Grafovski diskriminativni probablistički modeli - Bajesovske mreže

prema

Daphne Koller, Nir Friedman

Probabilistic Graphical Models Principles and Techniques

Istraživanje podataka u bioinformatici, 2021/2022.

G. Pavlović-Lažetić

Pregled

- Probabilistički grafovski modeli
- Definicija Bajesovske mreže (usmereni grafovski model) i lokalne nezavisnosti
- Globalne nezavisnosti Bajesovske mreže
- Relacija I-ekvivalentnosti
- Primene Bajesovske mreže u rešavanju biomedicinskih problema – primeri

Probabilistički grafovski modeli

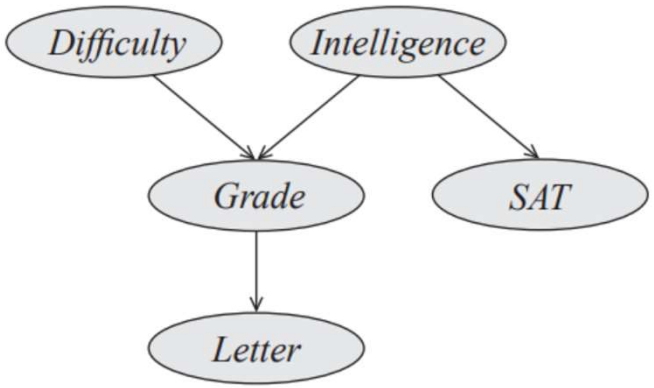
- Za modelovanje kompleksnih sistema (kao što su biološki sistemi) ne može se očekivati da se mogu identifikovati i adekvatno kvantifikovati sve promenljive koje određuju stanje tog sistema
- Potrebno je opise veza između promenljivih definisati u terminima verovatnoće
- *Probabilistički modeli* opisuju odnos između promenljivih njihovom *zajedničkom raspodelom*
- *Probabilistički grafovski modeli* koriste strukturu grafa za kompaktno predstavljanje visoko-dimenzionalnih zajedničkih raspodela

Probabilistički grafovski modeli

- Razmotrimo sledeći primer modelovanja kompanijskog sistema za zapošljavanje diplomiranih studenata (cilj: zaposliti inteligentne kandidate):
  - Model opisuje korelacije između sledećih 5 slučajnih promenljivih: SAT rezultat (Standardizovani test za prijem na američke fakultete), Letter (pismo preporuke, na osnovu ocene iz izabranog predmeta), Grade (ocena iz izabranog predmeta), Difficulty (težina predmeta) i Intelligence (inteligencija)
  - Svaka od slučajnih promenljivih SAT, Letter, Difficulty i Intelligence je binarna slučajna promenljiva (uzimaju vrednosti 1=high i 0=low), dok slučajna promenljiva Grade uzima vrednosti iz skupa {A, B, C}
  - Model treba da definiše zajedničku raspodelu ovih 5 slučajnih promenljivih
  - Dimenzionalnost prostora ove zajedničke raspodele je $2\times2\times3\times2\times2 = 48$
- Da bi se direktno modelovala ova zajednička raspodela, trebalo bi oceniti 48 parametara – po jedan parametar za svaku od mogućih kombinacija vrednosti slučajnih promenljivih

Probabilistički grafovski modeli

- Među nekim slučajnim promenljivim postoje zavisnosti (npr. Grade zavisi od Intelligence i Difficulty)
- Među nekim slučajnim promenljivim ne postoje zavisnosti (npr. SAT ne zavisi od Grade)
- Sledeći graf modeluje sve postojeće zavisnosti:
  - vrednost slučajne promenljive SAT direktno zavisi od vrednosti slučajne promenljive Inteligence
  - vrednost slučajne promenljive Grade direktno zavisi od vrednosti slučajnih promenljivih Difficulty i Inteligence
  - vrednost slučajne promenljive Letter direktno zavisi od vrednosti slučajne promenljive Grade
- Intuitivno značenje da "čvor zavisi direktno samo od svojih roditelja," je u centru semantike Bajesovskih mreža
- Pored direktnih zavisnosti, grafom su modelovane i indirektne zavisnosti između ovih slučajnih promenljivih (npr. vrednost slučajne promenljive Letter zavisi od vrednosti slučajne promenljive Inteligence)



