

Metody bioinformatyki (MBI)

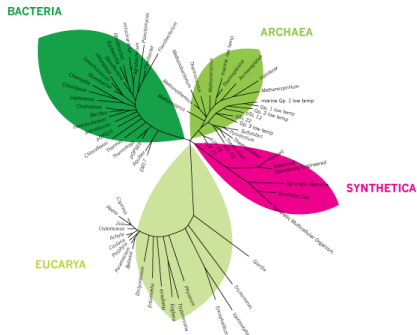
Wykład 13 - Struktury drugorzędowe, markery genetyczne,
haplotypy.

Robert Nowak

2024L

Biologia syntetyczna

Biologia syntetyczna – dziedzina inżynierii, projektowanie i realizacja sztucznych systemów biologicznych.



- ▶ chemiczna synteza DNA o zadanej sekwencji, 1970 r.
- ▶ synteza sztucznego genu, 1972 r.¹
- ▶ synteza sztucznego genomu
 - ▶ wirus, 7500 bp, 2002 r.²
 - ▶ bakteria, 1 Mbp, 2010 r.³

¹ Khora, Total synthesis of the structural gene for an alanine transfer ribonucleic acid from yeast

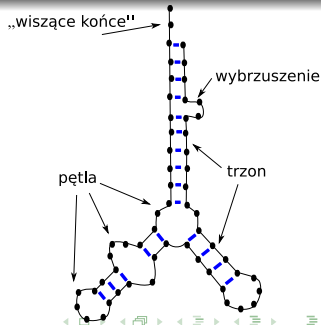
² Cello ..., Chemical synthesis of poliovirus cDNA: generation of infectious virus in the absence of natural template

Struktura drugorzędowa

Reprezentacje biopolimerów:

- ▶ struktura pierwszorzędowa - sekwencja symboli
- ▶ struktura drugorzędowa - uwzględnienie oddziaływania nukleotydów (DNA, RNA) lub aminokwasów (białka)
- ▶ struktura trzeciorzędowa - struktura atomów w przestrzeni 3D

- ▶ trzon (stem)
- ▶ pętle:
 - ▶ wybrzuszenie (bulge loop)
 - ▶ pętla wewnętrzna (interior loop)
 - ▶ pętla zewnętrzna (hairpin loop)
 - ▶ pętla wieloramienna (multi-branched loop)
- ▶ wiszący koniec

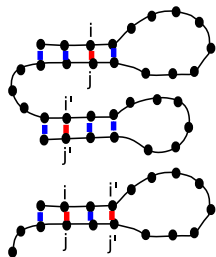


Struktura drugorzędowa - notacje

Struktura drugorzędowa: zbiór par (i, j) , takich że:

- ▶ $S = x_0x_2\dots x_{n-1}$ to reprezentacja pierwszorzędowa
- ▶ $0 \leq i < j < n$ (kolejność indeksów w parze),
- ▶ $j - i > 3$ (pętle zewnętrzne nie mogą być krótsze niż 4 symbole)
- ▶ jeżeli (i, j) oraz (i', j') są dwoma parami zasad to:

- ▶ $i < j < i' < j'$ (i, j) poprzedza (i', j')
- ▶ $i' < j' < i < j$ (i', j') poprzedza (i, j)
- ▶ $i < i' < j' < j$ (i, j) obejmuje (i', j')
- ▶ $i' < i < j < j'$ (i', j') obejmuje (i, j)



Algorytmy oparte o minimalizację energii swobodnej

- ▶ cząsteczka przyjmuje strukturę o najniższej energii
- ▶ energia w zależności od siły wiązania dla poszczególnych par
- ▶ energia dla struktury: suma dla poszczególnych par

$$E(S) = \sum_{i,j \in S} e(x_i, x_j)$$

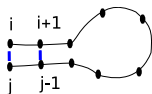
Przykładowa macierz energii:

$e(i,j)$	A	C	G	U
A	0	0	0	2
C	0	0	3	0
G	0	3	0	1
U	2	0	1	0

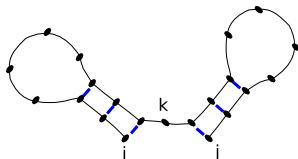
Algorytm Nussinov

- ▶ inicjacja - pętle zewnętrzne nie mogą być krótsze niż 4 nukleotydy

- ▶ tworzenie pary

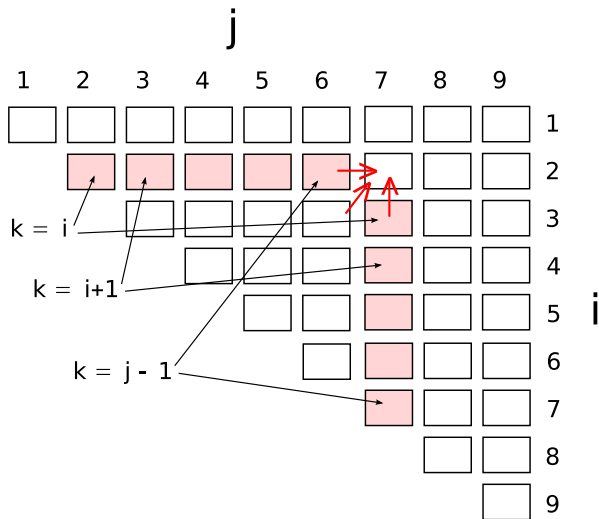


- ▶ podział sekwencji



$$F(i, j) = \max \begin{cases} 0 & j - i \leq 3 \\ F(i + 1, j - 1) + e(x_i, x_j) & \text{połączenie} \\ F(i + 1, j) & x_i \text{ bez pary} \\ F(i, j - 1) & x_j \text{ bez pary} \\ \max_{k: i \leq k < j} F(i, k) + F(k + 1, j) & \text{podział} \end{cases}$$

