Algorytmy w bioinformatyce sekwencji

markery genetyczne, haloptypy, algorytm EM

Robert Nowak

2024

Resekwencjonowanie

Resekwencjonowanie - etapy

Resekwencjonowanie:

- mapowanie odczytów na genom referencyjny
- oznaczanie duplikatów
- rekalibracja oceny jakości nukleotydów

Później - analiza wariantów

- możliwe analizy dla nadmiarowości (coverage) < 1
- możliwe analizy dla części genomu (np. exony)

Mapowanie odczytów na genom referencyjny

Plik SAM (Sequence Alighment/Map)

mapowanie wykorzystuje transofmatę Burrowsa-Wheelera

plik tekstowy (tabela), zawiera informację o mapowaniu odczytów na genom referencyjny, zawiera:

- identyfikator odczytu
- identyfikator kontiga (na który odczyt jest mapowany)
- pozycja (indeks) mapowania
- jakość mapowania, flagi, dodatkowe informacje (np. insercje i delecje)

Plik BAM: kompresja pliku SAM

Plik SAM - przykład

Obok podano fragment pliku SAM dla sekwencji referencyjnej ACTGG-GACCTA (indeks od 1). Wyznacz potencjalne warianty, podaj pokrycie i określ, czy jest on homo- czy heterozygotyczny.

POS	CIGAR	SEQ
2	5M1D1M	CTGGGC
4	3M1D3M	GTGCCT
5	2M1D4M	TGCCTA
5	2M1D4M	GGCCTA

A	C	T	G	G	G	A	C	C	T	A	
	C	T	G	G	G	*	C				5M1D1M
			G	T	G	*	C	C	T		3M1D3M
				T	G	*	C	C	T	A	2M1D4M
				G	G	*	C	C	T	A	2M1D4M

Plik SAM - przykład

Obok podano fragment pliku SAM dla sekwencji referencyjnej ACTGG-GACCTA (indeks od 1). Wyznacz potencjalne warianty, podaj pokrycie i określ, czy jest on homo- czy heterozygotyczny.

POS	CIGAR	SEQ
2	5M1D1M	CTGGGC
4	3M1D3M	GTGCCT
5	2M1D4M	TGCCTA
5	2M1D4M	GGCCTA

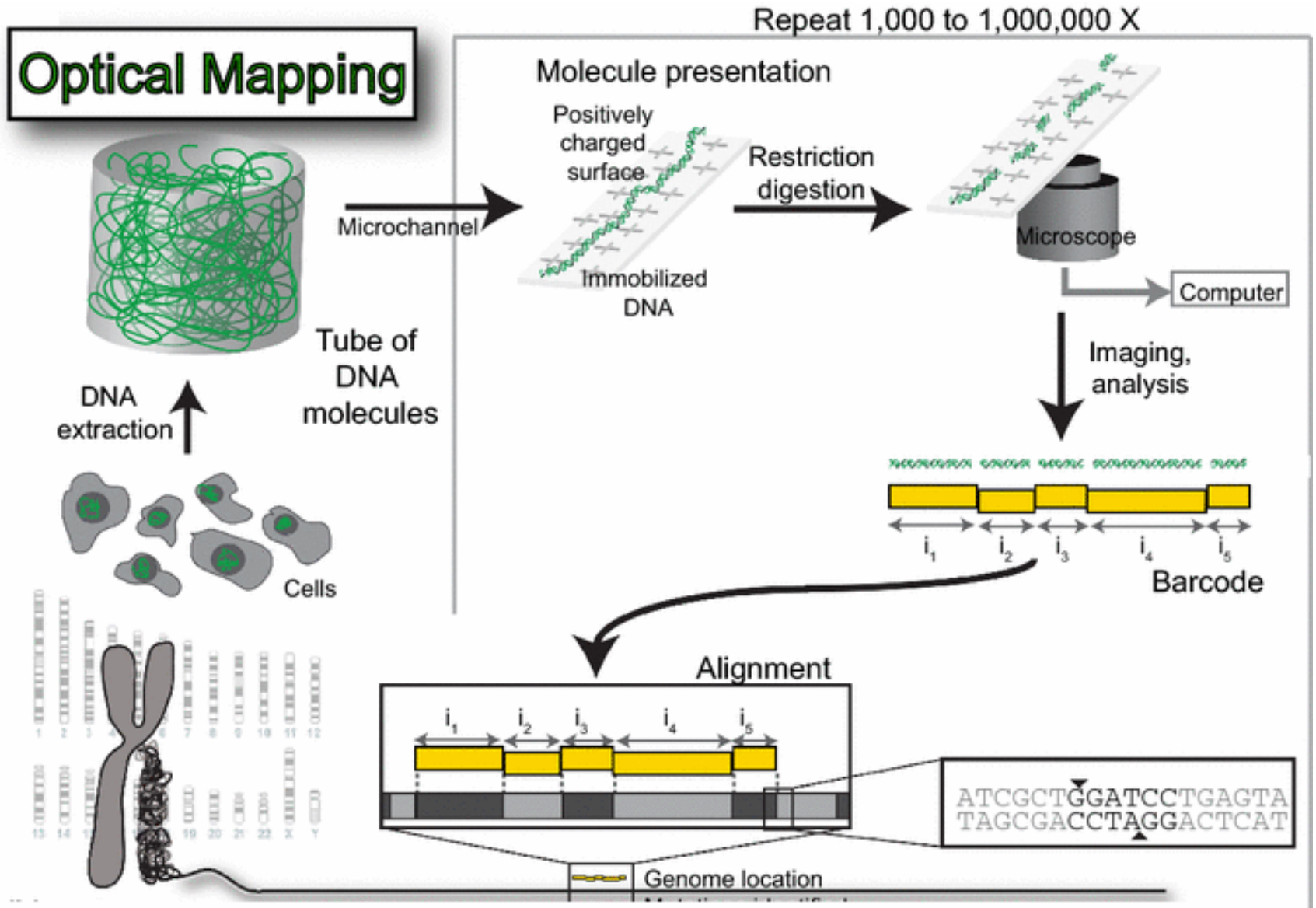
A	C	T	G	G	G	A	C	C	T	A	
	C	T	G	G	G	*	C				5M1D1M
			G	T	G	*	C	C	T		3M1D3M
				T	G	*	C	C	T	A	2M1D4M
				G	G	*	C	C	T	A	2M1D4M