Probabilistički grafovski modeli

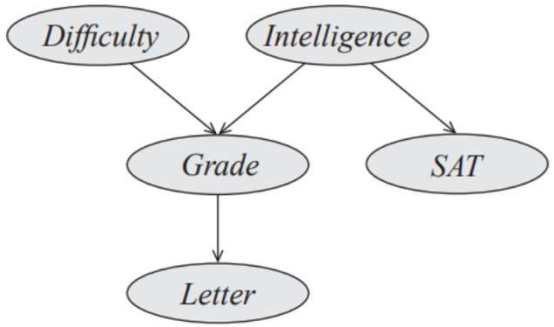
- Grafovska reprezentacija veze između slučajnih promenljivih može da se posmatra i iz perspektive (implicitne) *nezavisnosti* među slučajnim promenljivim
- Svojstva uslovne nezavisnosti označavaju se na sledeći način:
- $(X \perp Y \mid Z)$ ≡ vrednosti slučajnih promenljivih iz skupa X ne zavise od vrednosti slučajnih promenljivih iz skupa Y ukoliko su poznate vrednosti slučajnih promenljivih iz skupa Z (dužina ruku i veština čitanja!)
- Pretpostavke o nezavisnostima između slučajnih promenljivih se mogu iskoristiti za dobijanje jednostavnijeg oblika zajedničke raspodele
- Prema definiciji (uslovne) nezavisnosti slučajnih promenljivih važi
  - $(X \perp Y \mid \emptyset) \equiv P(X \mid Y) = P(X)$
  - $(X \perp Y \mid Z) \equiv P(X \mid Y, Z) = P(X \mid Z)$
- Posledica definicije uslovne verovatnoće $(P(X|Y)=P(X,Y) / P(Y)$ je tzv. pravilo lanca:

$$P(X_k|X_1,...,X_{k-1}) = P(X_1, X_2, ..., X_k) / P(X_1,..., X_{k-1}) \Rightarrow P(X_1, X_2, ..., X_k) = P(X_1,..., X_{k-1}) * P(X_k|X_1,...,X_{k-1})$$

- $$P(X_1, X_2, ..., X_k) = P(X_1) \cdot P(X_2|X_1) \cdot \cdot \cdot P(X_{k-1}|X_1,...,X_{k-2}) \cdot P(X_k|X_1,...,X_{k-1})$$

Probabilistički grafovski modeli

- Prethodni graf određuje sledeći skup *nezavisnosti*:
  - (Inteligence ⊥ Difficulty)
  - (Difficulty ⊥ Inteligence, SAT)
  - (SAT ⊥ Difficulty, Grade, Letter | Inteligence)
  - (Grade ⊥ SAT | Inteligence, Difficulty)
  - (Letter ⊥ Inteligence, Difficulty, SAT | Grade)



- Koristeći pravilo lanca i jednakosti koje slede iz navedenih nezavisnosti slučajnih promenljivih, dobija se sledeći (pojednostavljeni) oblik zajedničke raspodele:
- $P(I, D, G, S, L) = P(I) \cdot P(D|I) \cdot P(G|I,D) \cdot P(S|I,D,G) \cdot P(L|I,D,G,S)$

$= P(I) \cdot P(D) \cdot P(G|I,D) \cdot P(S|I) \cdot P(L|G)$

Probabilistički grafovski modeli

- Koliko je parametara sada potrebno oceniti?
- P(I) – 1 parametar: P(I=0) (P(I=1) = 1–P(I=0))
- P(D) – 1 parametar: P(D=0) (P(D=1) = 1–P(D=0))
- P(G|I, D) – 8 parametra:

  P(G=A|I=0, D=0), P(G=B|I=0, D=0), P(G=A|I=0, D=1), P(G=B|I=0, D=1),

  P(G=A|I=1, D=0), P(G=B|I=1, D=0), P(G=A|I=1, D=1), P(G=B|I=1, D=1)

  (P(G=C|I=0, D=0) = 1 - P(G=A|I=0, D=0) - P(G=B|I=0, D=0),

  P(G=C|I=0, D=1) = 1 - P(G=A|I=0, D=1) - P(G=B|I=0, D=1),

  P(G=C|I=1, D=0) = 1 - P(G=A|I=1, D=0) - P(G=B|I=1, D=0),

  P(G=C|I=1, D=1) = 1 - P(G=A|I=1, D=1) - P(G=B|I=1, D=1))
- P(S|I) – 2 parametra: P(S=0|I=0), P(S=0|I=1) (P(S=1|I=0) = 1 - P(S=0|I=0), P(S=1|I=1) = 1 - P(S=0|I=1))
- P(L|G) – 3 parametra: P(L=0|G=A), P(L=0|G=B), P(L=0|G=C)

