Wykrywanie interakcji między genami przy użyciu metod teorii informacji

Eliza Kaczorek

promotor: dr Paweł Teisseyre

Wydział Matematyki i Nauk Informacyjnych

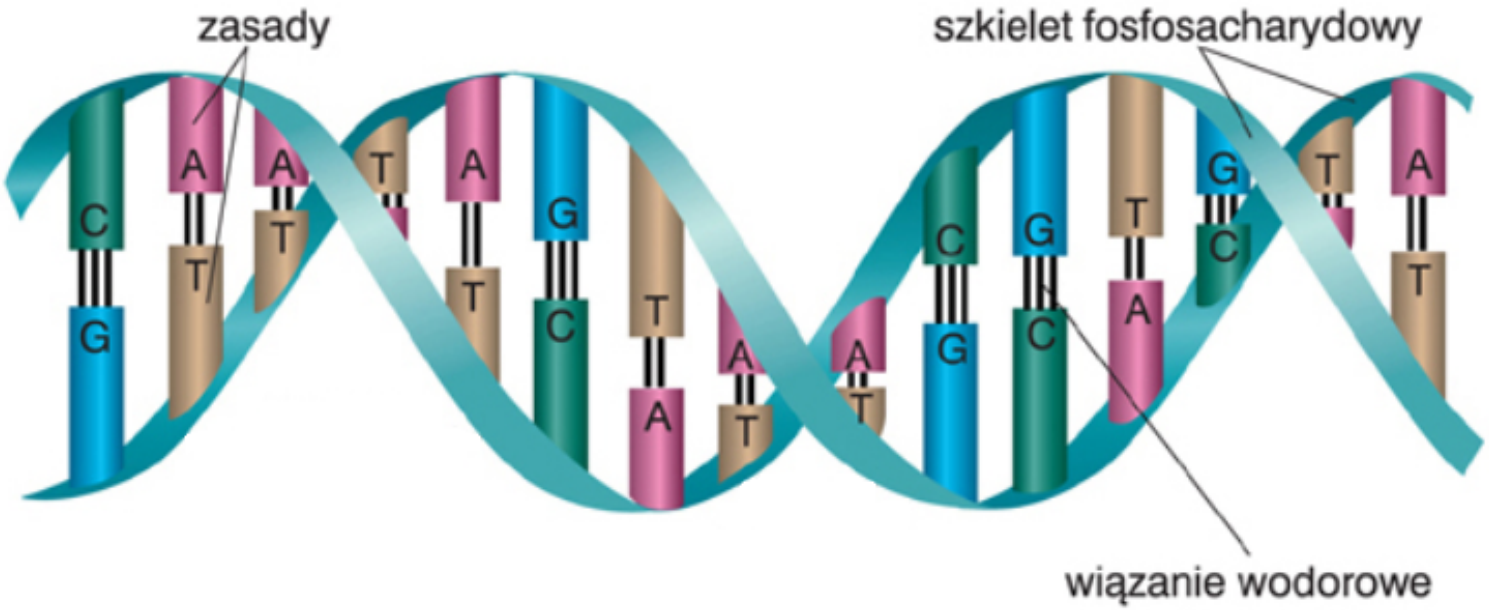
POLITECHNIKA WARSZAWSKA

4 czerwca 2020

Plan prezentacji

1. Budowa DNA, SNP
2. Klasyczne miary
3. Teoria informacji
4. Teorioinformacyjne miary interakcji
5. Eksperymenty
6. Podsumowanie