Są dwa warianty:

- wariant heterozygotyczny na pozycji 5, G>T
- delecja homozygotyczny na pozycji 7, A>_

Sekwencjonowanie optyczne, mapy restrykcyjne

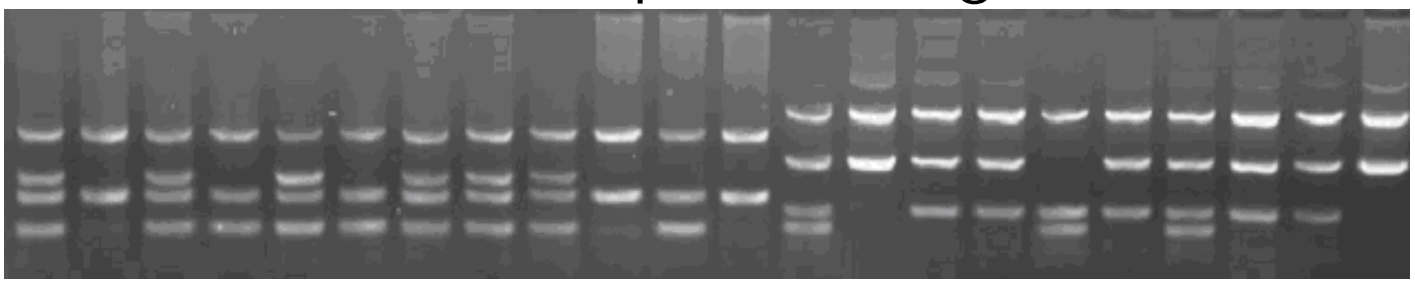
Sekwencjonowanie optyczne



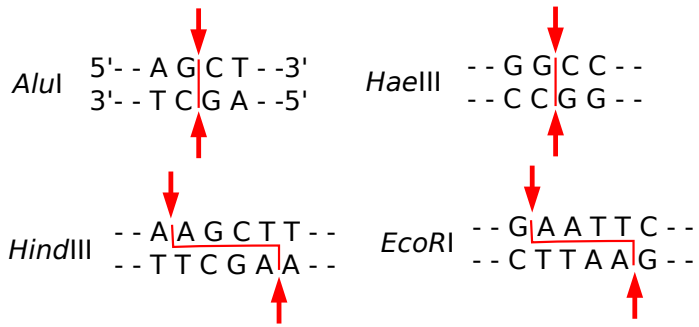
Mapy uproszczone - mapy restrykcyjne

Mapy restrykcyjne można uzyskać bez konieczności odczytywania sekwencji

- cząsteczki DNA trudno obserwować bezpośrednio
- elektroforeza - metoda pomiaru długości łańcucha



- enzym restrykcyjny tnie DNA w miejscu, które zawiera wzorzec specyficzny dla danego enzymu



Problem częściowego strawienia (partial digest problem)

$X = \{x_1, x_2, ..., x_n\}$ zbiór n pozycji

$\Delta X = \{x_j - x_i : i < j\}$ multi-zbiór odległości $|\Delta X| = \binom{n}{2}$

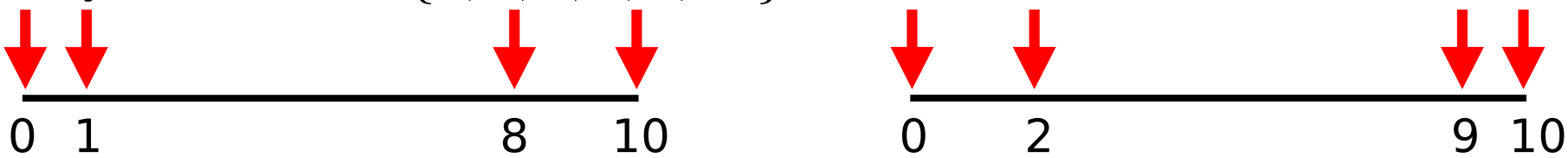
Przykład 1: $X = \{0, 5, 7\}, \Delta X = \{2, 5, 7\}$