Algorytm Nussinov - przykład



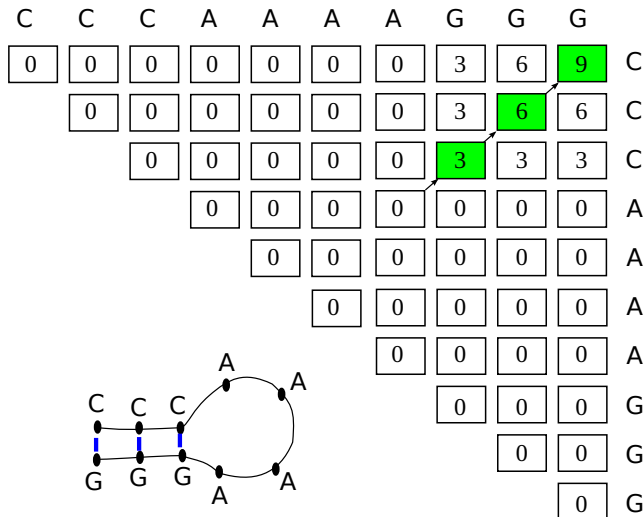
Algorytm Nussinov - przykład

C	C	C	A	A	A	A	G	G	G	
0	0	0								C
	0	0	0							C
		0	0	0						C
			0	0	0					A
				0	0	0				A
					0	0	0			A
						0	0	0		A
							0	0	0	G
								0	0	G
									0	G

Algorytm Nussinov - przykład

C	C	C	A	A	A	A	G	G	G	
0	0	0	0	0	0	0	3	6	9	C
	0	0	0	0	0	0	3	6	6	C
		0	0	0	0	0	3	3	3	C
			0	0	0	0	0	0	0	A
				0	0	0	0	0	0	A
					0	0	0	0	0	A
						0	0	0	0	A
							0	0	0	G
								0	0	G
									0	G

Algorytm Nussinov - przykład



Algorytm Nussinov - przykład (2)

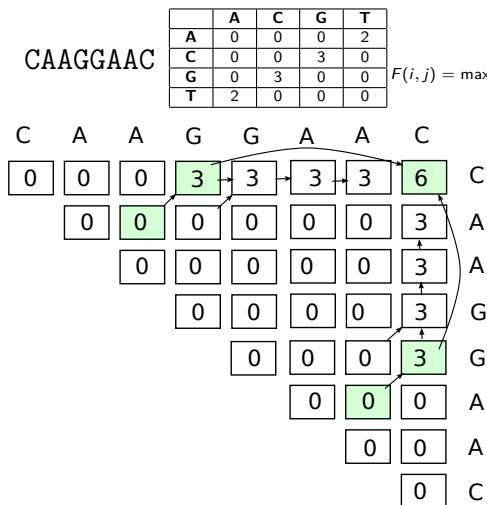
CAAGGAAC

	A	C	G	T
A	0	0	0	2
C	0	0	3	0
G	0	3	0	0
T	2	0	0	0

$$F(i, j) = \max \begin{cases} 0 & j - i \leq 2 \\ F(i + 1, j - 1) + e(x_i, x_j) \\ F(i + 1, j) \\ F(i, j - 1) \\ \max_{k:i \leq k < j} F(i, k) + F(k + 1, j) \end{cases}$$

C	A	A	G	G	A	A	C	
1	2	3	4	5	6	7	8	
0	0	0						1 C
	0	0	0					2 A
		0	0					3 A
			0	0				4 G
				0	0	0		5 G
					0	0	0	6 A
						0	0	7 A
							0	8 C

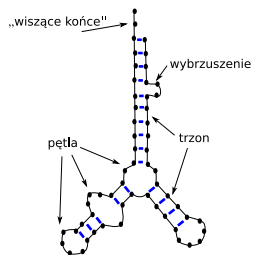
Algorytm Nussinov - przykład (2)



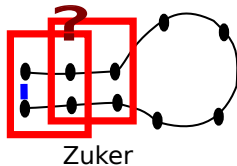
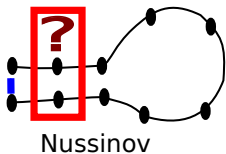
$$j - i \leq 2$$

Algorytm Zukera

- ▶ elementy stabilizujące
 - ▶ trzonki
- ▶ elementy destabilizujące:
 - ▶ pętle zewnętrzne (spinki)
 - ▶ pętle wewnętrzne (w tym wyrzuszenia)
 - ▶ pętle wieloramienne



analizie podlegają pary par nukleotydów (a nie pojedyncze nukleotydy)



Algorytm Zukera (2)