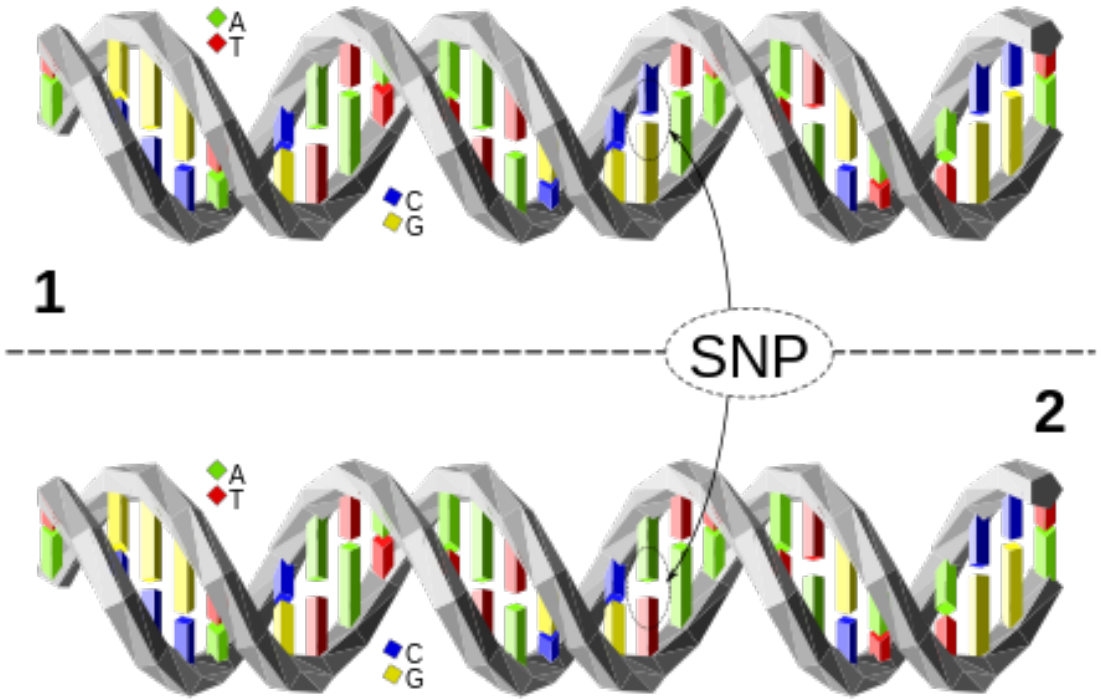
Budowa DNA



Rysunek: Podwójna helisa DNA.[1]

[1] Źródło ilustracji: [7].

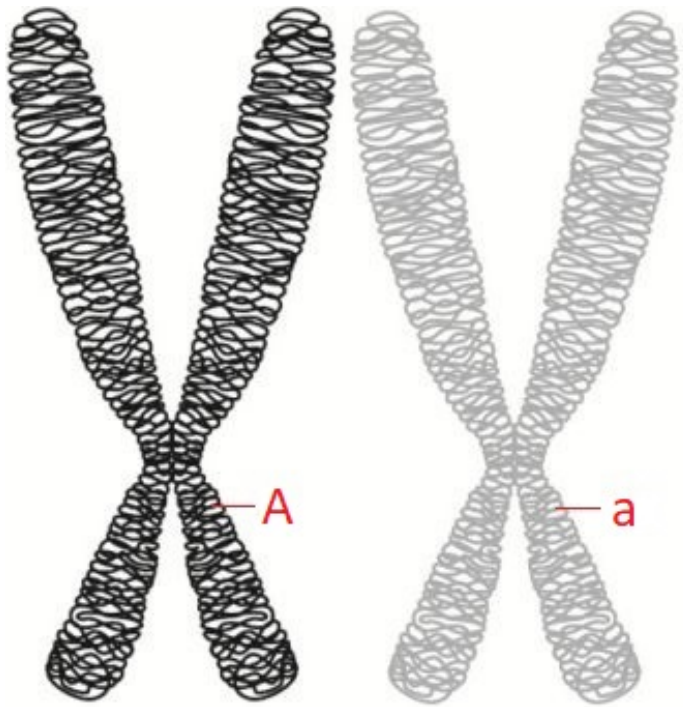
SNP (Single Nucleotide Polymorphism)



Rysunek: Polimorfizm pojedynczego nukleotydu.[2]

[2] Źródło ilustracji: [8].

Para chromosomów



Rysunek: Para chromosomów.[3]

[3] Źródło ilustracji: [9].

Notacja

X_1, X_2 - dyskretne zmienne losowe odpowiadające SNP-om,
$X_1 \in \{AA, Aa, aa\}$, $X_2 \in \{BB, Bb, bb\}$.

Y - dyskretna zmienna losowa opisująca występowanie choroby,
$Y \in \{0, 1\}$.

$p(x_i) := P(X_1 = x_i)$,
$p(x_i, x_j) := P(X_1 = x_i, X_2 = x_j)$,
$p(x_i|y_k) := P(X_1 = x_i|Y = y_k)$.

$X_1 \perp X_2$ - niezależność zmiennych X_1 i X_2.

Regresja logistyczna

Logarytm ilorazu szans dla addytywnego modelu logistycznego (M0):

$$\log\left[\frac{P(Y=1|X_1,X_2)}{P(Y=0|X_1,X_2)}\right] = \mu + \alpha_1 I(X_1 = Aa) + \alpha_2 I(X_1 = aa) + \beta_1 I(X_2 = Bb) + \beta_2 I(X_2 = bb),$$

gdzie $I(A)$ jest indykatorem zbioru A.

Regresja logistyczna

Logarytm ilorazu szans dla modelu logistycznego z interakcjami (M1):

$$\begin{aligned}
&\log\left[\frac{P(Y=1|X_1,X_2)}{P(Y=0|X_1,X_2)}\right] \\
&\quad = \mu + \alpha_1 I(X_1 = Aa) + \alpha_2 I(X_1 = aa) \\
&\qquad + \beta_1 I(X_2 = Bb) + \beta_2 I(X_2 = bb) \\
&\quad + \gamma_{11} I(X_1 = Aa, X_2 = Bb) + \gamma_{12} I(X_1 = Aa, X_2 = bb) \\
&\quad + \gamma_{21} I(X_1 = aa, X_2 = Bb) + \gamma_{22} I(X_1 = aa, X_2 = bb).
\end{aligned}$$

Miarą siły interakcji jest statystyka ilorazu wiarogodności

$$LRT(X_1, X_2, Y) := 2(L_{M1} - L_{M0}),$$

gdzie L_{M0}, L_{M1} są funkcjami log-wiarogodności modeli M0 i M1.