  (P(L=1|G=A) = 1–P(L=0|G=A), P(L=1|G=B) = 1–P(L=0|G=B), P(L=1|G=C) = 1–P(L=0|G=C))
- Ukupno 1+1+8+2+3 = 15 parametara, što je znatno manje od 48 parametra ukoliko bismo direktno modelovali zajedničku raspodelu

Probabilistički grafovski modeli

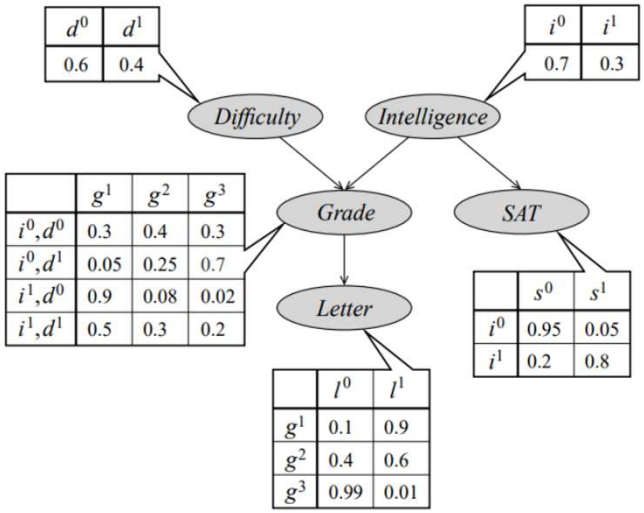
- Probabilistički grafovski modeli koriste određene pretpostavke o *nezavisnosti* slučajnih promenljivih (koje su indukovane *grafovskom reprezentacijom njihovih interakcija*) kako bi predstavili zajedničku raspodelu kao *proizvod uslovnih raspodela*, pri čemu svaka od uslovnih raspodela ima prostor značajno manje dimenzionalnosti u odnosu na prostor zajedničke raspodele
- Dve osnovne familije probabilističkih grafovskih modela su:
  - Bajesovske mreže – koriste usmerene grafove za predstavljanje interakcija između slučajnih promenljivih
  - Markovljeve mreže – koriste neusmerene grafove za predstavljanje interakcija između slučajnih promenljivih
- Ove dve vrste modela se razlikuju po pitanju skupova nezavisnosti koje mogu da predstave i faktorizaciji zajedničke raspodele koju ti skupovi nezavisnosti indukuju

Pregled

- Probabilistički grafovski modeli
- Definicija Bajesovske mreže i lokalne nezavisnosti
- Globalne nezavisnosti Bajesovske mreže
- Relacija I-ekvivalentnosti
- Primene Bajesovske mreže u rešavanju biomedicinskih problema – primeri

Definicija Bajesovske mreže i lokalne nezavisnosti

- Model Bajesovske mreže se sastoji od dve komponente:
  - *usmerenog grafa* čiji čvorovi predstavljaju slučajne promenljive, a grane direktne zavisnosti između slučajnih promenljivih
  - skupa lokalnih verovatnosnih modela koji opisuju prirodu zavisnosti slučajnih promenljivih od njihovih roditelja u grafu



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^1, d^0	0.9	0.08	0.02
i^1, d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

Definicija Bajesovske mreže i lokalne nezavisnosti

- Kako ovako definisan model Bajesovske mreže određuje zajedničku raspodelu?

- Na primer, događaj {I=1, D=0, G=B, S=1, L=0} sastoji se od osnovnih događaja

  {I=1}, {D=0}, {G=B|I=1, D=0}, {S=1|I=1} i {L=0|G=B}

- Koristeći lokalne modele koji su definisani Bajesovskom mrežom i osobinu da je verovatnoća preseka događaja jednaka proizvodu verovatnoća tih događaja

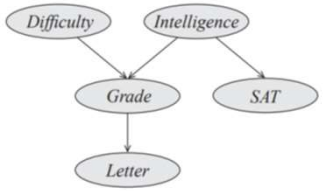
  $P(I=1, D=0, G=B, S=1, L=0) = P(I=1)P(D=0)P(G=B|I=1, D=0)P(S=1|I=1)P(L=0|G=B)$

  $= 0.3 \cdot 0.6 \cdot 0.08 \cdot 0.8 \cdot 0.4 = 0.004608$

- Isti postupak može da se primeni za bilo koji događaj iz prostora zajedničke raspodele