- ▶ wykorzystuje programowanie dynamiczne
- ▶ dostarczane są energie (wynikające z pomiarów) dla:
 - ▶ szpilek e_H
 - ▶ połączonych par e_S
 - ▶ pętli e_L
- ▶ oblicza macierz V - elementy (i, j) tworzą parę, oraz W - element (i, j) jest częścią pętli wieloramiennej

$$W(i, j) = \max \begin{cases} W(i + 1, j) \\ W(i, j - 1) \\ \max_{k: i \leq k < j} W(i, k) + W(k + 1, j) \\ V(i, j) \end{cases}$$

Algorytm Zukera (3)

$$V(i, j) = \max \left\{ \begin{array}{l} e_H(i, j) \\ \text{ostatnie wiązanie przed szpilką} \\ \\ e_S(i, i+1, j, j-1) + V(i+1, j-1) \\ \text{trzonek (połączenie)} \\ \\ \max_{i < i' < j' < j: i' - i + j' - j > 2} e_L(i, i', j, j') + V(i', j') \\ \text{pętla wewnętrzna} \end{array} \right.$$

Algorytmy minimalizacji energii swobodnej - podsumowanie

- ▶ Algorytm Nussinov
 - ▶ bada oddziaływania między pojedynczymi nukleotydami
 - ▶ uwzględnia ograniczenia dla 'szpilek'
- ▶ Algorytm Zukera
 - ▶ bada oddziaływania między dwiema parami nukleotydów
 - ▶ inaczej obliczana energia dla pętli, trzonka, itd.

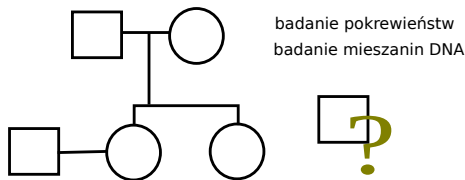
Problemy

- ▶ duża średnia liczba sub-optimalnych struktur, nie pozwala uwzględniać wszystkich możliwości
- ▶ nie uwzględniane oddziaływania trzeciorzędowe (pseudo-węzły, itp.)
- ▶ łańcuchy RNA niekoniecznie muszą przyjmować strukturę o minimalnej energii, konformacja może być wymuszona kinetyką tworzenia struktury drugorzędowej

Analiza markerów genetycznych

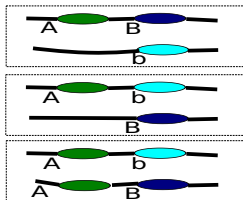
Genetyka populacyjna

- ▶ badanie różnic genetycznych pomiędzy osobnikami
- ▶ badanie chorób genetycznych
- ▶ kryminalistyka i medycyna sądowa



badanie chorób genetycznych

ABb ?



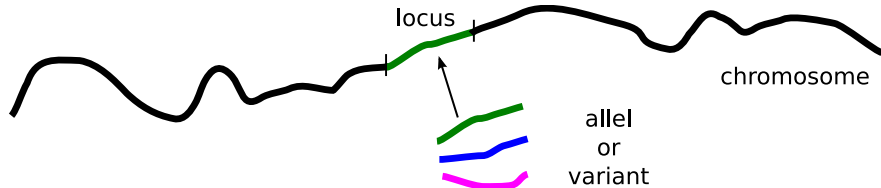
Definicje: chromosomy

- ▶ materiał genetyczny : DNA jądrowe - chromosomy, DNA mitochondrialny (mtDNA)
- ▶ liczba - zależna od organizmu (od 1 do 200)
- ▶ człowiek - w każdej komórce 46 chromosomów : 23 od matki, 23 od ojca
- ▶ długości od 50×10^6 do 260×10^6
- ▶ 22 pary autosomów
- ▶ 1 para chromosomów płciowych (kobiety XX, mężczyźni XY)



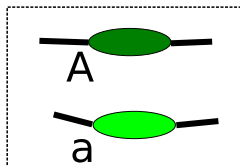
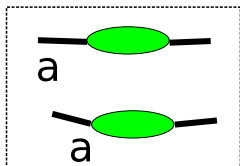
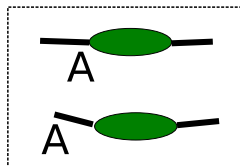
Definicje : locus, allel, wariant genu

- ▶ locus - miejsce na chromosomie
- ▶ allele - wariant genu
- ▶ haplotyp - zbiór wariantów (alleli), które są przekazywane razem



Definicje: homozygota, heterozygota

- ▶ homozygota: te same warianty genu
- ▶ heterozygota: różne warianty genu

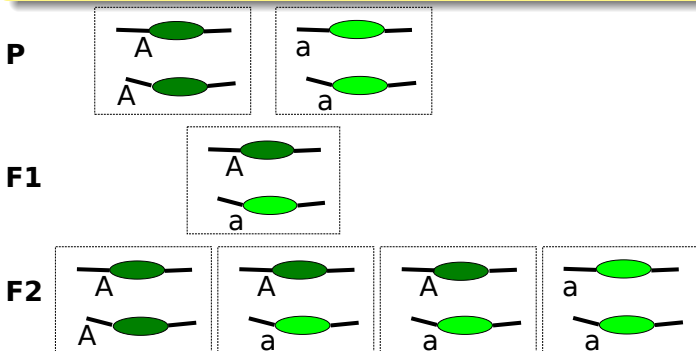


Dla n wariantów:

- ▶ n różnych homozygot
- ▶ $\frac{n*(n-1)}{2}$ różnych heterozygot

Pierwsze prawo Mendla (prawo czystości gamet)

każda komórka płciowa (gameta) posiada tylko jeden allel z danej pary alleli



W pokoleniu F1 nie obserwuje się rozszczepienia.

