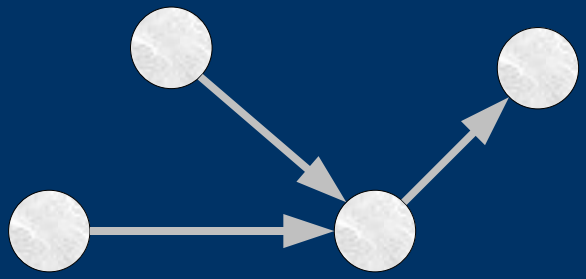
Eksploracja danych mikromacierzowych - sieci Bayesa



Inżynieria Danych, 30 listopada 2009, Tomasz Kułaga

Plan referatu

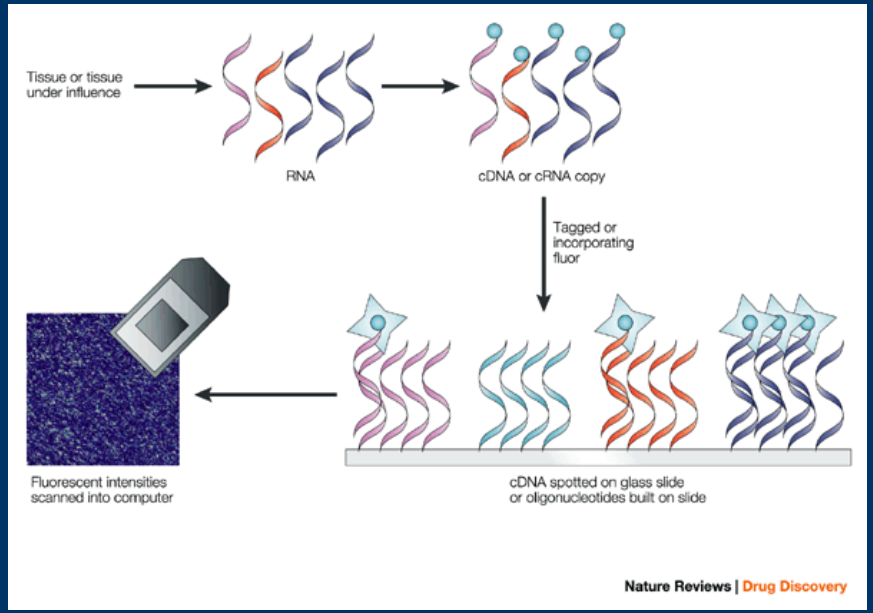
Mikromacierze

Model sieci Bayesa

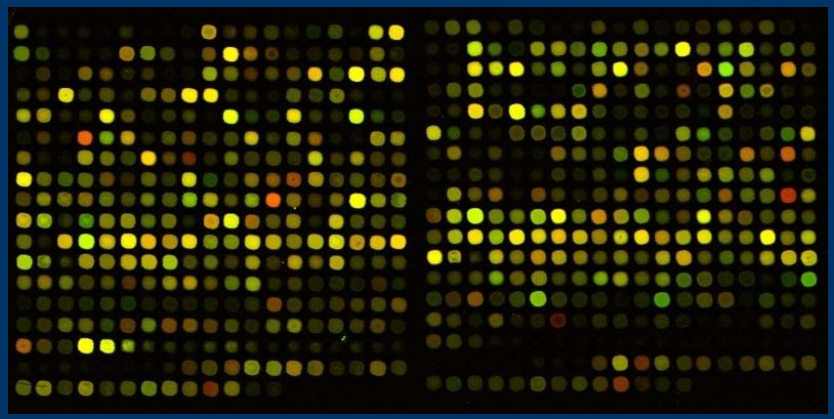
Metody Monte Carlo

Mikromacierz

Płytka z naniesionymi fragmentami sekwencji DNA. Fragmenty te są sondami, które wykrywają przez hybrydyzację komplementarne do siebie cząsteczki DNA lub RNA.



Kawałki te są znakowane, dzięki temu można później laserem lub mikroskopem uzyskać obraz ekspresji genów.



Mikromacierze

Platformy na których przeprowadza się eksperymenty mikromacierzowe mają różne wielkości, od kilku tysięcy do nawet dwóch milionów.

Ilość powtórzonych eksperymentów – próbek dochodzi w porywach do kilkuset (z reguły nie przekracza 50).

