (chronologicznie pierwszym) etapie składania sekwencji genomowej należy odtworzyć oryginalne uporządkowanie długich, nakładających się na siebie fragmentów DNA bez znajomości ich sekwencji. Sekwencje nukleotydowe fragmentów rozpoznawane są niezależnie w procesie sekwencjonowania i asemblacji
- Dane wejściowe stanowią multizbiory długości fragmentów (w mapowaniu restrykcyjnym, np. w problemach pojedynczego lub podwójnego trawienia) lub wyniki eksperymentu hybrydacyjnego na zbiorach fragmentów i unikalnych próbek (markerów)

3

Mapowanie przez hybrydyzację

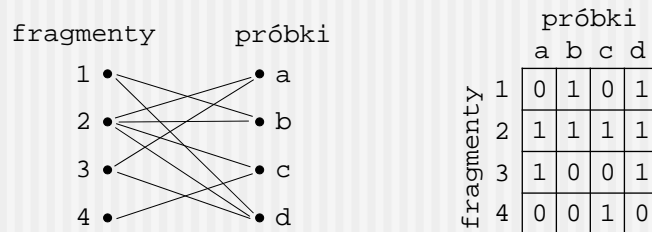
[S. Benzer, *Proc. Natl. Acad. Sci. USA* 45 (1959)]

- Metoda służąca identyfikacji topologii zapisu informacji genetycznej bakteriofaga. Wynik eksperymentu wskazał na jej linearną strukturę
- Podobna zasada porządkowania fragmentów stosowana jest obecnie w mapowaniu przez hybrydyzację za pomocą unikalnych próbek. Bezbłędna macierz hybrydyzacji może być reprezentowana poprzez graf interwałowy
- Grafy interwałowe odzwierciedlają relację nakładania się interwałów w przestrzeni liniowej. Dwa wierzchołki odpowiadające interwałom są połączone krawędzią, jeśli interwały przecinają się

4

Mapowanie przez hybrydyzację

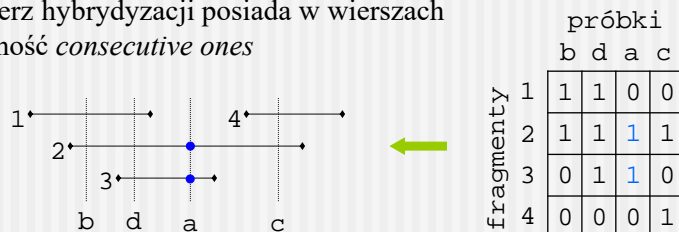
- Mapowanie za pomocą unikalnych próbek (bez błędów)
- W wyniku eksperymentu hybrydyzacji dłuższych fragmentów DNA o nieznanych sekwencjach z krótkimi unikalnymi próbkami otrzymujemy informację o ich zawieraniu



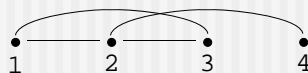
5

Mapowanie przez hybrydyzację

- W przypadku eksperymentu bez błędów macierz hybrydyzacji posiada w wierszach własność *consecutive ones*



- Graf skonstruowany na podstawie bezbłędnej macierzy hybrydyzacji jest grafem interwałowym



6

Mapowanie przez hybrydyzację

[F. Alizadeh i in., *J. Comput. Biol.* 2 (1995)]

- Błędy hybrydyzacji: chimery, błędy negatywne i pozytywne
- Macierz hybrydyzacji z błędami nie posiada już własności *consecutive ones*. Celem jest znalezienie takiego uporządkowania kolumn, dla którego liczba błędów w macierzy (czyli liczba przerw w ciągach jedynek) jest minimalna

		próbki				
		a	b	c	d	
fragmenty	1	0	1	0	1	macierz z błędami
	2	0	1	1	1	
	3	1	0	0	1	
	4	1	0	1	0	

		próbki				
		b	d	a	c	
fragmenty	1	1	1	0	0	rozwiązanie optymalne
	2	1	1	0	1	
	3	0	1	1	0	
	4	0	0	1	1	

7

Mapowanie przez hybrydyzację

[F. Alizadeh i in., *J. Comput. Biol.* 2 (1995)] – cd.