W pokoleniu F2 rozszczepienie w stosunku 3:1.

Pierwsze prawo Mendla - przykład

U bydła bezrożność dominuje nad występowaniem rogów. Bezrożnego buhaja skojarzono z 2 krowami. Cielę uzyskane ze skojarzenia z krową nr 1, rogatą było rogate, cielę uzyskane z krowy nr 2, bezrożnej było rogate. Podaj genotypy buhaja oraz krów.

Pierwsze prawo Mendla - przykład

U bydła bezrożność dominuje nad występowaniem rogów. Bezrożnego buhaja skojarzono z 2 krowami. Cielę uzyskane ze skojarzenia z krową nr 1, rogatą było rogate, cielę uzyskane z krowy nr 2, bezrożnej było rogate. Podaj genotypy buhaja oraz krów.

R - allel: bezrożność

r - allel: występowanie rogów

RR - bezrożne, **Rr** - bezrożne, **rr** - rogate

Pierwsze prawo Mendla - przykład

U bydła bezrożność dominuje nad występowaniem rogów. Bezrożnego buhaja skojarzono z 2 krowami. Cielę uzyskane ze skojarzenia z krową nr 1, rogatą było rogate, cielę uzyskane z krowy nr 2, bezrożnej było rogate. Podaj genotypy buhaja oraz krów.

R - allel: bezrożność

r - allel: występowanie rogów

RR - bezrożne, **Rr** - bezrożne, **rr** - rogate

buhaj **Rr**

krowa 1 **rr**, cielę **rr**

krowa 2 **Rr**, cielę **rr**

Równowaga Hardy'ego-Weinberga

Locus: 2 warianty A i a

- ▶ P_A - częstość allele A
- ▶ $P_a = 1 - P_A$ - częstość allele a

Dla potomków:

$$\begin{aligned}P_{Aa} &= P(\{\text{mat} = A \text{ i oj} = a\} \cup \{\text{mat} = a \text{ i oj} = A\}) \\ &= P(\text{mat} = A)P(\text{oj} = a) + P(\text{mat} = a)P(\text{oj} = A) \\ &= 2P_AP_a\end{aligned}$$

$$P_{AA} = (P_A)^2$$

$$P_{aa} = (P_a)^2$$

Równowaga Hardy'ego-Weinberga (2)

	A	a
A		
a		

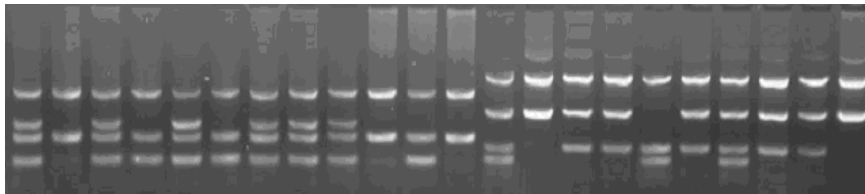
- ▶ nieskończenie duża populacja
- ▶ krzyżowanie w sposób losowy (brak preferencji w doborze partnerów)
- ▶ brak mutacji
- ▶ te same częstości wariantów dla różnych płci

Markery genetyczne

- ▶ markery STR (short tandem repeat)
- ▶ markery SNP (single nucleotide polymorphism)

Wykorzystanie:

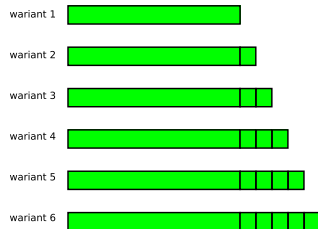
- ▶ wykrywanie chorób genetycznych
- ▶ badanie pokrewieństw
 - ▶ potwierdzanie / wykluczanie rodzicielstwa (używając genotypu potencjalnego rodzica, używając genotypów krewnych, itd.)
 - ▶ badanie pokrewieństw pomiędzy kuzynami
- ▶ kryminalistyka (badanie śladów, badanie mieszanin)



Markery genetyczne (2)

STR

- ▶ warianty różniące się długością
- ▶ kilka - kilkanaście wariantów dla danego locus
- ▶ prosty odczyt



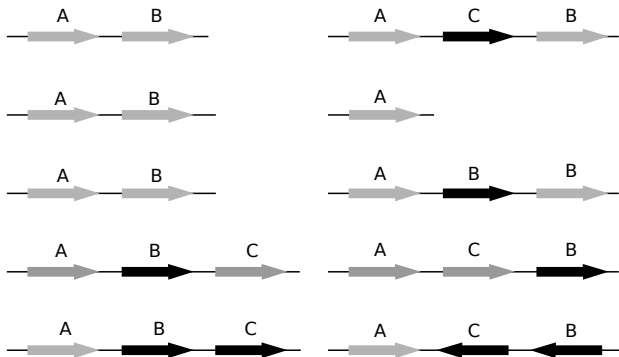
SNP

- ▶ warianty różniące się sekwencją jednego nukleotydu
- ▶ zazwyczaj dwa warianty
- ▶ złożony odczyt
- ▶ miliony znanych różnic (<http://hapmap.ncbi.nlm.nih.gov/>)