Miara LD

Przykładowa miara pochodząca z całej klasy miar określanych wspólną nazwą *Linkage Disequilibrium* zaproponowana przez Yanga i innych (2009):

$$LD(X_1, X_2, Y) := \sum_{i,j} \frac{\left(\delta_{ij}^{(1)} - \delta_{ij}^{(0)}\right)^2}{p(x_i, x_j)},$$

gdzie $\delta_{ij}^{(1)} = p(x_i, x_j|1) - p(x_i|1)p(x_j|1)$,
$\delta_{ij}^{(0)} = p(x_i, x_j|0) - p(x_i|0)p(x_j|0)$.

Entropia

Entropia zmiennej losowej X_1 o zbiorze wartości $\{x_i\}_{i \in I}$ to

$$H(X_1) := -\sum_i p(x_i) \log p(x_i).$$

Własności

Niech zmienna X_1 przyjmuje N wartości ze zbioru $\{x_1, ...x_N\}$.

- *Wtedy* $0 \leqslant H(X_1) \leqslant \log N$.
- *Jeżeli* $p(x_1) = p(x_2) = ... = p(x_N) = \frac{1}{N}$, *to* $H(X_1)$ *jest maksymalna.*

Informacja wzajemna

Informacją wzajemną nazywamy

$$I(X_1, X_2) := \sum_{i,j} p(x_i, x_j) \log \frac{p(x_i, x_j)}{p(x_i)p(x_j)}.$$

Własności

- $I(X_1, X_2) \geqslant 0$.
- $I(X_1, X_2) = 0$ *wtedy i tylko wtedy, gdy* $X_1 \perp X_2$.

Warunkowa informacja wzajemna

$$I(X_1, X_2|Y) := \sum_{k} p(y_k) \sum_{i,j} p(x_i, x_j|y_k) \log \frac{p(x_i, x_j|y_k)}{p(x_i|y_k)p(x_j|y_k)}.$$

Informacja interakcyjna

Informacją interakcyjną nazywamy

$$II(X_1, X_2, Y) := I\left((X_1, X_2), Y\right) - I(X_1, Y) - I(X_2, Y).$$

Alternatywnie możemy zdefiniować informację interakcyjną jako

$$II(X_1, X_2, Y) = I(X_1, X_2|Y) - I(X_1, X_2).$$

IG_2

W pracy Fana i innych (2011) została zaproponowana następująca miara:

$$IG_2 := I(X_1, X_2|Y = 1) - I(X_1, X_2).$$

Twierdzenie

Zachodzą następujące relacje:

(i) $IG_2 = II(X_1, X_2, Y) \iff I(X_1, X_2|Y = 0) = I(X_1, X_2|Y = 1).$

(ii) $(X_1, X_2) \perp Y \implies IG_2 = 0.$

(iii) *Załóżmy, że* $X_1 \perp X_2$. *Wtedy* $II(X_1, X_2, Y) = 0 \implies IG_2 = 0.$

W pracy skonstruowano przykład, który pokazuje, że przeciwna implikacja do (*iii*) nie zachodzi.

IG_1

Dong i inni (2008) w swojej pracy zaproponowali znormalizowaną miarę interakcji IG_1.

$$IG_1 := \frac{I((X_1, X_2), Y) - \max[I(X_1, Y), I(X_2, Y)]}{\min[H(Y|X_1), H(Y|X_2)]}.$$

Twierdzenie

Zachodzą następujące związki:

(i) *Załóżmy, że* $X_1 \perp Y$ *lub* $X_2 \perp Y$. *Wtedy* $IG_1 = 0 \iff II(X_1, X_2, Y) = 0$.

(ii) $(X_1, X_2) \perp Y \implies IG_1 = 0$.

Odległość interakcyjna *ID*

Ignac i inni (2014) zaproponowali *odległość interakcyjną* jako miarę interakcji:

$$ID := \frac{I(X_1, X_2|Y)}{\max\left[H(X_1|Y), H(X_2|Y)\right]} - \frac{I(X_1, X_2)}{\max\left[H(X_1), H(X_2)\right]}.$$

Twierdzenie

Zachodzą następujące zależności:

(I) $II(X_1, X_2, Y) = 0 \implies ID \geqslant 0.$

(II) $II(X_1, X_2, Y) > 0 \implies ID > 0.$

(III) *Jeżeli zachodzi którykolwiek z warunków*

(i) $X_1 \perp Y$ *i* $X_2 \perp Y$,

(ii) $X_1 \perp X_2$,

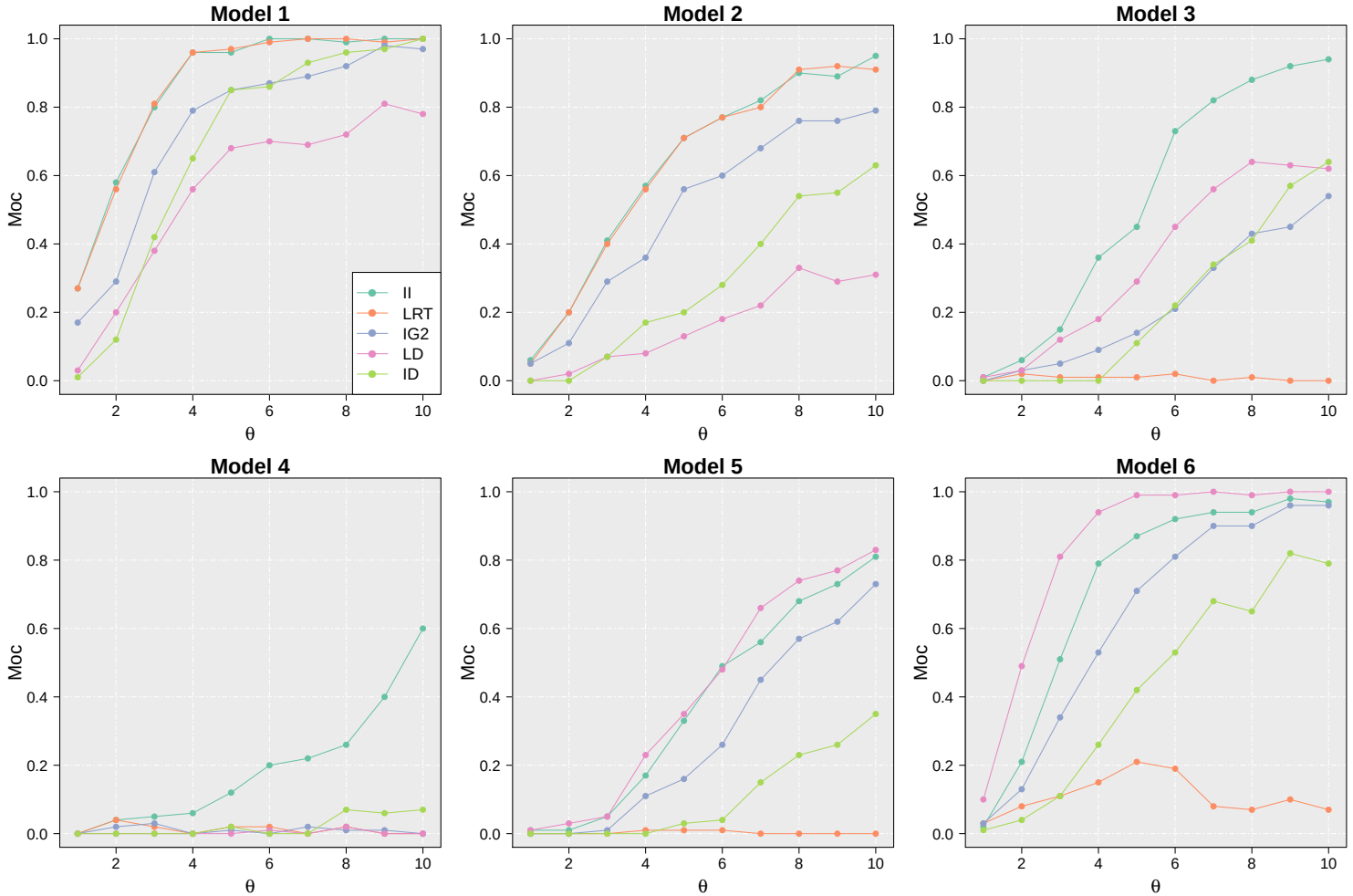
(iii) $X_2 \perp X_2|Y$,

to $\operatorname{sgn}(II(X_1, X_2, Y)) = \operatorname{sgn}(ID)$.