Definicija Bajesovske mreže i lokalne nezavisnosti

- Prema definiciji, struktura Bajesovske mreže je graf koji modeluje *zavisnosti* između slučajnih promenljivih
- Dualna perspektiva: grafom je implicitno određen skup *nezavisnosti* koje važe između slučajnih promenljivih
- Primer:



- Pretpostavka je da profesor formira pismo preporuke samo na osnovu ocene koju je student dobio kod njega na ispitu. Tada, ukoliko je poznata vrednost slučajne promenljive Grade, na ocenu slučajne promenljive Letter ne utiču informacije o vrednostima preostalih slučajnih promenljivih, tj. $(L \perp D, I, S \mid G)$
- Slično, ako znamo da rezultat na SAT ispitu zavisi samo od inteligencije studenta i pri tom je poznata vrednost slučajne promenljive Intelligence, tada na ocenu slučajne promenljive SAT ne utiču informacije o vrednostima preostalih slučajnih promenljivih, tj. $(S \perp D, G, L \mid I)$
- Da li ova zapažanja predstavljaju šablon - da je svaka slučajna promenljiva nezavisna od preostalih slučajnih promenljivih kada su poznate vrednosti njenih roditelja?

Definicija Bajesovske mreže i lokalne nezavisnosti

- Neka su poznate vrednosti slučajnih promenljivih I = 1 i D=1, tj. neka imamo informaciju da je student inteligentan i da je ispit bio težak. Ako bismo pored toga imali informaciju da je student dobio jako pismo preporuke, šanse da je student dobio ocenu A bi trebalo da porastu.
- Slučajna promenljiva Grade nije nezavisna od slučajne promenljive Letter i kada su poznate vrednosti njenih roditelja
- I kada su poznate vrednosti roditelja, slučajna promenljiva i dalje može da zavisi od svojih potomaka
- Od ostalih slučajnih promenljivih zavisnost ne postoji ((G $\perp$ S | I, D), (I $\perp$ D) i (D $\perp$ I, S))

Definicija Bajesovske mreže i lokalne nezavisnosti

- *Obrazac*: Roditeljski čvorovi zaklanjaju verovatnosni uticaj uzročne prirode. Drugim rečima, kada su poznate vrednosti roditelja, nikakve informacije koje se direktno ili indirektno odnose na njihove roditelje ili druge pretke ne utiču na ocenu vrednosti posmatrane slučajne promenljive. Međutim, informacije o potomcima još uvek mogu biti relevantne za predviđanje vrednosti slučajne promenljive.
- *Definicija*: Struktura Bajesovske mreže *G* je usmereni aciklični graf čiji čvorovi predstavljaju slučajne promenljive X1, X2,..., Xn. Neka je Pa_{Xi} skup roditelja čvora Xi, a $NonDescendants_{Xi}$ skup slučajnih promenljivih koje nisu potomci od Xi u grafu *G*. Tada *G* određuje sledeći skup nezavisnosti: $(Xi \perp NonDescendants_{Xi} \mid Pa_{Xi})$ za svako Xi
- Naziva se *skupom lokalnih nezavisnosti* grafa *G* i označava sa $I_l(G)$.
- Lokalne nezavisnosti su osnovna ali ne i jedina vrsta nezavisnosti određena strukturom Bajesovske mreže

Raspodela P se faktoriše pomoću grafa *G* ≡ Raspodela P zadovoljava lokalne nezavisnosti indukovane grafom *G*

- *Definicija:* Neka je *G* struktura Bajesovske mreže nad slučajnim promenljivim X1, X2,..., Xn. Kažemo da se raspodela P definisana na istom skupu slučajnih promenljivih *faktoriše* pomoću grafa *G* ako važi

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid \mathrm{Pa}_{X_i})$$

- Ova jednakost se naziva *pravilo lanca za Bajesovske mreže*. Pojedinačni faktori P(Xi | Pa_{Xi}) nazivaju se *lokalni probabilistički modeli*.
- *Definicija*: Bajesovska mreža je par B = (*G*, P) gde se P faktoriše pomoću *G* i raspodela P je određena skupom lokalnih probabilističkih modela pridruženih grafu *G*.

zajednička raspodela P faktoriše se pomoću grafa *G* ≡ P zadovoljava lokalne nezavisnosti indukovane grafom *G*

Pregled

- Probabilistički grafovski modeli
- Naivni Bajesov klasifikator
- Definicija Bajesovske mreže i lokalne nezavisnosti
- Globalne nezavisnosti Bajesovske mreže
- Relacija I-ekvivalentnosti
- Primene Bajesovske mreže u rešavanju biomedicinskih problema – primeri