- Poszukiwanie minimalnej liczby przerw w macierzy odpowiada poszukiwaniu cyklu komiwojażera w grafie (metoda heurystyczna)
- Wagi krawędzi to odległość Hamminga pomiędzy próbkami

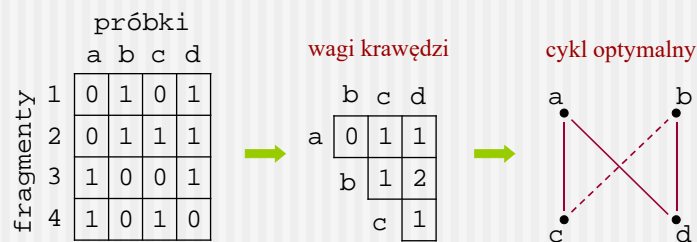


8

Mapowanie przez hybrydyzację

[D. Greenberg i S. Istrail, *J. Comput. Biol.* 2 (1995)]

- Nagradzanie zgodności zamiast karania niezgodności
- Wierzchołki reprezentują próbki, a krawędzie opisane są liczbą fragmentów, które zawierają obie próbki. Rozwiązaniem jest cykl Hamiltona o najwyższej wadze



9

Mapowanie przez hybrydyzację

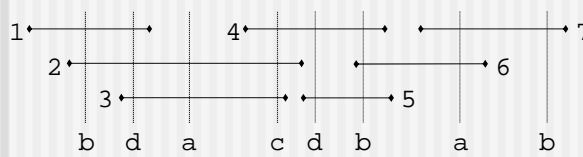
- Jeśli próbki nie są unikalne w obrębie badanego fragmentu genomu, nie można oczekiwać własności *consecutive ones* macierzy hybrydyzacji. W macierzy nadal jednej próbce odpowiada jedna kolumna, tym razem jednak stowarzyszone z nią fragmenty nie muszą wystąpić obok siebie w genomie
- Zestaw próbek hybrydujących z fragmentem tworzy jego profil (ang. *fingerprint*). Uszeregowanie fragmentów można odtworzyć, opierając się na przypuszczeniu, że im większe jest ich wzajemne nałożenie, tym bardziej podobne są ich profile (tym większą liczbę próbek mają wspólną)

10

Mapowanie przez hybrydyzację

[D. Gusfield, *Algorithms on Strings, Trees, and Sequences* (1997)]

- Problem mapowania przez hybrydyzację z nieunikalnymi próbkami jest trudny obliczeniowo, rozwiązany został m.in. heurystyką zachłanną



	a	b	c	d
1	0	1	0	1
2	1	1	1	1
3	1	0	1	1
4	0	1	1	1
5	0	1	0	1
6	1	1	0	0
7	1	1	0	0

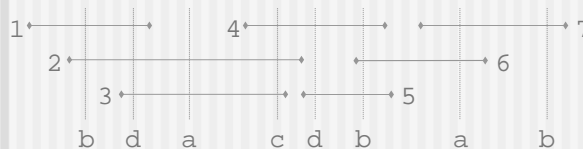
11

Mapowanie przez hybrydyzację

[D. Gusfield, *Algorithms on Strings, Trees, and Sequences* (1997)]

- W kolejnych krokach dobierane są w pary i łączone w kontig dwa fragmenty o najbardziej pasujących profilach

**1+5, 6+7, (1+5)+4, (1+5+4)+2,
(1+5+4+2)+3, (6+7)+(1+5+4+2+3)**



	a	b	c	d
1	0	1	0	1
2	1	1	1	1
3	1	0	1	1
4	0	1	1	1
5	0	1	0	1
6	1	1	0	0
7	1	1	0	0

12

Mapowanie restrykcyjne

- Innym rodzajem mapowania jest podejście bez markerów, z użyciem tylko enzymów restrykcyjnych (np. jednego lub dwóch), które trawią badany fragment dwuniciowego DNA. Multizbiory długości odcinków wynikowych wystarczają do odtworzenia oryginalnego uporządkowania odcinków
- Wśród kombinatorycznych modeli problemów mapowania restrykcyjnego można wymienić problem podwójnego trawienia, problem częściowego trawienia, uproszczony problem częściowego trawienia

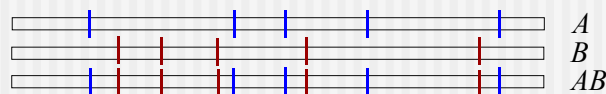
13

Mapowanie restrykcyjne

- Problem mapowania za pomocą dwóch enzymów restrykcyjnych — problem podwójnego trawienia (ang. *double digest problem*, DDP)

Instancja: Multizbiory A , B i AB długości odcinków pochodzących z trawienia kopii fragmentu DNA dwoma enzymami restrykcyjnymi.

Rozwiązanie: Mapa miejsc restrykcyjnych w badanym fragmencie, tzn. takie uszeregowanie elementów multizbioru AB , które umożliwi pokrycie wskazanych miejsc cięcia poprzez uszeregowanie elementów multizbiorów A i B .