Dane te są ogólnodostępne np. przez stronę NCBI http://www.ncbi.nlm.nih.gov/geo/

Mikromacierze – wstępna obróbka

Zanim zacznie się analizę, dane z mikromacierzy trzeba poddać wstępnej obróbce:
- Analiza zeskanowanego obrazu - obliczenie intensywności sygnału dla każdej sondy.
- Odjęcie sygnału tła.
- Normalizacja danych - modyfikacja wartości ekspresji w celu dostosowania do całości eksperymentu lub do zamierzonego rozkładu, niezbędna dla porównywania intensywności między macierzami

Mikromacierze

Załóżmy, że mamy już przetworzone dane mikromacierzowe dla kilku eksperymentów.

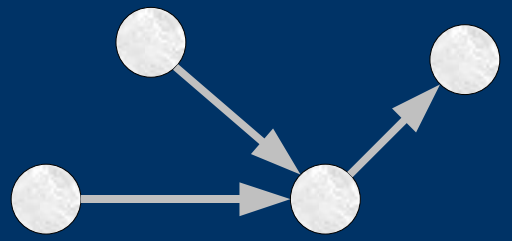
I co dalej?...

Data table	
ID_REF	VALUE
1	0.0843
2	-0.1221
3	0.2185
4	-0.1755
5	-0.2813
6	-0.03873
7	0.1602
8	-0.3796
9	-0.1773
10	0.1878
11	-0.2561
12	0.08678
13	-0.2296

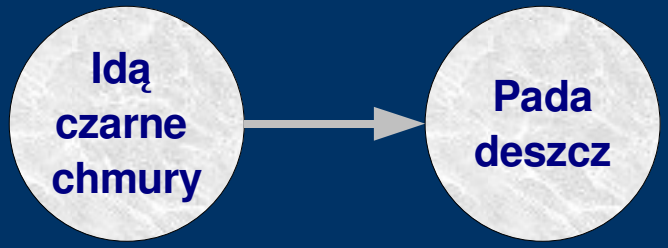
Analiza danych mikromacierzowych

Metody obliczeniowe stosowane dla mikromacierzy pozwalają na takie analizy, jak:
- znajdowanie genów, różniących się ekspresją między próbkami (test T, ANOVA)
- analiza zmian ekspresji w czasie
- klasyfikacja i grupowanie genów ze względu na profil ekspresji w próbkach (klastrowanie, analiza głównych składowych)
- model sieci Bayesa dla obrazowania (dynamicznych lub statycznych) zależności między genami

Pojęcie sieci Bayesa



Przejrzysty model do obrazowania zależności przyczynowo-skutkowych pomiędzy badanymi wielkościami jako (acykliczny) graf skierowany (fakt, że idą czarne chmury ma wpływ na to, czy spadnie deszcz)

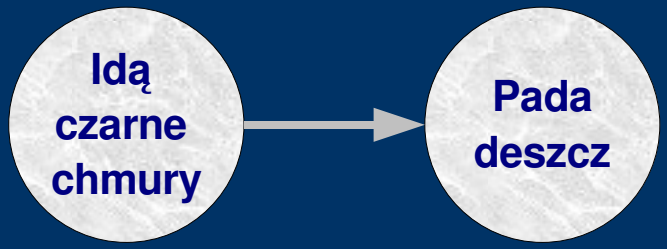


Wierzchołki w grafie to zmienne losowe a krawędzie obrazują zależność warunkową.

W dalszych rozważaniach zawęzimy się do przypadku, gdzie zmienne losowe są dyskretne.

Pojęcie sieci Bayesa

Na pełny opis sieci Bayesa składają się oprócz grafu także rozkłady warunkowe



Jaka jest szansa, że pojawią się czarne chmury?

