Algorytmy kombinatoryczne w bioinformatyce

wykład 6: 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 hybrydyzacyjnego na zbiorach fragmentów i unikalnych próbek (markerów)

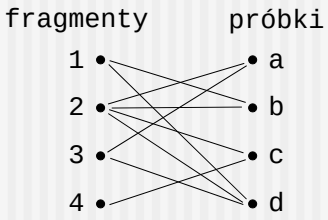
Mapowanie przez hybrydyzację

[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

Mapowanie przez hybrydyzację

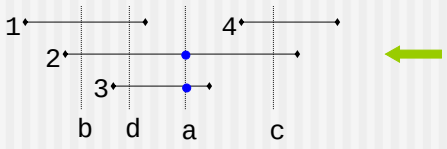
- Mapowanie za pomocą unikalnych próbek (bez błędów)
- W wyniku eksperymentu hybrydyzacji dłuższych fragmentów DNA o nieznanych sekwencjach z krótkimi unikalnymi próbkami otrzymujemy informację o ich zawieraniu



fragmenty \ próbki	a	b	c	d
1	0	1	0	1
2	1	1	1	1
3	1	0	0	1
4	0	0	1	0

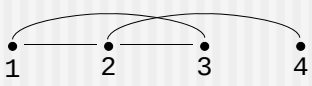
Mapowanie przez hybrydyzację

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



fragmenty \ próbki	b	d	a	c
1	1	1	0	0
2	1	1	1	1
3	0	1	1	0
4	0	0	0	1

- Graf skonstruowany na podstawie bezbłędnej macierzy hybrydyzacji jest grafem interwałowym



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

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

macierz z błędami

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

rozwiązanie optymalne

Mapowanie przez hybrydyzację

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

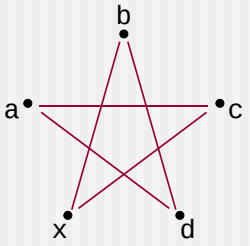
- Poszukiwanie minimalnej liczby przerw w macierzy odpowiada poszukiwaniu cyklu komiwojażera w grafie (metoda heurystyczna)
- Wagi krawędzi to odległość Hamminga pomiędzy próbkami

fragmenty	próbki a	b	c	d	x
1	0	1	0	1	0
2	0	1	1	1	0
3	1	0	0	1	0
4	1	0	1	0	0

odległość Hamminga

	b	c	d	x
a	4	2	3	2
b		2	1	2
c			3	2
d				3

cykl optymalny (1 z 4)



Mapowanie przez hybrydyzację

- Zadanie

Stosując metodę Alizadeha i in. należy tak uporządkować kolumny poniższej macierzy, aby liczba błędów (przerw w ciągach jedynek) była minimalna.

fragmenty	próbki a	b	c	d	e	f	g
1	0	1	1	0	0	1	1
2	1	1	0	1	1	1	1
3	0	1	1	1	0	0	0
4	1	1	1	1	1	0	0

W tym przypadku uzyskane rozwiązanie jest jednym z możliwych rozwiązań optymalnych dla tej instancji.

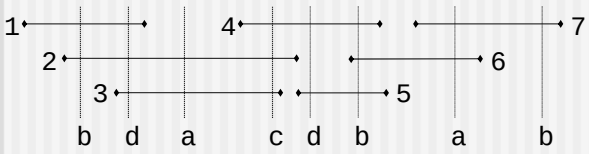
Mapowanie przez hybrydyzację

- 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 hybrydyzujących 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)

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



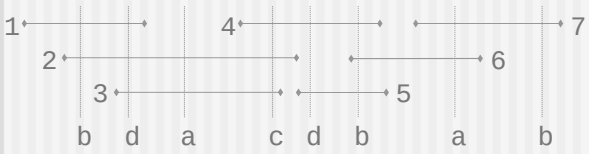
	a	b	c	d
1	0	1	0	1
2	1	1	1	1
3	1	0	1	1
4	0	1	1	1
5	0	1	0	1
6	1	1	0	0
7	1	1	0	0

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

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

Mapowanie restrykcyjne

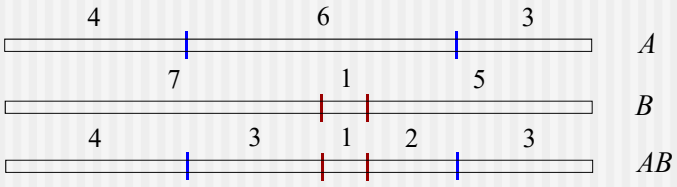
- Problem mapowania za pomocą dwóch enzymów restrykcyjnych — problem podwójnego trawienia (ang. *double digest problem*, DDP)

  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*.

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

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.

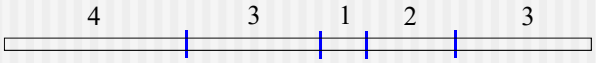
  Rozwiązanie: Mapa miejsc restrykcyjnych w badanym fragmencie taka, że odległości pomiędzy wszystkimi parami miejsc (również końcami fragmentu) pokryją się z multizbiorem A.

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 doborem właściwego czasu trwania reakcji
- Problem przy założeniu braku błędów w instancji jest otwarty z punktu widzenia złożoności obliczeniowej

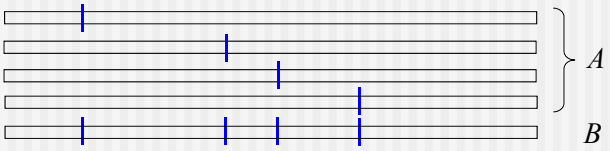
Mapowanie restrykcyjne

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

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

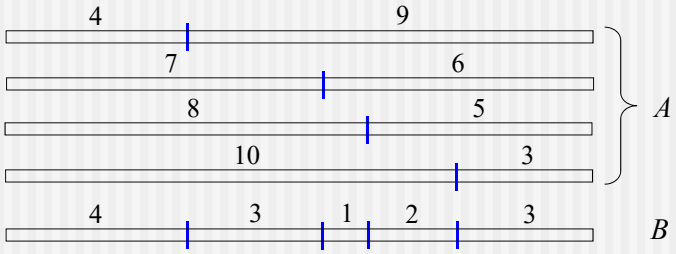
  Instancja: Multizbiory A i B długości odcinków pochodzących odpowiednio z krótkiego i długiego trawienia kopii fragmentu DNA jednym enzymem restrykcyjnym.

  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.



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

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

Mapowanie restrykcyjne

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

- Dwa dopasowywane profile cięć reprezentowane są przez sekwencje A i B odpowiednio n i m miejsc restrykcyjnych:
  $A = (r_0, p_0, r_1, p_1, \ldots, r_{n+1}, p_{n+1})$,
  $B = (s_0, q_0, s_1, q_1, \ldots, s_{m+1}, q_{m+1})$,
  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

Mapowanie restrykcyjne

[M. Waterman i in., *Nucleic Acids Research* 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, $\lambda=1$, $w_0=w_1=0{,}5$

Mapowanie restrykcyjne

[M. Waterman i in., *Nucleic Acids Research* 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 $O(n^2m^2)$

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, Δ=0.5, λ=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

Mapowanie restrykcyjne

- Przykład — cd.

  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)