Modyfikacje dużych fragmentów

- ▶ insercje
- ▶ delecje
- ▶ duplikacje
- ▶ transpozycje
- ▶ inwersje



warianty genetyczne dla człowieka - statystyki

Znane modyfikacje (<http://www.hapmap.org>)

SNP:	$11 * 10^6$
małe inderscje/delecje (do 10000nt):	$3 * 10^6$
rearanżacje:	$3 * 10^4$

Dla zdrowego człowieka (<http://www.1000genomes.org/>)

- ▶ SNP (wszystkich $3 * 10^6$), w sekwencjach kodujących:

wstawienie kodonu stop (nonsense)	1057
utrata kodonu stop	77
zmiana kodonu (missense)	≈ 68000
bez zmiany kodonu (silent)	≈ 60000

- ▶ small indel (do 10000nt): $362 * 10^3$

- ▶ rearanżacje (zajmują ok. 10% genomu)

usunięcia	≈ 16000
insercje	4775
duplikacje	407

Badanie pokrewieństw za pomocą analizy markerów

Pojedyncze locus może wykluczyć pokrewieństwo

	matka	dziecko	X1	X2	X3
wariant 1	-	-	+	-	-
wariant 2	+	-	+	-	-
wariant 3	+	+	-	+	+
wariant 4	-	-	-	-	-
wariant 5	-	+	-	-	+
wariant 6	-	-	-	-	-

$$PI(W, X) = \frac{X(W)}{Y(W)} = \frac{\text{X dostarczył allel}}{\text{ktoś inny dostarczył allel}}$$

$$X(W) = \begin{cases} 1 & \text{homozygota } WW \\ 0.5 & \text{heterozygota } WZ \\ 0 & \text{brak wariantu} \end{cases}$$

$Y(W)$ - częstość W w populacji

$$Y(w5) = 0.12$$

$$PI(w5, X1) = 0$$

$$PI(w5, X2) = 0$$

$$PI(w5, X3) = 4.17$$

Badanie pokrewieństw (2)

$$CPI(X) = \prod_{i=0}^n PI(loci_i, X) \text{ (combined paternity index)}$$

gdzie PI to paternity index (omówiony wcześniej)

$$PP(X) = \frac{CPI(X)}{CPI(X) + 1} \text{ (probability of paternity)}$$

CPI	PP(%)	opis
>399	99.8 - 99.9	praktycznie pewne
>99	99.1 - 99.8	bardzo wysokie prawdopodobieństwo
>19	95 - 99	wysokie prawdopodobieństwo
>9	90 - 95	prawdopodobne

badanie pokrewieństw (3) - grupy krwi

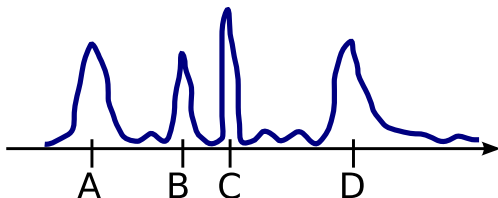
3 warianty: 0, A, B

fenotyp, grupa krwi	genotyp
0	00
A	A0, AA
B	B0, BB
AB	AB

Grupy krwi dzieci:

-	0	A	B	AB
0	0	0,A	0,B	A,B
A	0,A	0,A	0,A,B,AB	A,B,AB
B	0,B	0,A,B,AB	0,B	A,B,AB
AB	A,B	A,B,AB	A,B,AB	A,B,AB

Badanie mieszanin



- ▶ locus: 4 warianty (A, B, C, D), 6 różnych genotypów

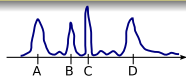
osobnik 1	osobnik 2
A B	C D
A C	B D
A D	B C
B C	A D
B D	A C
C D	A B

- ▶ 2 osobniki w mieszaninie:

Badanie mieszanin, obliczanie prawdopodobieństwa hipotez

$$LR = \frac{P(E|H_p)}{P(E|H_d)}$$

gdzie: H_p hipoteza oskarżyciela, H_d hipoteza obrońcy, $P(E|H_p)$ prawdopodobieństwo warunkowe obserwacji



Częstości alleli

p_a	0.1
p_b	0.2
p_c	0.3
p_d	0.4

osobnik1	osobnik2	$P(os1 \cap os2)$
A B	C D	0.0096
A C	B D	0.0096
A D	B C	0.0096
B C	A D	0.0096
B D	A C	0.0096
C D	A B	0.0096

$$P(E) = 0.0576$$

- ▶ H_p : ofiara:AB, podejrzany:CD; H_d : ofiara:AB, podejrzany:?

$$P(E|H_p) = 1, P(E|H_d) = 2p_c p_d, LR = 4.17$$

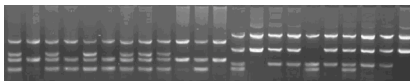
- ▶ H_p : ofiara:?, podejrzany:CD; H_d : ofiara:?, podejrzany:?

$$P(E|H_p) = 2p_a p_b, P(E|H_d) = P(E), LR = 0.69$$

Analiza haplotypów

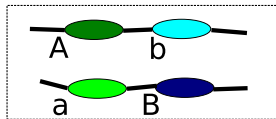
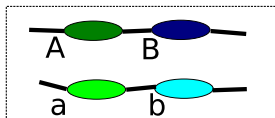
Analiza haplotypów