Globalne nezavisnosti Bajesovske mreže

- Da li pored lokalnih nezavisnosti postoje još neke nezavisnosti koje zadovoljava svaka raspodela P koja se faktoriše pomoću grafa G?
- Ekvivalentno, da li se još neke nezavisnosti mogu ‘očitati’ direktno sa grafa G?
- Ukoliko postoje takve nezavisnosti, one svakako moraju biti posledica skupa lokalnih nezavisnosti koje indukuje taj graf G.
- Analiza obrnutog problema: kada vrednost slučajne promenljive X *može* da utiče na ocenu slučajne promenljive Y, pri uslovu da je vrednost slučajne promenljive Z poznata?
- Postoje sledeći slučajevi:
  - direktna povezanost - postoji grana između X i Y
  - indikrektna povezanost - postoji staza između X i Y preko čvora Z
  - opšta povezanost - postoji staza između X i Y preko većeg broja čvorova
- *Definicija*: Kažemo da čvorovi X1,..., Xk formiraju *stazu* ukoliko između svaka dva čvora Xi i Xi+1 postoji grana (bilo kojeg usmerenja).

Globalne nezavisnosti Bajesovske mreže

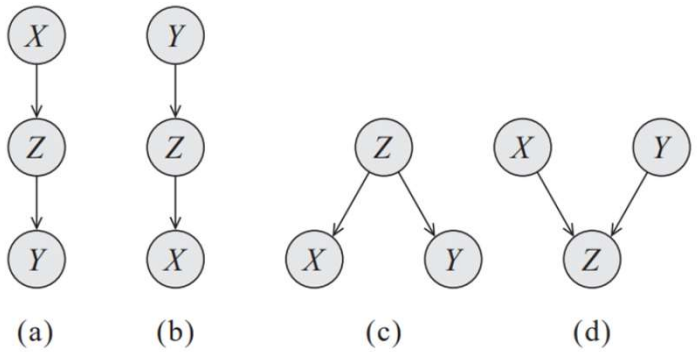
- *Direktna povezanost*

  Za proizvoljnu strukturu grafa G takvu da sadrži granu X→Y ili Y→X uvek postoji raspodela P koja se faktoriše pomoću G i u kojoj su slučajne promenljive X i Y korelisane, bez obzira na ostale slučajne promenljive.

  Stoga, nijedna nezavisnost ($X \perp Y \mid Z$) za bilo koje Z ne može da važi za *sve* raspodele koje se faktorišu pomoću G.

- *Indirektna povezanost*

  Na sledećoj slici su prikazana 4 slučaja povezanosti čvorova X i Y preko čvora Z. Analiziraćemo svaki od njih pojedinačno.



Globalne nezavisnosti Bajesovske mreže

- *Indirektna povezanost – slučaj (a)*

Kako $Z \in Pa_Y$ i $X \in NonDescendants_Y$, na osnovu lokalne nezavisnosti za čvor Y imamo da važi $(Y \perp X \mid Z)$, i simetrično $(X \perp Y \mid Z)$.

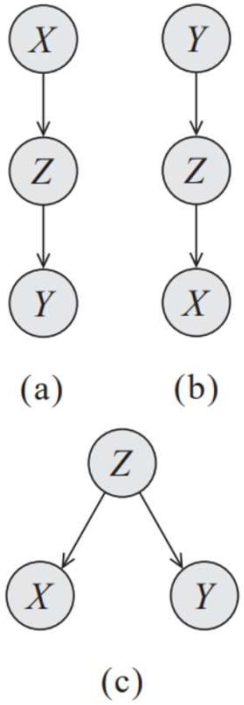
Stoga, nezavisnost $(X \perp Y \mid Z)$ važi za *sve* raspodele koje se faktorišu pomoću G.

- *Indirektna povezanost – slučaj (b)*

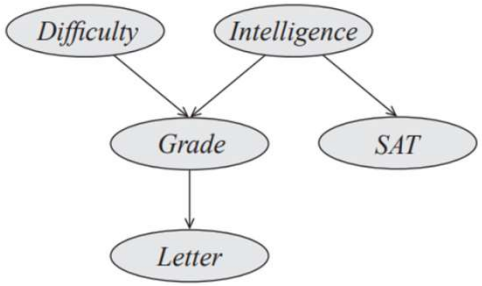
Kako $Z \in Pa_X$ i $Y \in NonDescendants_X$, na osnovu lokalne nezavisnosti za čvor X imamo da važi $(X \perp Y \mid Z)$, i simetrično $(Y \perp X \mid Z)$.