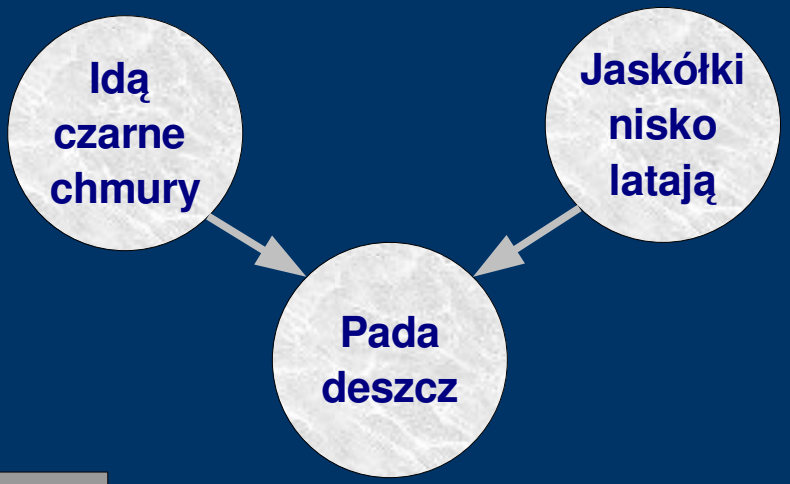
	Prawd
Idą czarne chmury	0.1
Nie idą	0.9

Jakie jest prawdopodobieństwo, że spadnie deszcz, jeśli czarne chmury faktycznie idą lub jeśli ich nie ma?

	Prawd	
Idą czarne chmury	Tak	Nie
Pada deszcz	0.9	0.3
Nie pada	0.1	0.7

Pojęcie sieci Bayesa

Rozkłady warunkowe – bardziej skomplikowany przykład



	Prawd
Idą czarne chmury	0.1
Nie idą	0.9

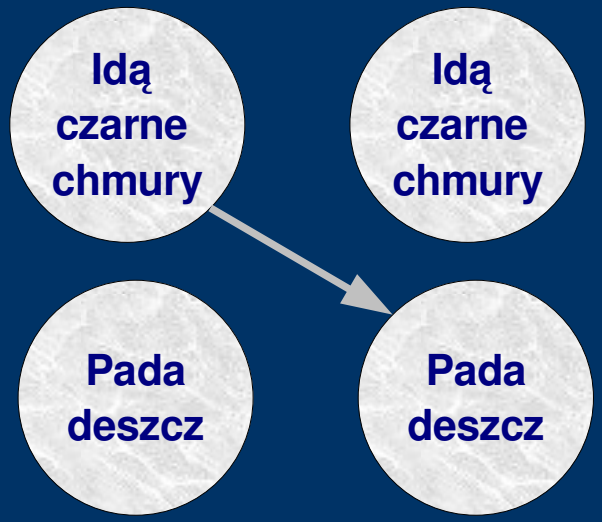
	Prawd
Jaskółki nisko latają	0.05
Nie latają	0.95

	Prawd			
Idą czarne chmury	Tak	Tak	Nie	Nie
Jaskółki nisko latają	Tak	Nie	Tak	Nie
Pada deszcz	0.99	0.7	0.8	0.1
Nie pada	0.01	0.3	0.2	0.9