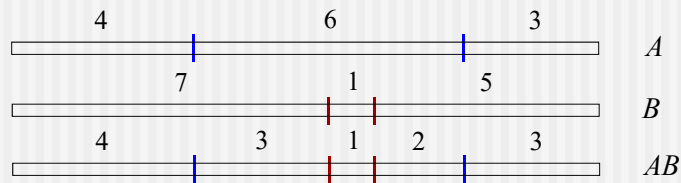
14

Mapowanie restrykcyjne

- DDP — przykład

$A = \{3, 4, 6\}$, $B = \{1, 5, 7\}$, $AB = \{1, 2, 3, 3, 4\}$

Rozwiązanie:



- Problem ten nawet przy założeniu braku błędów w instancji jest trudny obliczeniowo

15

Mapowanie restrykcyjne

[K. Danna i D. Nathans, *Proc. Natl. Acad. Sci. USA* 68 (1971)]

- Problem mapowania za pomocą jednego enzymu restrykcyjnego — problem częściowego trawienia (ang. *partial digest problem*, PDP)

Instancja: Multizbiór A długości odcinków pochodzących z trawienia kopii fragmentu DNA jednym enzymem restrykcyjnym w różnych przedziałach czasowych.

Rozwiązanie: Mapa miejsc restrykcyjnych w badanym fragmencie taka, że odległości pomiędzy wszystkimi parami miejsc (również końcami fragmentu) pokryją się z multizbiorem A .

16

Mapowanie restrykcyjne

- PDP — przykład

$A = \{1, 2, 3, 3, 3, 4, 4, 5, 6, 6, 7, 8, 9, 10, 13\}$

Rozwiązanie:



- Mapowanie metodą częściowego trawienia jest trudne w realizacji ze względu na liczbę przeprowadzanych eksperymentów laboratoryjnych i problem z doбором właściwego czasu trwania reakcji
- Problem przy założeniu braku błędów w instancji jest otwarty z punktu widzenia złożoności obliczeniowej

17

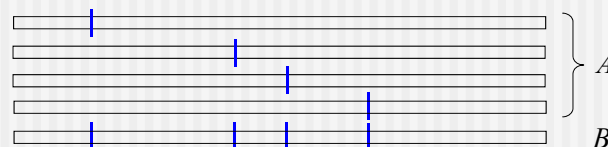
Mapowanie restrykcyjne

[J. Błazewicz i in., *Bioinformatics* 17 (2001)]

- Uproszczony problem częściowego trawienia (ang. *simplified partial digest problem*, SPDP)

Instancja: Multizbiory A i B długości odcinków pochodzących odpowiednio z krótkiego i długiego trawienia kopii fragmentu DNA jednym enzymem restrykcyjnym.

Rozwiązanie: Mapa miejsc restrykcyjnych w badanym fragmencie, tzn. takie uszeregowanie elementów multizbioru B , które umożliwi pokrycie wskazanych miejsc cięcia poprzez uszeregowanie par elementów z A .



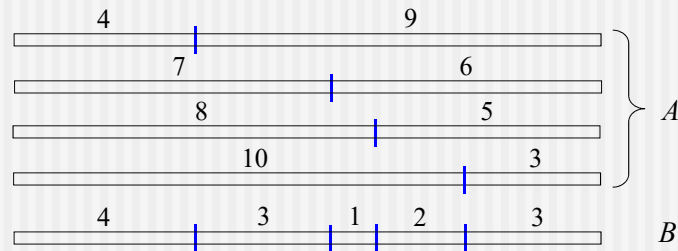
18

Mapowanie restrykcyjne

- SPDP — przykład

$A = \{3, 4, 5, 6, 7, 8, 9, 10\}$, $B = \{1, 2, 3, 3, 4\}$

Rozwiązanie:



- Problem przy założeniu braku błędów w instancji jest trudny obliczeniowo

19

Mapowanie restrykcyjne

- Mapowanie restrykcyjne może być także rozumiane jako dopasowywanie do siebie dwóch fragmentów genomu na podstawie ich profili cięć enzymami restrykcyjnymi (*fingerprinting*). Podobne fragmenty poddane trawieniu tych samych enzymów dadzą w efekcie podobny obraz długości uzyskanych odcinków
- Dopasowanie takie może być uzyskane algorytmem programowania dynamicznego w podobny sposób jak w problemie dopasowania dwóch sekwencji. Dane o trawieniu zapisuje się w postaci sekwencji naprzemiennie identyfikatorów enzymów oraz odległości odcinków od znakowanego końca

20

Mapowanie restrykcyjne

[M. Waterman i in., *Nucleic Acids Res.* 12 (1984)]