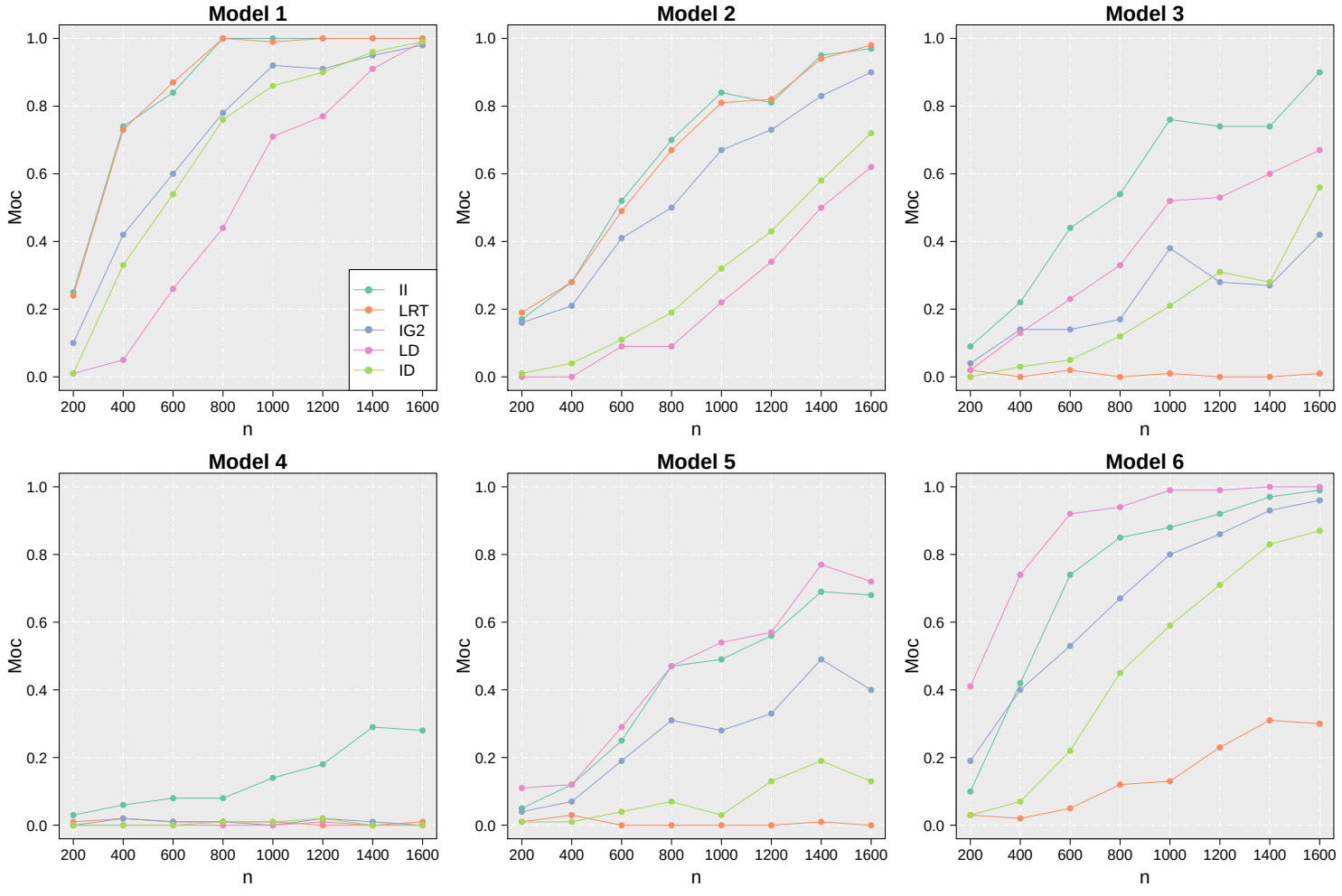
Eksperyment symulacyjny

	Model 1				Model 2			
	bb	*Bb*	*BB*		*bb*	*Bb*	*BB*
aa	$\gamma(1+\theta)^4$	$\gamma(1+\theta)^2$	γ	*aa*	$\gamma(1+\theta)$	$\gamma(1+\theta)$	γ
Aa	$\gamma(1+\theta)^2$	$\gamma(1+\theta)$	γ	*Aa*	$\gamma(1+\theta)$	$\gamma(1+\theta)$	γ
AA	γ	γ	γ	*AA*	γ	γ	γ
	Model 3				Model 4			
	bb	*Bb*	*BB*		*bb*	*Bb*	*BB*
aa	$\gamma(1+\theta)^4$	$\gamma(1+\theta)^3$	$\gamma(1+\theta)^2$	*aa*	$\gamma(1+\theta)^2$	$\gamma(1+\theta)^2$	$\gamma(1+\theta)$
Aa	$\gamma(1+\theta)^3$	$\gamma(1+\theta)^2$	$\gamma(1+\theta)$	*Aa*	$\gamma(1+\theta)^2$	$\gamma(1+\theta)^2$	$\gamma(1+\theta)$
AA	$\gamma(1+\theta)^2$	$\gamma(1+\theta)$	γ	*AA*	$\gamma(1+\theta)$	$\gamma(1+\theta)$	γ
	Model 5				Model 6			
	bb	*Bb*	*BB*		*bb*	*Bb*	*BB*
aa	$\gamma(1+\theta)^3$	$\gamma(1+\theta)^3$	$\gamma(1+\theta)^2$	*aa*	$\gamma(1+\theta)^4$	$\gamma(1+\theta)^3$	$\gamma(1+\theta)^2$
Aa	$\gamma(1+\theta)$	$\gamma(1+\theta)$	γ	*Aa*	$\gamma(1+\theta)^3$	γ	γ
AA	$\gamma(1+\theta)$	$\gamma(1+\theta)$	γ	*AA*	$\gamma(1+\theta)^2$	γ	γ

Tabela: Szanse choroby dla modeli uwzględniających 2 SNP-y.