Przykład 2: $X = \{0, 5, 7, 9\}, \Delta X = \{2, 2, 4, 5, 7, 9\}$

PDP: dla danego ΔX znaleźć (wszystkie) X

- typowo 1 rozwiązanie
- $H(n)$ - liczba rozwiązań dla zbioru n elementowego, $H(n) \leq \frac{1}{2} n^{1.233}$

Przykład: $\Delta X = \{1, 2, 7, 8, 9, 10\}$



$$X_1 = \{0, 1, 8, 10\}, X_2 = \{0, 2, 9, 10\}$$

PDP strawienia - rekurencja z nawrotami

```
bool partialDigest(set L) { //L - zbiór różnic
  width = deleteMax(L);
  X = { 0, width };
  return place(L,X);
}
bool place(set L, set X) { //L - zbiór różnic, X - zbiór miejsc
  if ( empty L ) return true;
  x = deleteMax(L); //nowy fragment umieszczony od lewej
  if ( delta(x, X ) in L ) { //oblicza różnice odległości
    new_X = X + x;
    new_L = L - delta(x, X);
    if( place(new_L, new_X) ) return true;
  }
  x = width - x; //nowy fragment umieszczony od prawej
  if ( delta(x, X in L ) {
    new_X = X + x; new_L = L - delta(x, X);
    if( place(new_L, new_X) ) return true;
  }
  return false;
}
```

PDP - rekurencja z nawrotami (2)

Przykład $L = \{1, 2, 7, 8, 9, 10\}$

width = 10 place()	X={0,10}, L={1,2,7,8,9} x = 9, delta(x,X) = {1,9} place() return true	X = {0,9,10}, L = {2,7,8} x = 8, delta(x,X) = {1,2,8} x = 2, delta(x,X) = {2,7,8} place() return true	X = {0,2,9,10}, L = { } return true

- złożoność pesymistyczna:

$$O(2^n n \log n)$$

- złożoność średnia

$$O(n^2 \log n)$$

PDP - problemy podobne

- uproszczony problem częściowego strawienie (SPDP), przeprowadzane są dwa doświadczenia: częściowe i pełne trawienie, problem NP-trudny (złożoność wykładnicza)[1]
- znakowany problem częściowego strawienia (labeled PDP), końce cząsteczki znakowane (np. radioaktywnie), następnie częściowe trawienie, złożoność wielomianowa[2]
- cięcie dwoma enzymami (Double Digest Problem, DDP), przeprowadzane są trzy doświadczenia dla dwóch różnych enzymów, trawienie pełne jednym enzymem, trawienie pełne drugim, trawienie pełne jednym i drugim, problem NP-trudny

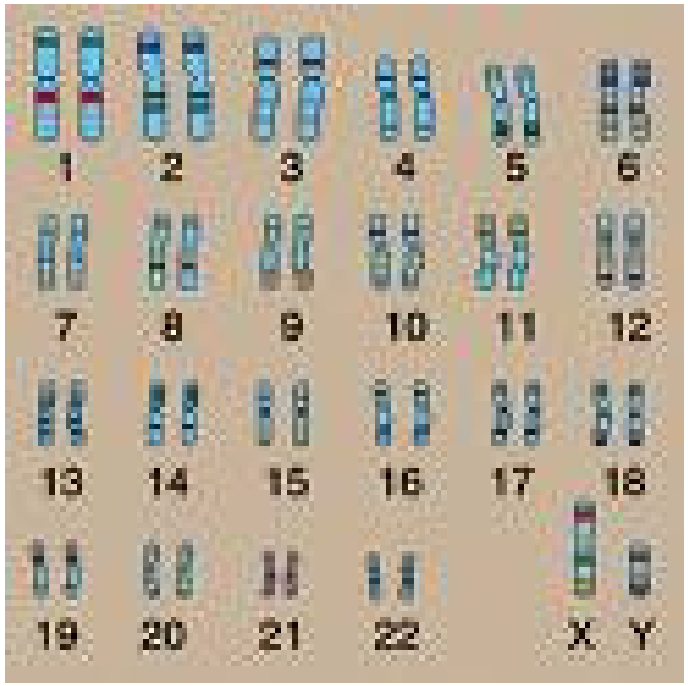
[1] J.Blazewicz, M.Kasprzyk, On the complexity of DNA Simplified Partial Digest Problem (2006)

[2] G.Pandurangan, H.Ramesh, The restriction mapping problem revised (2002)