Dynamiczne sieci Bayesa

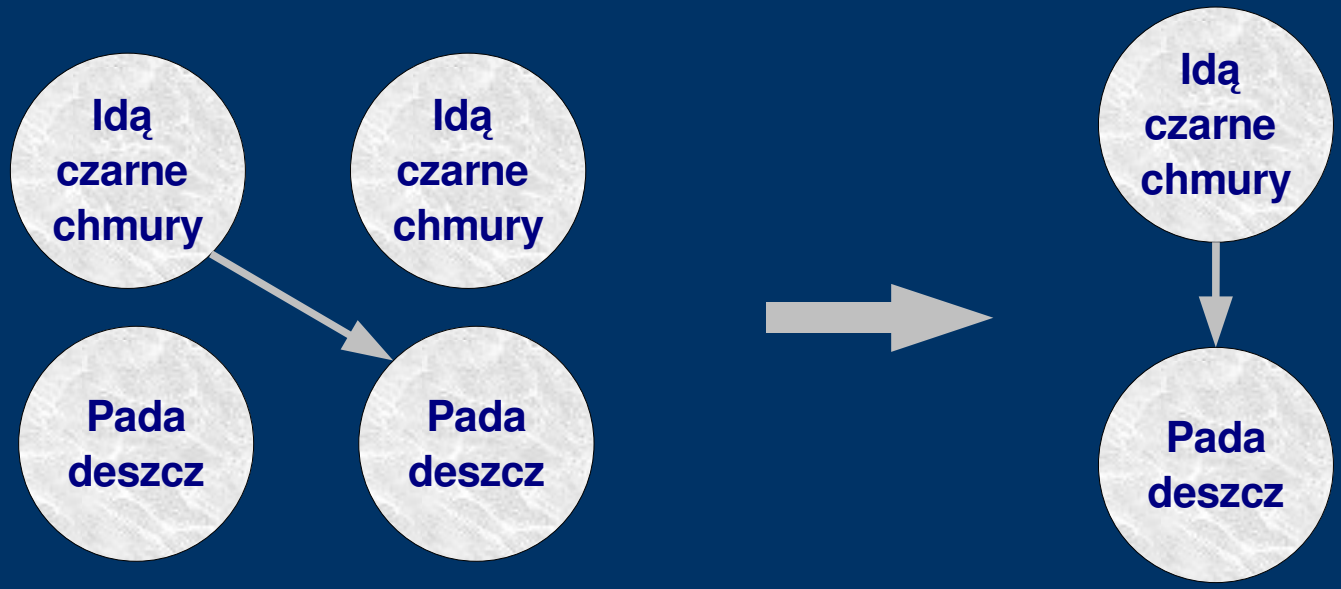
Często zdarza się, że w eksperymentach mikromacierzowych ekspresja genów mierzona jest przez jakiś czas. Np. tuż po podaniu leku jak i 3, 6, 9, 12 i 24 godziny po. Dane tworzą szereg czasowy.

Można to przedstawić używając modelu dynamicznej sieci Bayesa. W najprostszym przypadku interesujemy się tylko zależnościami między dwoma sąsiednimi krokami czasowymi. Zaniedbane zostają połączenia pomiędzy bardziej oddalonymi punktami jak i w obrębie jednego kroku czasowego.



Dynamiczne sieci Bayesa

Aby uprościć ten model wraca się z powrotem do jednego kroku czasowego 'zwijając graf'.



Uwaga: powstały graf nie musi być już acykliczny – zmienna losowa może wpływać na siebie samą w następnym kroku czasowym (samoregulacja genu).

Dynamiczne sieci Bayesa

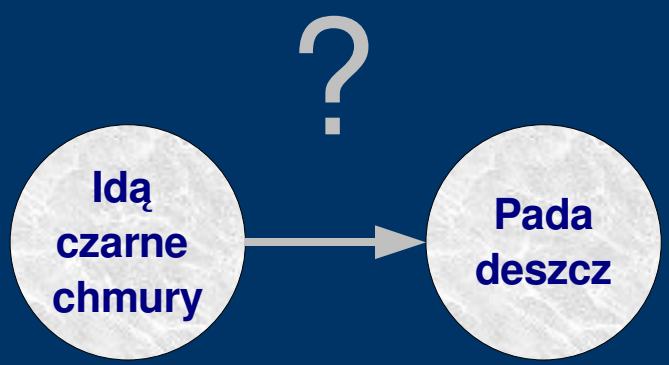
Trzeba pamiętać o pewnych ograniczeniach:
- Pomijamy interakcje wewnątrz kroku czasowego jak i pomiędzy tymi bardziej oddalonymi. Takie założenie może być oczywiście błędne. Można tego nie zakładać, ale złożoność zagadnienia bardzo znacząco wzrośnie – jeszcze do tego wrócimy.
- W najprostszym przypadku zakładamy, że dynamiczna sieć jest tak naprawdę stała (tak samo wygląda ona pomiędzy momentem tuż po podaniu leku i 3 godziny później, jak i pomiędzy 3 a 6 godzin później). Znów można tego nie zakładać i modelować każdą z sieci z osobną, ale skutkuje to zubożeniem ilości obserwacji, którymi dysponujemy.

Jaki model?

Główny problem, to znalezienie odpowiedniego grafu, który jak najlepiej wyjaśniałby rozpatrywane zjawisko. Jest jeszcze kwestia rozkładów warunkowych, ale to drugorzędna sprawa.

Jaki model?

Szukamy najlepszego (najlepszych) modelu, który wyjaśniałby zgromadzone przez nas dane



Idą czarne chmury	Pada deszcz
tak	tak
tak	tak
nie	nie
nie	tak
tak	nie
nie	nie
nie	tak
nie	nie

Jaki model?