- Dwa dopasowywane profile cięć reprezentowane są przez sekwencje A i B odpowiednio n i m miejsc restrykcyjnych:
 $A = (r_0, p_0, r_1, p_1, \dots, r_{n+1}, p_{n+1})$,
 $B = (s_0, q_0, s_1, q_1, \dots, s_{m+1}, q_{m+1})$,
w których r_i oraz s_i to identyfikatory enzymów, a p_i oraz q_i to odległości odcinków od znakowanego końca, przy: $r_0=s_0=\alpha$,
 $p_0=q_0=0$, $r_{n+1}=s_{m+1}=\beta$, $p_{n+1}=N$, $q_{m+1}=M$, gdzie α i β to końce sekwencji a N i M to długości fragmentów
- Celem algorytmu jest przekształcenie jednego profilu w drugi przy najmniejszym koszcie

21

Mapowanie restrykcyjne

[M. Waterman i in., *Nucleic Acids Res.* 12 (1984)] – cd.

- W rozwiązywanym problemie dopuszczone są wstawienia bądź usunięcia miejsc restrykcyjnych, także zmiana długości odcinków, natomiast wyklucza się zmianę identyfikatorów enzymów restrykcyjnych
- Koszt wstawienia lub usunięcia miejsca związany jest ze zmienną λ , koszt zmiany długości odcinka ze zmienną μ , gdzie $\mu = w_0 + x \cdot w_1$ a x to różnica w długości wyrażona w kbp (tysiącach par zasad). Jeśli x jest małe ($x \leq \Delta$), przyjmuje się brak kary za zmianę długości ($\mu=0$)
- Przykładowe wartości z artykułu: $\Delta = 0,5$ kbp, $\lambda=1$, $w_0=w_1=0,5$

22

Mapowanie restrykcyjne

[M. Waterman i in., *Nucleic Acids Res.* 12 (1984)] – cd.

- Algorytm programowania dynamicznego wypełnia tablicę odległości D o $n+2$ wierszach i $m+2$ kolumnach

$$D(0,0) = 0,$$

jeśli $0 < i < n+1$ lub $0 < j < m+1$:

$$D(i,0) = D(i,m+1) = D(0,j) = D(n+1,j) = \infty,$$

jeśli $r_i \neq s_j$:

$$D(i,j) = \infty,$$

wpp. dla $0 < k \leq i$, $0 < l \leq j$:

$$D(i,j) = \min[D(i-k,j-l) + \lambda(k+l-2) + w_0 + w_1|p_i - p_{i-k} - q_j + q_{j-l}|].$$

- Wynik odczytywany jest w $D(n+1,m+1)$. Złożoność $O(n^2m^2)$

23

Mapowanie restrykcyjne

- Przykład (wartości p_i , q_i i Δ w kbp)

A = (α , 0, E_1 , 8.4, E_2 , 15.2, E_1 , 19.0, β , 24.0)

B = (α , 0, E_1 , 7.7, E_2 , 9.5, E_2 , 14.8, E_1 , 20.3, β , 25.0)

$n=3$, $m=4$, $\Delta=0.5$, $\lambda=1$, $w_0=w_1=0.5$

		j					
		0	1	2	3	4	5
i	0	0	∞	∞	∞	∞	∞
	1	∞	0.85	∞	∞	9.45	∞
	2	∞	∞	3.85	1.85	∞	∞
	3	∞	8.15	∞	∞	3.20	∞
	4	∞	∞	∞	∞	∞	3.20

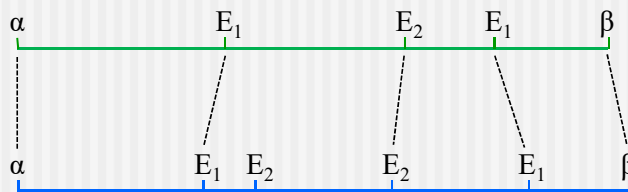
24

Mapowanie restrykcyjne

- Przykład – cd.