Analiza markerów genetycznych

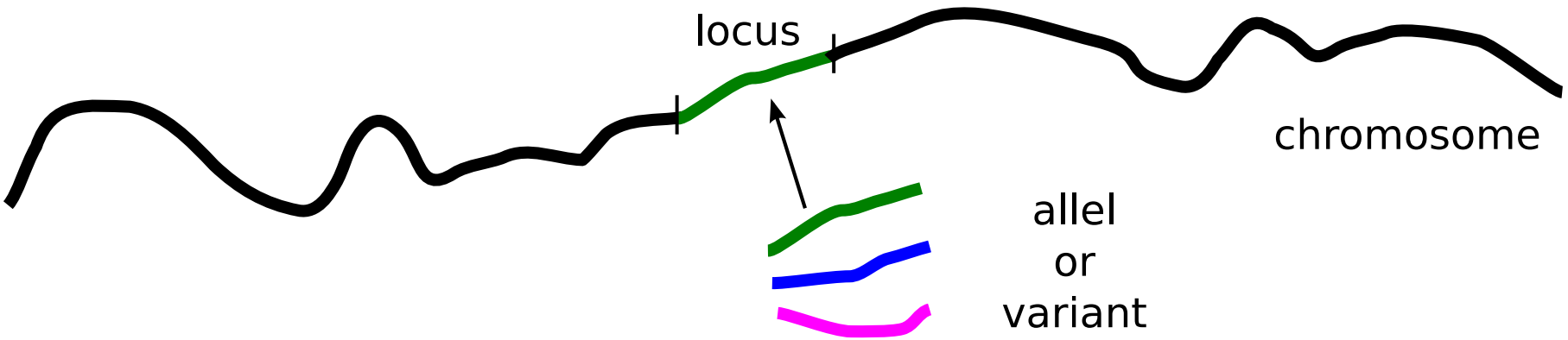
Genetyka populacyjna, chromosomy

- 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)



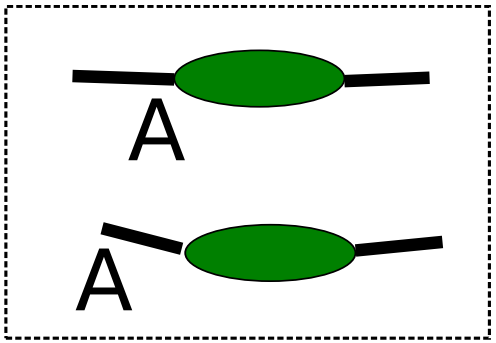
Definicje : locus, allel, wariant genu

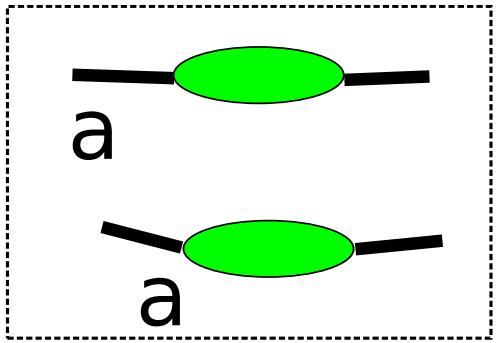
- locus - miejsce na chromosomie
- allele - wariant genu
- haplotyp - zbiór wariantów (alleli), które są przekazywane razem

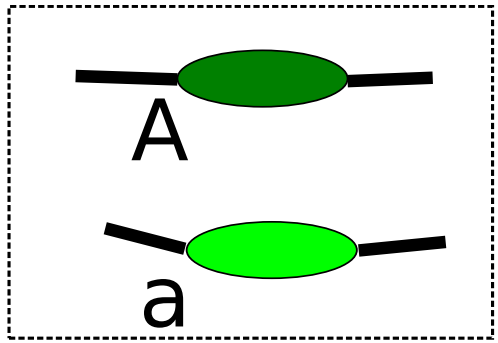


Definicje: homozygota, heterozygota

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







Dla n wariantów:

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

Równowaga Hardy'ego-Weinberga

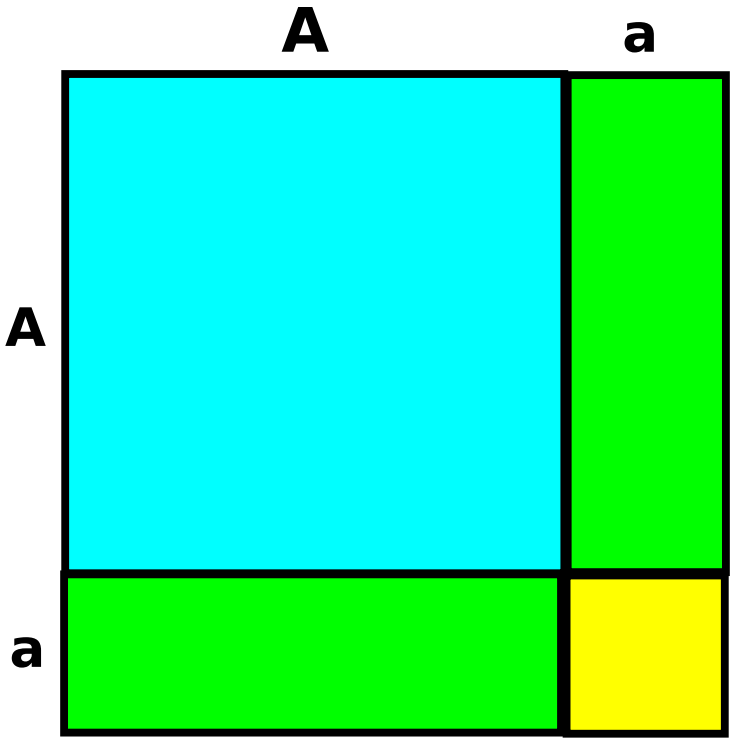
Locus: 2 warianty A i a

- P_A - częstość allela A
- $P_a = 1 - P_A$ - częstość allela a

Dla potomków:

$$\begin{aligned} P_{Aa} &= P(\{\text{mat} = A \text{ i oj} = a\} \cup \{\text{mat} = a \text{ i oj} = A\}) \\ &= P(\text{mat} = A)P(\text{oj} = a) + P(\text{mat} = a)P(\text{oj} = A) \\ &= 2P_A P_a \\ P_{AA} &= (P_A)^2 \\ P_{aa} &= (P_a)^2 \end{aligned}$$