Ocena modelu

Czyniąc pewne założenia (niezależność parametrów modelu, rozkład apriori dla parametrów wybrany jako rozkład Dirichleta) i stosując wnioskowanie Bayesowskie dochodzimy do wzoru na prawdopodobieństwo uzyskania naszych obserwacji przy użyciu naszego modelu. W tym wzorze nieznane parametry już nie występują.

$$P(D|M)=\prod_{i=1}^{n}\prod_{\pi_i}\frac{\Gamma(a(\pi_i))}{\Gamma(a(\pi_i)+d(\pi_i))}\prod_{x_i}\frac{\Gamma(a(\pi_i,x_i)+d(\pi_i,x_i))}{\Gamma(a(\pi_i,x_i))}$$

gdzie: D – zbiór obserwacji, M – oceniany model,
– funkcja Gamma-Eulera – dość złożona obliczeniowo,
a – wektor obserwacji apriori, d – wektor faktycznie zanotowanych obserwacji.

Jaki model?

Wyprowadziliśmy więc wzór na P(D|M). Korzystając ze wzoru na prawodopodobieństwo całkowite otrzymujemy

$$P(M|D) = \frac{P(D|M)P(M)}{\Sigma_m P(D|m)P(m)}$$

Pojawia się pewien problem. Żeby policzyć coś takiego, musimy wcześniej policzyć P(D|m) dla wszystkich modeli m.

Pytanie do sali: Ile jest wszystkich modeli? Innymi słowy, ile różnych grafów (acyklicznych) można zbudować dla n wierzchołków?
(n dla mikromacierzy to przynajmniej kilka tysięcy)

Jaki model?

Liczba możliwych modeli, które rozważamy jest super-eksponencjalna, tzn. 2^{n*n}.

To co możemy łatwo zrobić, to porównać dwie oceny modeli, która jest większa i o ile (albo ile razy). Dzięki temu można skupić się na heurystycznych algorytmach przeszukujących, np.:
greedy search,
best-fit search
albo na metodach Monte-Carlo.

Redukcja złożoności

Można próbować redukować złożoność problemu przeprowadzając najpierw np. klasteryzację a dopiero później szukać jakichś zależności między klastrami.

Jednak trzeba być ostrożnym – budując klaster posługujemy się później tak naprawdę jakimś reprezentantem tego klastra (np, średnia wartość, jakiś środkowy 'osobnik').
Jeśli wykażemy jakąś zależność między tymi reprezentantami wcale nie musimy mieć racji w sensie całego klastra.
Część zależności będzie także w oczywisty sposób niedostrzegalna.

Metody Monte Carlo (MCMC)

Stosując metody Monte Carlo można m.in. uzyskać próbkę z rozkładu, który znamy tylko z dokładnością do stałej normującej (całka lub suma nie wynosi 1).

Idea symulacji:
- zaczynamy z pewnego, bliżej nieokreślonego modelu (może być wybrany arbitralnie lub losowo),
- przechodzimy do następnego posługując się macierzą przejścia,
- po odpowiednio długiej symulacji stwierdzamy, że ostatni uzyskany model (lub wszystkie uzyskane do tej pory modele) pochodzi z naszego prawdziwego rozkładu

Metody Monte Carlo (MCMC)

Dwa najpopularniejsze algorytmy (a właściwie jeden):

Próbkowanie Gibbs'a
- Mając ustalony model (graf), wybieramy dwa wierzchołki a i b.
- Liczymy ocenę dwóch modeli – z krawędzią a->b i bez niej.
- Losujemy, który z tych modeli będzie naszym następnym w symulacji z rozkładu dwupunktowego, proporcjonalnego do ocen.

Metropolis-Hastings
- Mając ustalony model, usuwamy lub dodajemy losową krawędź.
- Akceptujemy tą zmianę z pewnym prawdopodobieństwem, które zależy od ocen modelu obecnego i proponowanego

Próbkowanie Gibbs'a jest specjalną odmianą Metropolisa-Hastingsa

Czego szukamy?

Możliwe faktyczne pytania:
- czy aktywność genu A reguluje w jakiś sposób ekspresję genu B?
- jakie geny wpływają na ekspresję genu A?
- które geny mają wpływ na pojawienie się danego schorzenia?
- jaka jest ścieżka aktywności genów? itp...