Rysunek: Moc w zależności od efektu genotypu θ. Liczność próbek $n = 1000$, prewalencja $P(Y = 1) = 0.1$, MAF $q = 0.2$.



Rysunek: Moc w zależności od liczności próbek n. Efekt genotypu $\theta = 6$, prewalencja $P(Y = 1) = 0.1$, MAF $q = 0.2$.

Eksperyment - dane rzeczywiste

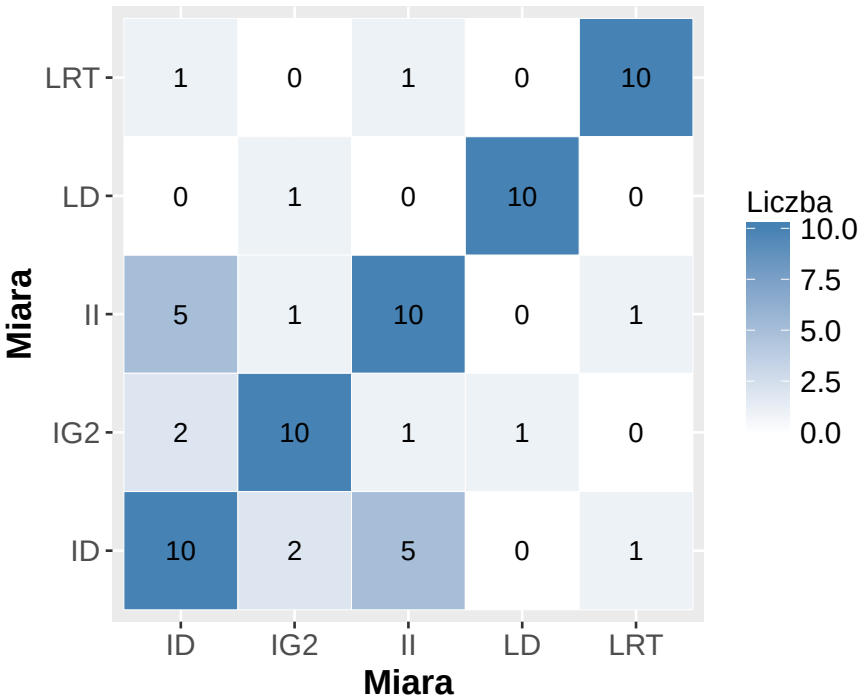
Eksperyment na danych rzeczywistych ma na celu sprawdzenie zgodności działania rozważanych miar interakcji, tzn. zweryfikowania, czy miary te wykrywają te same interakcje.

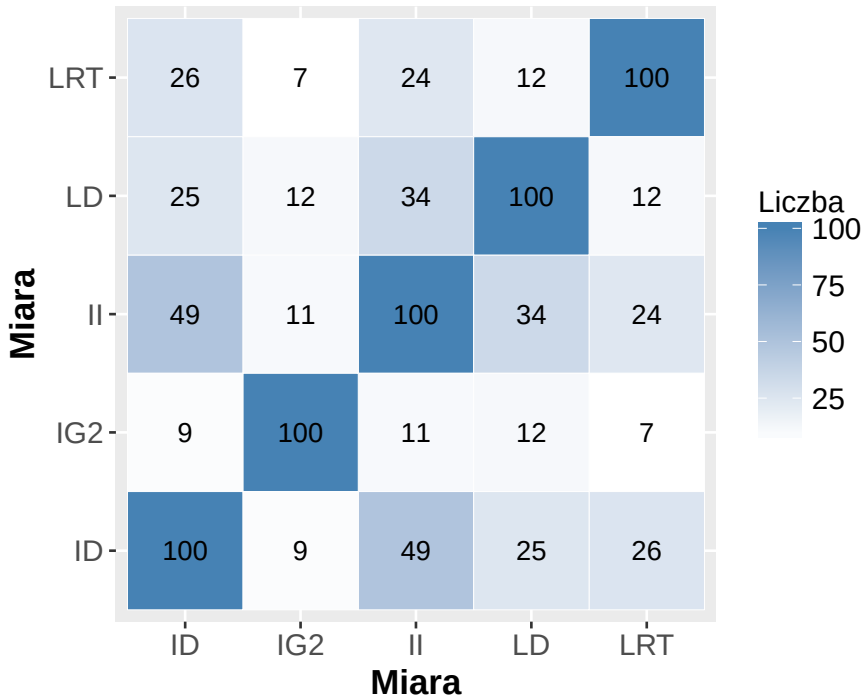
Wykorzystano zbiór dotyczący 208 pacjentów, w tym 121 chorych na raka trzustki i 87 kontrolnych. Zbiór ten zawiera informacje o 901 SNP-ach.

10 pierwszych par SNP-ów z rankingu dla każdej z miar.