Równowaga Hardy'ego-Weinberga (2)



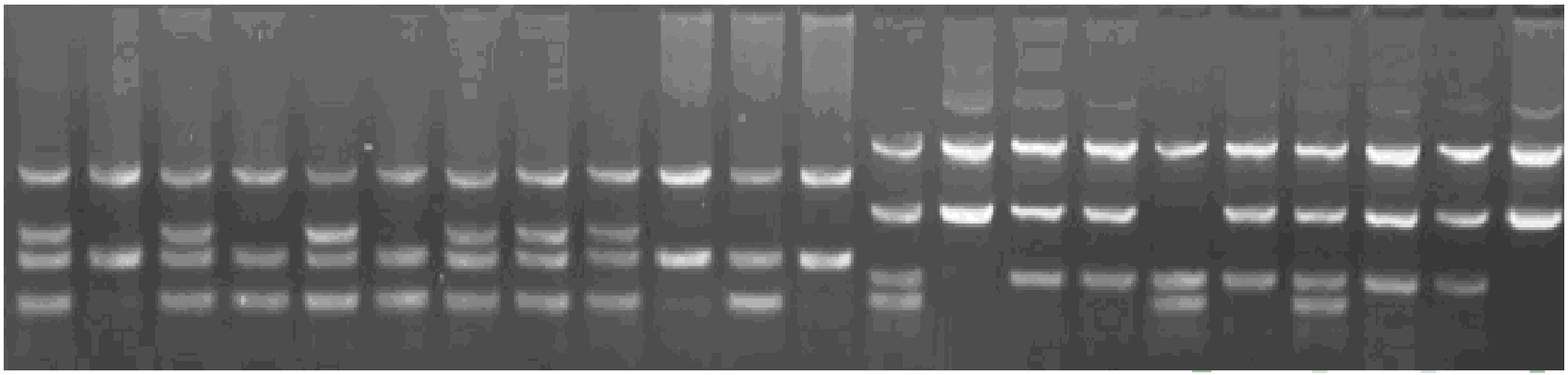
- nieskończenie duża populacja
- krzyżowanie w sposób losowy (brak preferencji w doborze partnerów)
- brak mutacji
- te same częstości wariantów dla różnych płci

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

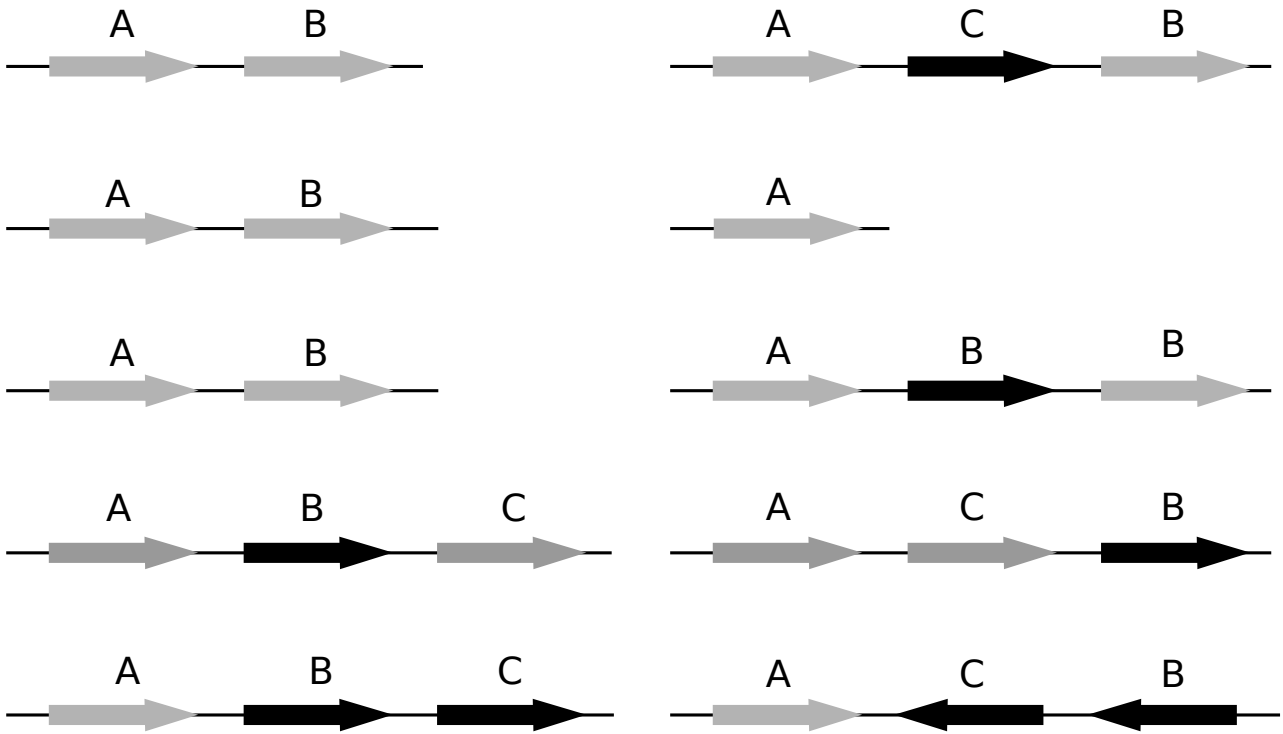
- warianty różniące się długością
- kilka - kilkanaście wariantów dla danego locus
- prosty odczyt