Dodatkowe korzyści ze stosowania MCMC:
- Łatwiej odpowiedzieć na powyższe pytania dysponując próbką z rozkładu modeli
- Liczymy wartość oczekiwaną odpowiedniej zmiennej losowej (cechy modelu). Np. zliczamy w ilu modelach wystąpiła krawędź od genu A do genu B.

Dodatkowe aspekty - dyskretyzacja

Surowe dane z eksperymentów z mikromacierzami mają charakter ciągły (ekspresja genów).

W prezentowanym podejściu sieci Bayesa zakładają dyskretne wartości węzłów (gen aktywny, nieaktywny, itd... im mniej tym mniejsza złożoność modelu).

Pytanie jak to robić? Jak wybierać progowe wartości dla klasyfikowania genu jako aktywny, nieaktywny itp.

Różne progi będą prowadzić do innych wniosków.

Dodatkowe aspekty – ocena zbieżności algorytmu

Pytanie: W którym momencie uznajemy, że algorytm wyprodukował zadowalającą nas próbkę.

Możliwe podejście – test chi-kwadrat na przynależność do danego rozkładu. Polega on na porównaniu otrzymanych rozkładów dla kilku niezależnych przebiegów algorytmu.

Istotny problem redukcji złożoności – ma to swoje znaczenie dla jakości testu chi-kwadrat. Chodzi o to, żeby nie było modeli, które są bardzo rzadko odwiedzane. Im mniejsza przestrzeń modeli, tym lepiej.

Dodatkowe aspekty – niezależność

Problem zależności wygenerowanej próbki:

Metody MCMC generują próbkę, gdzie kolejno wygenerowane modele są w oczywisty sposób zależne.

Chcemy tego uniknąć, aby uzyskać lepszą próbkę. Wymagane też do testu chi-kwadrat.

Można to uzyskać przez branie co k-tego modelu z wygenerowanej próbki – problem estymacji parametru k (np. szukanie wartości własnych macierzy przejścia -> metody numeryczne)

Dodatkowe aspekty – niezależność

Mając naszą próbkę $M_1,...,M_n$ aproksymujemy macierz przejścia $(P_{i,j})_{i.j=1,...,n}$

$$P_{i,j} = \frac{c_{i,j}}{c_{i,.}}$$

gdzie $c_{i,j}$ - ilość przejść wykonanych z modelu i do modelu j,
$c_{i,.}$ - ilość przejść wykonanych z modelu i do jakiegokolwiek innego.

Mając macierz P liczymy dla niej drugą co do wielkości wartość własną (pierwsza odpowiada rozkładowi stacjonarnemu i jest równa 1). Wartość określa prędkość, z jaką znika zaburzenie wektora własnego odpowiadającego rozkładowi stacjonarnemu.

Jeśli zadowala nas błąd na poziomie 1% to wybieramy co k-ty model, gdzie k spełnia

$^{k} \leq 1\%$.

Eksploracja danych mikromacierzowych - sieci Bayesa

Dziękuję za uwagę

Referencje:

S. El Adlouni, A. C. Favre and B. Bobée (2005). *Comparison of methodologies to assess the convergence of Markov chain Monte Carlo methods*. Computational Statistics & Data Analysis 50 (2006), pages 2685 – 2701.

S. P. Brooks, P. Giudici and A. Philippe (2003). *Nonparametric Convergence Assessment for MCMC Model Selection.* Journal of Computational and Graphical Statistics, volume 12, number 1, pages 1 – 22.

M. K. Cowles and B. P. Carlin (1996). *Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review.* Journal of the American Statistical Association, volume 91, number 434, pages 883-904.

T. Kułaga (2006). *The Markov Blanket Concept in Bayesian Networks and Dynamic Bayesian Networks and Convergence Assessment in Graphical Model Selection Problems.* Praca magisterska w IM UJ.

C. Riggelsen (2005). *MCMC Learning of Bayesian Network Models by Markov Blanket Decomposition*. ECML, pages 329 – 340.

Źródła obrazków:

http://www.nature.com/nrd/journal/v1/n12/images/nrd961-f1.gif

http://en.wikipedia.org/wiki/File:Affymetrix-microarray.jpg