Popularne metody genotypowania (PCR)



- ▶ brak informacji o haplotypach (o fazie)
- ▶ łatwo ustalić mając dane od wielopokoleniowych rodzin

AaBb



Metoda rekonstrukcji ⁴

- ▶ badanie genotypów, które są homozygotą dla każdego locus
genotyp $A_1A_1B_1B_1C_1C_1$ zawiera haplotypy $A_1B_1C_1$
- ▶ badanie genotypów, które są heterozygotą tylko na jednym locus
genotyp $A_1A_1B_1B_1C_1C_2$ zawiera haplotypy $A_1B_1C_1$ i $A_1B_1C_2$
- ▶ badanie pozostałych genotypów
 - ▶ czy mogą być utworzone przez istniejące haplotypy
 - ▶ dodawanie nowych haplotypów do listy znanych
 - ▶ kontynuacja aż wszystkie genotypy będą rozpatrzone albo nie można dodać nowych haplotypów

⁴Clark, Inference of haplotypes from PCR-amplified samples of diploid populations, 1990

Rekonstrukcja haplotypów (2)

Rekonstrukcja haplotypów - przykład

$A_1A_1B_1B_1C_1C_1$, $A_1A_1B_1B_1C_1C_2$, $A_1A_2B_1B_1C_1C_1$

- ▶ haplotyp $A_1B_1C_1$ na podstawie 1 (homozygota)
- ▶ haplotyp $A_1B_1C_2$ na podstawie 2 (jedno locus jest heterozygotą)
- ▶ haplotyp $A_2B_1C_1$ na podstawie 3 (bo $A_1A_2B_1B_1C_1C_1$ może być utworzone przez $A_1B_1C_1$ i $A_2B_1C_1$)

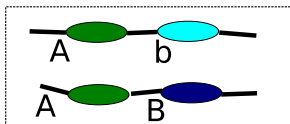
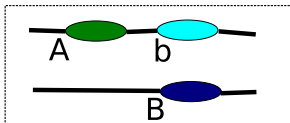
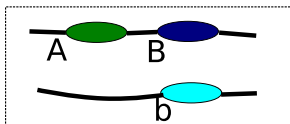
Rekonstrukcja haplotypów - problemy

- ▶ brak genotypów które są homozygotą na każdym locus albo na każdym oprócz jednego
- ▶ duża liczba nierozpatrzonych genotypów po zakończeniu algorytmu

Nieme warianty w analizie haplotypów

polimorfizm polegający na tym, że gen może nie wystąpić, więc w niektórych przypadkach brak informacji o tym, czy homozygota, czy heterozygota

ABb



Analiza danych populacyjnych

- ▶ nie potrzebuje danych od wielopokoleniowych rodzin
- ▶ znajduje najbardziej prawdopodobne haplotypy
- ▶ na podstawie prawdopodobieństw haplotypów wnioskuje się o prawdopodobieństwach układu haplotypów dla danego genotypu

Założenia:

- ▶ równowaga Hardy'ego-Weinberga
- ▶ losowy dobór osobników do próby
- ▶ ten sam rozkład prawdopodobieństw haplotypów w obrębie próby

Analiza probabilistyczna danych populacyjnych

S obserwacja, G różnych genotypów, n osobników

$$S = \langle n_1, n_2, \dots, n_G \rangle, \sum_{j=0}^G n_j = n$$

Osobnicy są dobierane niezależnie

$$P(S|g_1, g_2, \dots, g_G) = \frac{n!}{n_1! * n_2! * \dots * n_G!} * \prod_{j=1}^G g_j^{n_j} = \alpha \prod_{j=1}^G g_j^{n_j}$$

Szacowanie prawdopodobieństwa haplotypów h_j .

$$\arg \max_{h_1, \dots, h_H} P(S|h_1, \dots, h_H) = \arg \max_{h_1, \dots, h_H} \prod_{j=1}^G \left(\sum_{i=0}^{r_j} z_{mn} \right)^{n_j}$$

$$\text{gdzie } z_{mn} = \begin{cases} h_m^2 & \text{dla } m = n \\ 2h_m h_n & \text{dla } m \neq n \end{cases}$$

Analiza złożoności

Liczba układów (par) haplotypów:

$$R = \frac{1}{2}H * (H + 1), H = \prod_{i=1}^k l_i, \quad \text{gdzie } \begin{array}{l} k \text{ liczba loci,} \\ l_i \text{ liczba wariantów dla loci} \end{array}$$

Liczba genotypów:

$$G = \prod_{i=1}^k \frac{(l_i - \delta_i)(l_i + 1 - \delta_i) + 2\delta_i}{2}, \delta_i = \begin{cases} 1 & \text{loci ma niemy allel} \\ 0 & \end{cases}$$

Liczba układów haplotypów dla genotypu j :

$$r_j = \begin{cases} 2^{s_j-1} * 3^{t_j} & \text{dla } s_j > 0 \\ \frac{3^{t_j+1}}{2} & \text{dla } s_j = 0 \end{cases}$$

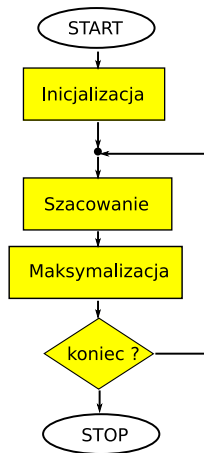
s_j liczba obserwowanych heterozygot

t_j liczba obserwowanych wariantów, które nie są nieme dla loci posiadających niemy wariant