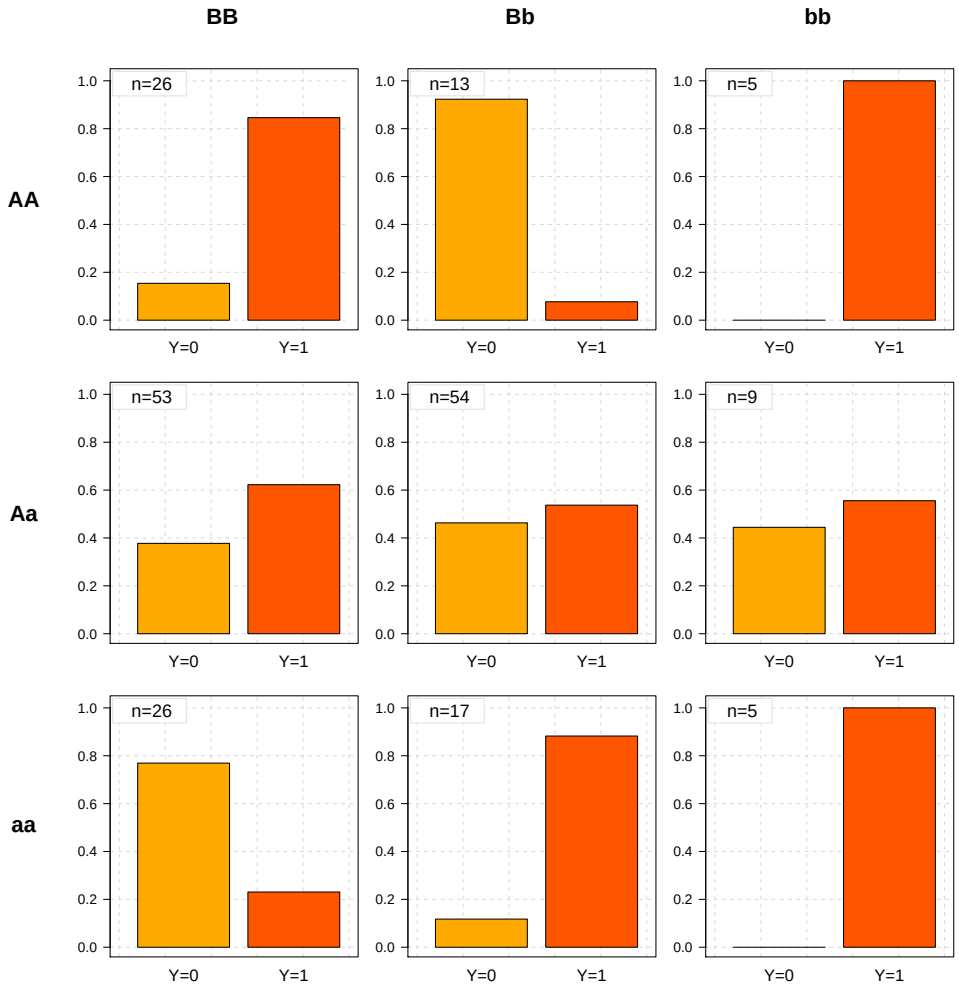
Nr w rankingu	$II(X_1, X_2, Y)$	LRT	IG_2	LD	ID
1	rs1131854 - rs7374	rs2010963 - rs8234	rs2291166 - rs678312	rs8920 - rs9785	rs1048895 - rs3212651
2	rs14804 - rs7201	rs1042506 - rs6115	rs1131854 - rs7374	rs1049612 - rs3217926	rs1131854 - rs7374
3	rs1058213 - rs6115	rs3217926 - rs3771527	rs4761924 - rs6603	rs4761924 - rs6603	rs285162 - rs3750105
4	rs3732253 - rs6115	rs1049612 - rs3771527	rs210135 - rs2272857	rs6108 - rs938	rs13117 - rs2066718
5	rs2010963 - rs8234	rs1061474 - rs2278688	rs2703092 - rs4728164	rs1054204 - rs1059829	rs2010963 - rs8234
6	rs1131854 - rs4252125	rs12727 - rs2227564	rs2703092 - rs4731575	rs3317 - rs448475	rs1058213 - rs6115
7	rs3744262 - rs3764574	rs1042124 - rs1138374	rs3736510 - rs7374	rs1042328 - rs1688029	rs3732253 - rs6115
8	rs3744262 - rs4801200	rs2278688 - rs3798577	rs1295685 - rs797821	rs1304037 - rs17561	rs14804 - rs7201
9	rs3200894 - rs3752095	rs3798577 - rs7224	rs3792215 - rs4252745	rs397768 - rs459552	rs13385 - rs3025053
10	rs1061624 - rs7711	rs3771527 - rs7224	rs571247 - rs613870	rs42427 - rs459552	rs2291166 - rs678312