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`)

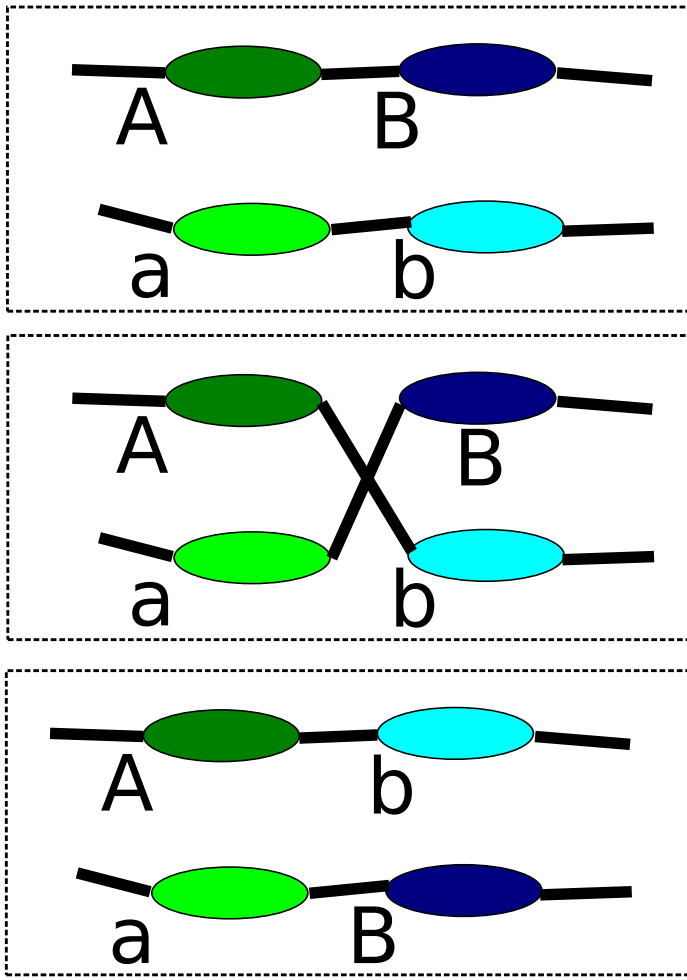
- SNP : $11 * 10^6$
- indel (do $10000nt$) : $3 * 10^6$
- rearanżacje: $3 * 10^4$ (zajmują ok. 10% genomu)

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) : $68 * 10^3$
  - bez zmiany kodonu (silent) : $60 * 10^3$
- small indel (do $10000nt$): $362 * 10^3$
- usunięcia: $16 * 10^3$
- insercje: 4775
- duplikacje: 407

Mapa genetyczna i fizyczna

- Mapa genetyczna - mapa lokalizacji genów lub markerów na chromosomie powstała poprzez badanie osobników w krzyżówce testowej. Jednostka – centymorgany (cM).
- Mapa fizyczna - mapa powstała poprzez odczyt sekwnecji. Jednostka - nt (nukleotydy) lub bp (pary zasad).



cM - prawdopodobieństwo rozdzielenia w jednym pokoleniu podczas rekombinacji (cross-over) wynosi 1%

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

$$\begin{aligned} Y(w5) &= 0.12 \\ PI(w5,X1) &= 0 \\ PI(w5,X2) &= 0 \\ PI(w5,X3) &= 4.17 \end{aligned}$$

Badanie pokrewieństw (2)

PI: paternity index

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

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

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

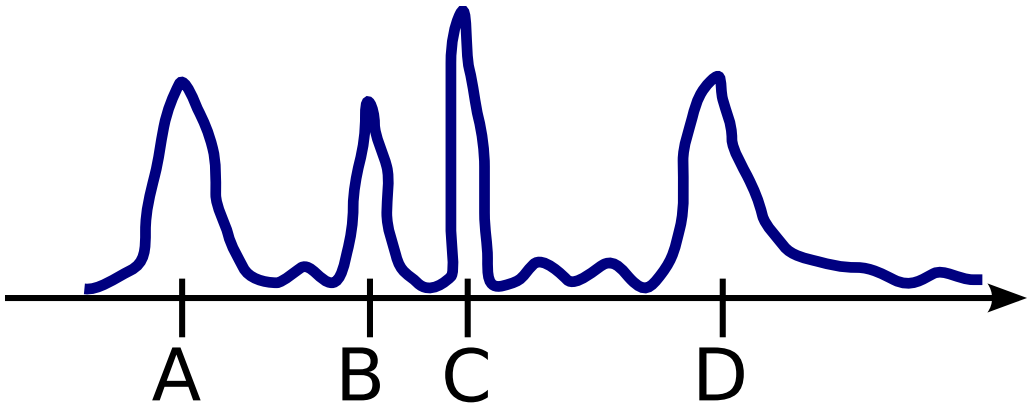
3 warianty: 0, A, B

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



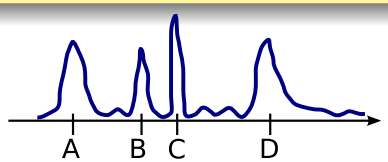
- locus: 4 warianty (A, B, C, D), 6 różnych genotypów
- 2 osobniki w mieszaninie:

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

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

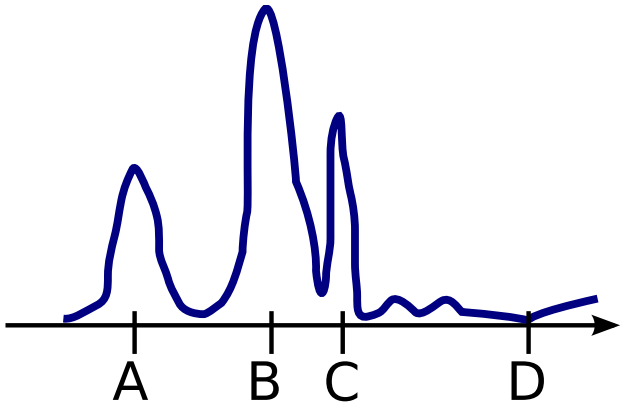
p_a	0.1
p_b	0.2
p_c	0.3
p_d	0.4

osobnik$_1$	osobnik$_2$	$P(o_1 \cap o_2)$
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:?
  $P(E|H_p) = 2p_a p_b$, $P(E|H_d) = P(E)$, $LR = 0.69$

Badanie mieszanin, uwzględnienie ilości materiału



uwzględnienie wysokości (pola powierzchni):

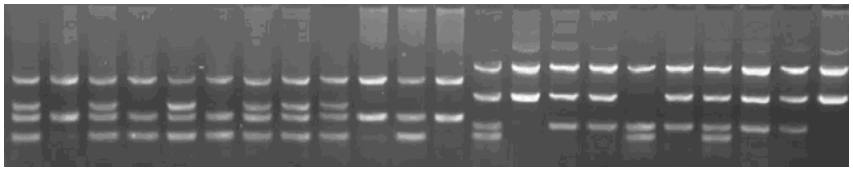
osobnik 1	osobnik 2
A B	B C
A C	B B
B B	A C
B C	A B

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

Analiza haplotypów

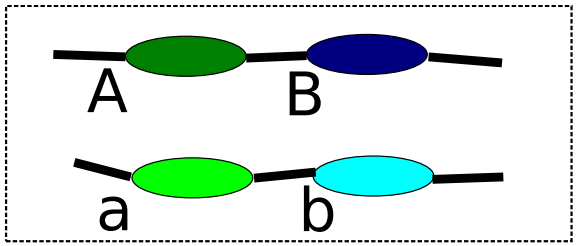
Analiza haplotypów

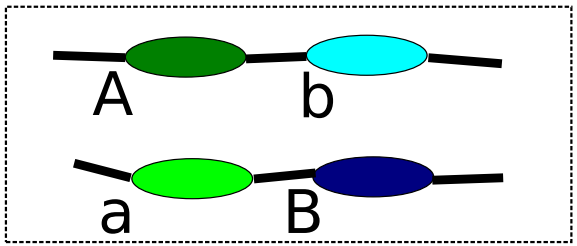
Popularne metody genotypowania (PCR)



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

AaBb ?





Metoda rekonstrukcji [3]

- 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

[3] Clark, 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 $A_2B_1C_1$)

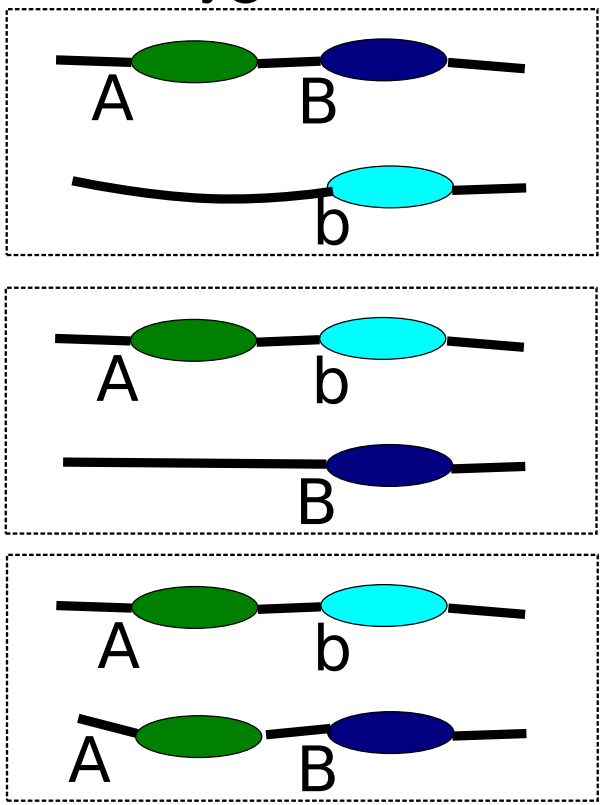
Rekonstrukcja haplotypów - problemy

- 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
$S =< n_1, n_2, ..., n_G >$, $\sum_{j=0}^{G} n_j = n$
Osobnicy są dobieranie niezależnie

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

Szacowanie prawdopodobieństwa haplotypów h_i.