Algorytm EM użyty do odtwarzania haplotypów ⁵

Algorytm maksymalizacji oczekiwań (Expectation Maximization)

- ▶ cykliczne powtarzanie:
 - ▶ przewidywania parametrów (krok E)
 - ▶ maksymalizacja funkcji celu (krok M)
- ▶ kryterium stopu: brak zmian w kolejnych cyklach
- ▶ optymalizacja lokalna
- ▶ szybka zbieżność



⁵Excoffier, Slatkin, Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population, 1995

Algorytm EM - przykład z monetami

Obserwujemy 5 serii rzutów monetą A lub monetą B

seria	wynik	ilość O	ilość R
1	O R R R O O R O O R	5	5
2	R O O O O O O O R O	8	2
3	R O R O R O O O O R	6	4
4	O O R R R R R R O R	3	7
5	O O O O O R O O O O	9	1

- ▶ Jakie jest prawdopodobieństwo wyrzucenia orła przez monetę A ?
- ▶ Jakie jest prawdopodobieństwo wyrzucenia orła przez monetę B ?

Algorytm EM - przykład z monetami (2)

Obserwujemy 5 serii rzutów monetą A lub monetą B

seria	wynik	ilość O	ilość R
1	O R R R O O R O O R	5	5
2	R O O O O O O O R O	8	2
3	R O R O R O O O O R	6	4
4	O O R R R R R R O R	3	7
5	O O O O O R O O O O	9	1

Jeżeli wiemy, że seria 1, 3 i 4 była monetą A, zaś seria 2 i 5 monetą B:

$$\blacktriangleright \theta(A) = \frac{5+6+3}{30} \approx 0.47$$

$$\blacktriangleright \theta(B) = \frac{8+9}{20} \approx 0.85$$

Algorytm EM - przykład z monetami (3)

Nie wiemy, która seria monetą A, a która monetą B.

seria	wynik										ilość O	ilość R
1	O	R	R	R	O	O	R	O	O	R	5	5
2	R	O	O	O	O	O	O	O	R	O	8	2
3	R	O	R	O	R	O	O	O	O	R	6	4
4	O	O	R	R	R	R	R	R	O	R	3	7
5	O	O	O	O	O	R	O	O	O	O	9	1

losujemy początkowe wartości, np: $\theta_0(A) = 0.6$, $\theta_0(B) = 0.5$

Wtedy 1 seria monetą A:

$$\binom{10}{5} \theta_0(A)^5 (1 - \theta_0(A))^{10-5} = \alpha 0.6^5 (1 - 0.6)^5 \approx 0.000796 \alpha$$

gdzie $\alpha = \binom{10}{5}$. 1 seria monetą B to $\approx 0.000977 \alpha$.

$$P_0(A) = \frac{0.000796}{0.000796 + 0.000977} \approx 0.45, P_0(B) \approx 0.55.$$

Algorytm EM - przykład z monetami (4)

seria	wynik										n_O	n_R	$P_0(A)$	$P_0(B)$
1	O	R	R	R	O	O	R	O	O	R	5	5	0.45	0.55
2	R	O	O	O	O	O	O	O	R	O	8	2	0.73	0.27
3	R	O	R	O	R	O	O	O	O	R	6	4	0.55	0.45
4	O	O	R	R	R	R	R	R	O	R	3	7	0.27	0.73
5	O	O	O	O	O	R	O	O	O	O	9	1	0.80	0.20

► monetą A wyrzuciliśmy

orły: $5 * 0.45 + 8 * 0.73 + 6 * 0.55 + 3 * 0.27 + 9 * 0.80 = 19.40$

reszki: $5 * 0.45 + 2 * 0.73 + 4 * 0.55 + 7 * 0.27 + 1 * 0.80 = 8.60$

więc $\theta_1(A) = \frac{19.4}{19.4+8.6} = 0.67$.

► $\theta_1(B) = \frac{11.6}{11.6+10.4} = 0.53$

Powtarzamy iteracje obliczając $\theta_2(A), \theta_2(B), \theta_3(A), \theta_3(B), \dots$

Algorytm EM uwzględniając nieme warianty⁶

- ▶ inicjacja:

$$z_{mn}^{(0)} = \frac{1}{r_j}, \text{ gdzie układ } mn \text{ daje genotyp } j$$

- ▶ krok M:

$$z_{mn}^{(t+1)} = \frac{n_j}{n} * \frac{z_{mn}^{(t)}}{g_j^{(t)}}, \text{ gdzie } mn \text{ daje genotyp } j, g_j^{(t)} = \sum_x z_x^{(t)}$$

- ▶ krok E:

$$z_{mn}^{(t+1)} = \begin{cases} (h_m^{(t)})^2 & \text{dla } m = n \\ 2 h_m^{(t)} h_n^{(t)} & \text{dla } m \neq n \end{cases} \quad h_m^{(t)} = \frac{1}{2} \left(\sum_i z_{im}^{(t)} + \sum_j z_{mj}^{(t)} \right)$$