A = (α , 0, E₁, 8.4, E₂, 15.2, E₁, 19.0, β , 24.0)

B = (α , 0, E₁, 7.7, E₂, 9.5, E₂, 14.8, E₁, 20.3, β , 25.0)



25

Mapowanie optyczne

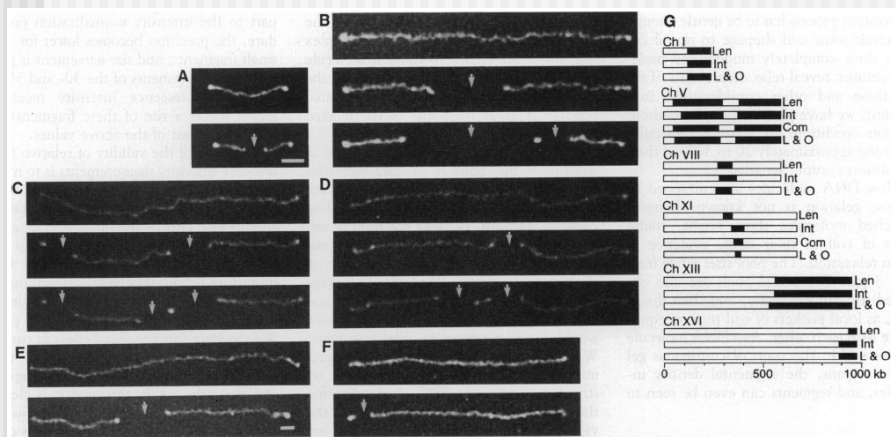
[D. Schwartz i in., *Science* 262 (1993)]

- Mapowanie optyczne z użyciem enzymu restrykcyjnego
- Badana cząsteczka jest oznaczana fluorescencyjnie i usztywniana, reakcja trawienia obserwowana jest pod mikroskopem
- Obraz rejestrowany co 20 sek. umożliwia zaobserwowanie podziału i ustalenie kolejności poszczególnych fragmentów w cząsteczce. Długości fragmentów szacowane są na drodze wizualnego porównania ze wzorcem — ocena jasności i/lub długości

26

Mapowanie optyczne

[D. Schwartz i in., *Science* 262 (1993)] – cd.



27

Mapowanie optyczne

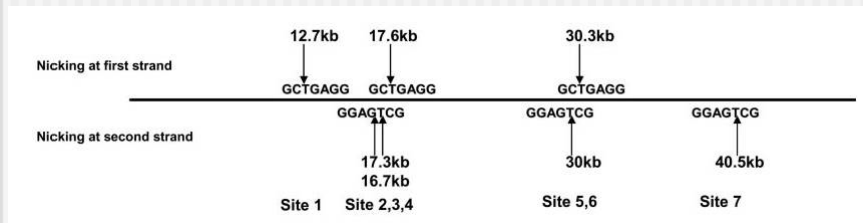
[M. Xiao i in., *Nucleic Acids Res.* 35 (2007)]

- Mapowanie optyczne w celu zidentyfikowania pozycji próbek (*fingerprinting*)
- Badana cząsteczka jest oznaczana innym barwnikiem niż znacznik, który przyłącza się do cząsteczki w miejscach wcześniej uszkodzonych odpowiadających próbce. Obraz całości analizowany jest pod mikroskopem
- Podejście może służyć także w trzecim etapie składania sekwencji genomowej. Aby uzyskać bardziej jednoznaczny wynik eksperymentu, można użyć więcej niż jeden rodzaj próbek

28

Mapowanie optyczne

[M. Xiao i in., *Nucleic Acids Res.* 35 (2007)] – cd.

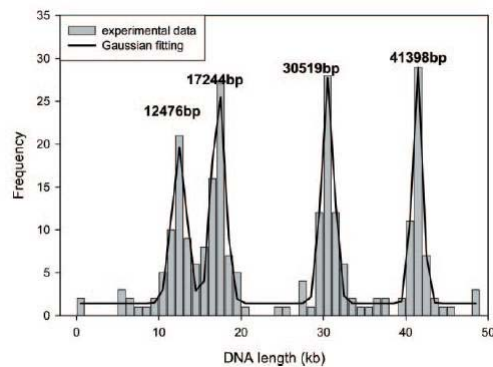
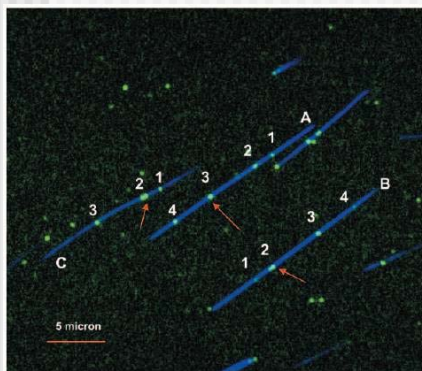


Pozycje próbek w badanej cząsteczce — stan rzeczywisty.
Nacięcia dokonywane są enzymem rozpoznającym odpowiedni fragment.

29

Mapowanie optyczne

[M. Xiao i in., *Nucleic Acids Res.* 35 (2007)] – cd.



Obraz „uliniowanego” materiału po eksperymencie (po lewej)
oraz wykres uzyskany przez nałożenie wyników dla 81 molekuł (po prawej).

30