Liczba tych samych par wyłonionych przez pary miar interakcji



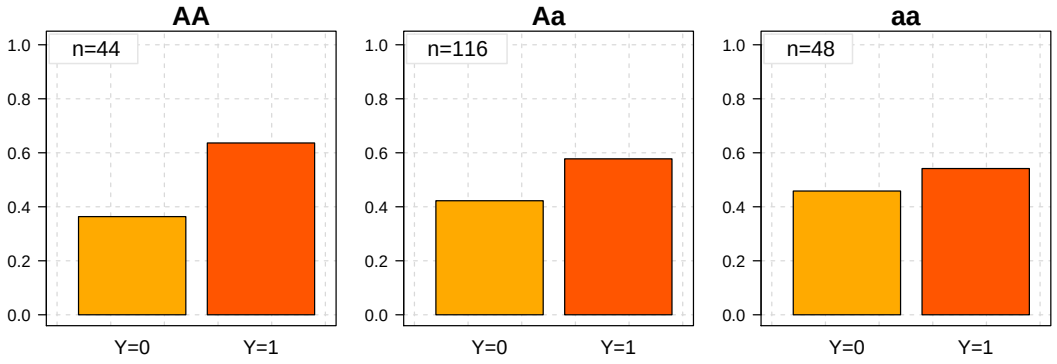


Analiza pierwszej pary w rankingu dla //

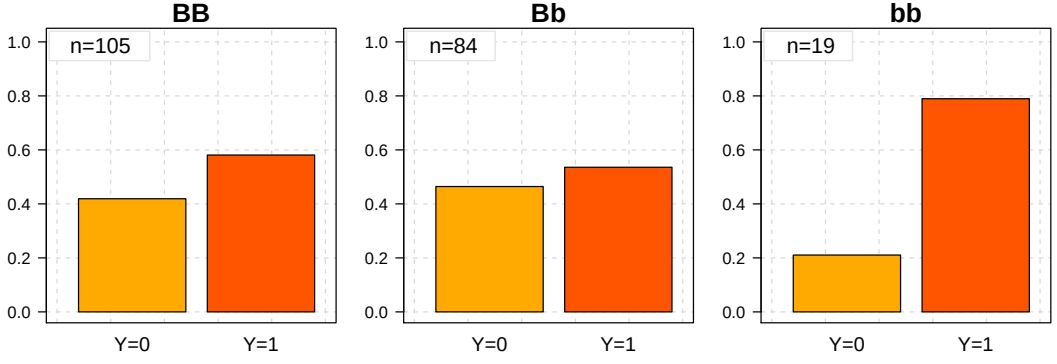


Rysunek: Wykresy rozkładu warunkowego Y pod warunkiem (X_1, X_2), czyli genotypów występujących na tych SNP-ach.

Analiza pierwszej pary w rankingu dla //



Rysunek: Wykresy rozkładu warunkowego Y pod warunkiem X_1.



Rysunek: Wykresy rozkładu warunkowego Y pod warunkiem X_2.

Podsumowanie

- Udowodniono nowe związki informacji interakcyjnej z innymi miarami opartymi na teorii informacji (twierdzenia przedstawione w prezentacji).
- Uznano, że miara IG_1 nie jest właściwą miarą interakcji, ponieważ może wskazywać na jej istnienie w przypadku występowania tylko efektów głównych.
- W eksperymencie symulacyjnym pokazano, że najbardziej skuteczną miarą interakcji jest informacja interakcyjna. Natomiast miara LRT często zupełnie nieefektywna.
- Wykazano małą zgodność w wykrywaniu interakcji między miarami, co potwierdziło różnice teoretyczne.
- Informacja interakcyjna została uznana za najbardziej uniwersalną miarę, wykrywającą różne typy interakcji.

Źródła

[1] Dong C., Chu X., Wang Y., Wang Y., Jin L., Shi T., Huang W., Li Y. (2008). *Exploration of gene-gene interaction effects using entropy-based methods*. European Journal of Human Genetics, 16, 229-235.

[2] Fan R., Zhong M., Wang S., Zhang Y., Andrew A., Karagas M., Chen H., Amos C., Xiong M., Moore J. (2011). *Entropy-based information gain approaches to detect and to characterize gene-gene and gene-environment interaction/corrections od complex diseases*. Genetic Epidemiology, 35, 706-721.

[3] Ignac T. M., Skupin A., Sakhanenko N. A., Galas D. J. (2014). *Discovering Pair-Wise Genetic Interactions: An Information Theory-Based Approach*. PLoS ONE 9(3): e92310.

[4] Mielniczuk J., Teisseyre P. (2018), *A deeper look at two concepts of measuring gene-gene interactions: logistic regression and interaction information revisited*. Genetic Epidemiology, 42, 187-200.

[5] Yang Y., He C., Ott J. (2009). *Testing association with interactions by partitioning chi-squares*. Annals of Human Genetics, 73, 109-117.

[6] `http://snpsyn.biolab.si/examples.html`, 10.04.2018.

[7] `https://encyklopedia.pwn.pl/haslo/deoksyrybonukleinowy-kwas;3891842.html`, 06.03.2018.

[8] `https://pl.wikipedia.org/wiki/Polimorfizm_pojedynczego_nukleotydu`, 25.02.2018.

[9] `https://portal.bioslone.pl/images/portal/etapy%20tworzenia%20biwalentu.jpg`, 21.04.2018.