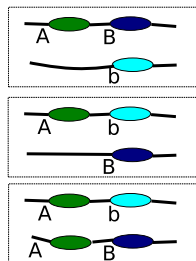
⁶Nowak, Ploski, NullHap - a versatile application to estimate haplotype frequencies from unphased genotypes in the presence of null-alleles, 2008

Analiza haplotypów: przykład

2 loci, każde z nich dwa allele **A**, **a**, **B**, **b**, a niemy allele

genotyp	układ haplotypów
AB	AB/AB, AB/aB
ABb	AB/Ab, AB/ab, Ab/aB
Ab	Ab/Ab, Ab/ab
aB	aB/aB
aBb	aB/ab
ab	ab/ab

ABb ?



$$\begin{array}{l}
 Z_{AB/AB} = 0.5 \\
 \text{Inicjacja: } Z_{AB/Ab} = 0.33 \\
 Z_{AB/aB} = 0.5 \\
 Z_{AB/ab} = 0.33
 \end{array}
 \qquad
 \begin{array}{l}
 Z_{Ab/Ab} = 0.5 \\
 Z_{Ab/aB} = 0.33 \\
 Z_{Ab/ab} = 0.5
 \end{array}
 \qquad
 \begin{array}{l}
 Z_{aB/aB} = 1 \\
 Z_{aB/ab} = 1 \\
 Z_{ab/ab} = 1
 \end{array}$$

Analiza haplotypów: przykład (2), krok M

Przykładowa obserwacja (25 osobników):

genotyp	liczba obserwacji	genotyp	liczba obserwacji
AB	3	aB	1
ABb	8	aBb	4
Ab	5	ab	4

$$z'_{AB/AB} = \frac{n_{AB}}{n} * \frac{z_{AB/AB}}{g_{AB}} = 0.06$$

$$z'_{AB/Ab} = \frac{n_{ABb}}{n} * \frac{z_{AB/Ab}}{g_{ABb}} = 0.11$$

$$z'_{AB/aB} = \frac{n_{AB}}{n} * \frac{z_{AB/aB}}{g_{AB}} = 0.06$$

$$z'_{AB/ab} = \frac{n_{ABb}}{n} * \frac{z_{AB/ab}}{g_{ABb}} = 0.11$$

$$z'_{Ab/Ab} = \frac{n_{Ab}}{n} * \frac{z_{Ab/Ab}}{g_{Ab}} = 0.1$$

genotyp	układ haplotypów
AB	AB/AB, AB/aB
ABb	AB/Ab, AB/ab, Ab/aB
Ab	Ab/Ab, Ab/ab
aB	aB/aB
aBb	aB/ab
ab	ab/ab

$$z'_{Ab/aB} = \frac{n_{ABb}}{n} * \frac{z_{Ab/aB}}{g_{ABb}} = 0.11$$

$$z'_{Ab/ab} = \frac{n_{Ab}}{n} * \frac{z_{Ab/ab}}{g_{Ab}} = 0.1$$

$$z'_{aB/aB} = \frac{n_{aB}}{n} = 0.04$$

$$z'_{aB/ab} = 0.16$$

$$z'_{ab/ab} = 0.16$$

Analiza haplotypów: przykład (3), krok E

$$h_{AB} = z_{AB/AB} + \frac{z_{AB/Ab} + z_{AB/aB} + z_{AB/ab}}{2} = 0.197$$

$$h_{Ab} = z_{Ab/Ab} + \frac{z_{AB/Ab} + z_{Ab/aB} + z_{Ab/ab}}{2} = 0.26$$

$$h_{aB} = z_{aB/aB} + \frac{z_{AB/aB} + z_{Ab/aB} + z_{aB/ab}}{2} = 0.203$$

$$h_{ab} = z_{ab/ab} + \frac{z_{AB/ab} + z_{Ab/ab} + z_{aB/ab}}{2} = 0.343$$

$$z_{AB/AB} = 0.039 \quad z_{Ab/Ab} = 0.066 \quad z_{aB/aB} = 0.041$$

$$z_{AB/Ab} = 0.1 \quad z_{AB/aB} = 0.1 \quad z_{Ab/ab} = 0.14$$

$$z_{AB/aB} = 0.08 \quad z_{Ab/ab} = 0.18 \quad z_{aB/ab} = 0.12$$

$$z_{AB/ab} = 0.14$$

Analiza haplotypów: przykład (4), kolejne kroki

$$g_{AB} = z_{AB/AB} + z_{AB/aB} = 0.12$$

wynik:

$$g_{ABb} = z_{AB/Ab} + z_{AB/ab} + z_{Ab/aB} = 0.34$$

$$h_{AB} = 0.2$$

$$g_{Ab} = z_{Ab/Ab} + z_{Ab/ab} = 0.24$$

$$h_{Ab} = 0.2$$

$$g_{aB} = z_{aB/aB} = 0.041$$

$$h_{aB} = 0.2$$

$$g_{aBb} = z_{aB/ab} = 0.14$$

$$h_{ab} = 0.4$$

$$g_{ab} = z_{ab/ab} = 0.12$$

wnioski

- ▶ najczęstszy haplotyp (w grupie) to ab (40%)
- ▶ gdy obserwujemy ABb, to
 - ▶ układ AB/Ab z częstością 25%
 - ▶ układ AB/ab z częstością 50%
 - ▶ układ Ab/aB z częstością 25%

Dziękuję