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

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

Analiza złożoności

Liczba układów (par) haplotypów:

$$R = \frac{1}{2}H \ * (H+1), \text{ gdzie } H = \prod_{i=1}^{k} l_i$$

- k liczba analizowanych loci
- l_i liczba wariantów (alleli) dla każdego loci i

Liczba genotypów:

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

Analiza złożoności

Liczba układów haplotypów dla genotypu j:

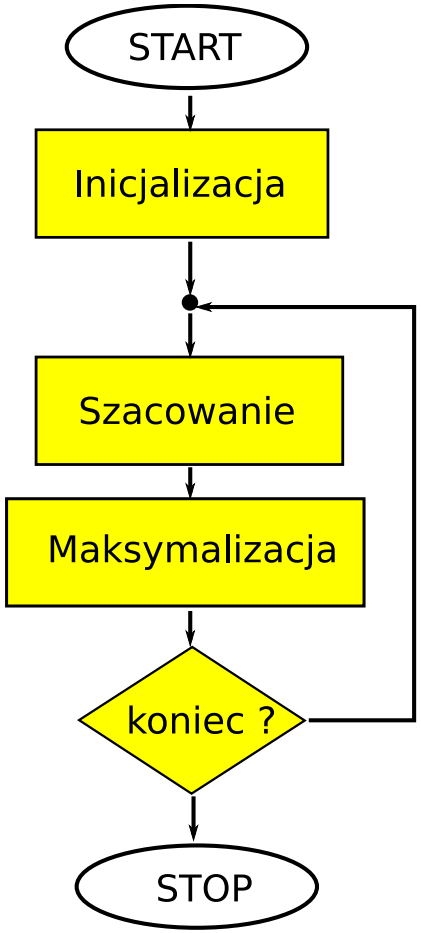
$$r_j = \begin{cases} 2^{s_j-1} * 3^{t_j} & \text{dla } s_j > 0 \\ \frac{3^{t_j}+1}{2} & \text{dla } s_j = 0 \end{cases}$$

- s_j liczba obserwowanych heterozygot
- t_j liczbę obserwowanych wariantów, które nie są nieme dla loci posiadających niemy wariant

Algorytm EM użyty do odtwarzania haplotypów [4]

Algorytm maksymalizacji oczekiwań (algorytm EM)

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



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

Algorytm EM uwzględniając nieme warianty[5]

- inicjacja:

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

- krok M:

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

- krok E:

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

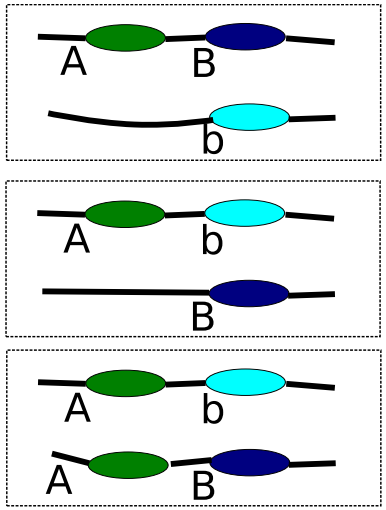
[5] Nowak, Ploski, NullHap - a versatile application to estimate haplotype frequencies from unphased genotypes in the presence of null alleles, 2008

Analiza haplotypów: przykład

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

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 ?



Inicjacja:

$$z_{AB/AB} = 0.5 \quad z_{Ab/Ab} = 0.5 \quad z_{aB/aB} = 1$$
$$z_{AB/Ab} = 0.33 \quad z_{Ab/aB} = 0.33 \quad z_{aB/ab} = 1$$
$$z_{AB/aB} = 0.5 \quad z_{Ab/ab} = 0.5 \quad z_{ab/ab} = 1$$
$$z_{AB/ab} = 0.33$$

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

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

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

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

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

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

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

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

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

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

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

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

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

$$
\begin{aligned}
g_{AB} &= z_{AB/AB} + z_{AB/aB} = 0.12 \\
g_{ABb} &= z_{AB/Ab} + z_{AB/ab} + z_{Ab/aB} = 0.34 \\
g_{Ab} &= z_{Ab/Ab} + z_{Ab/ab} = 0.24 \\
g_{aB} &= z_{aB/aB} = 0.041 \\
g_{aBb} &= z_{aB/ab} = 0.14 \\
g_{ab} &= z_{ab/ab} = 0.12
\end{aligned}
$$

wynik:

$$
\begin{aligned}
h_{AB} &= 0.2 \\
h_{Ab} &= 0.2 \\
h_{aB} &= 0.2 \\
h_{ab} &= 0.4
\end{aligned}
$$

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%

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:

- $\theta(A) = \frac{5+6+3}{30} \approx 0.47$
- $\theta(B) = \frac{8+9}{20} \approx 0.85$

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 $\theta_2(A), \theta_2(B), \theta_3(A), \theta_3(B), \ldots$

Dziękuję