- *Indirektna povezanost – slučaj (c)*

Kako $Z \in Pa_Y$ i $X \in NonDescendants_Y$, na osnovu lokalne nezavisnosti za čvor Y imamo da važi $(Y \perp X \mid Z)$, i simetrično $(X \perp Y \mid Z)$.



Globalne nezavisnosti Bajesovske mreže



- *Indirektna povezanost – slučaj (d)*

- Analiziraćemo ovaj slučaj na prethodnom primeru Bajesovske mreže
  - X = Difficulty, Y = Intelligence, Z = Grade
  - Neka je poznata vrednost slučajne promenljive Grade = C. Ako bismo pored toga imali informaciju da je student inteligentan, šanse da je ispit (na kome je student dobio ocenu C) težak bi trebalo da porastu.
  - Neka nam nije poznata vrednost slučajne promenljive Grade, već njenog potomka Letter. Neka je student dobio slabo pismo preporuke. Slabo pismo preporuke je indikator da je student dobio lošu ocenu na ispitu kod profesora koji piše pismo preporuke, što je dovoljno da slučajne promenljive Difficulty i Intelligence postanu zavisne.
- Dakle, zaključujemo da slučajna promenljiva Difficulty (X) nije nezavisna od slučajne promenljive Intelligence (Y) kada je poznata vrednost slučajne promenljive Grade (Z) ili nekog njenog potomka.

Globalne nezavisnosti Bajesovske mreže

- Formalno, kada uticaj (u smislu zavisnosti) može da ‘teče’ od X do Y preko Z kažemo da je staza X ⇋ Z ⇋ Y **aktivna**. Na osnovu prethodnog razmatranja po slučajevima imamo da je:
  - X → Z → Y je aktivna akko vrednost Z nije poznata
  - X ← Z ← Y je aktivna akko vrednost Z nije poznata
  - X ← Z → Y je aktivna akko vrednost Z nije poznata
  - X → Z ← Y je aktivna akko je vrednost Z ili nekog od njegovih potomaka poznata
- Struktura podgrafa koja odgovara stazi X → Z ← Y naziva se *v-struktura.*

Globalne nezavisnosti Bajesovske mreže

- Opšta povezanost: staza proizvoljne dužine $X = X1 \leftrightharpoons X2 \leftrightharpoons \cdots \leftrightharpoons Xn = Y$
- Da bi uticaj (u smislu zavisnosti) mogao da ‘teče’ od X do Y, potrebno je da svaka od staza $Xi\text{-}1 \leftrightharpoons Xi \leftrightharpoons Xi+1$ bude aktivna. Formalno,
- *Definicija*: Neka je *G* struktura Bajesovske mreže i $X1 \leftrightharpoons X2 \leftrightharpoons \cdots \leftrightharpoons Xn$ staza u *G.* Neka je Z podskup skupa promenljivih sa poznatim vrednostima. Staza $X1 \leftrightharpoons X2 \leftrightharpoons \cdots \leftrightharpoons Xn$ je aktivna ako:
  - Za svaku v-strukturu $Xi\text{-}1 \rightarrow Xi \leftarrow Xi+1$ u okviru staze imamo da je Xi ili neki od njegovih potomaka u skupu Z (poznat)
  - Nijedan od preostalih čvorova na stazi nije u Z
- Ako u grafu postoji više od jedne staze od čvora X do čvora Y, tada da bi uticaj (u smislu zavisnosti) mogao da ‘teče’ od X do Y, potrebno je da barem jedna od tih staza bude aktivna.

Globalne nezavisnosti Bajesovske mreže

- *Definicija*: Neka je *G* struktura Bajesovske mreže i **X**, **Y** i **Z** skupovi čvorova iz *G.* Kažemo da su **X** i **Y** *d-razdvojeni* (directed separation) pomoću **Z** ako ne postoji aktivna staza između bilo kojih čvorova X ∈ **X** i Y ∈ **Y** kada su vrednosti promenljivih iz skupa **Z** date

  Oznaka je $\text{d-sep}_G(X; Y \mid Z)$

- Na osnovu prethodnog, ako su X i Y d-razdvojeni pomoću Z, tada važi $(X \perp Y \mid Z)$
- Skup nezavisnosti koje su indukovane d-razdvajanjima u grafu *G* obeležavamo sa *I(G)* i nazivamo *globalnim skupom nezavisnosti grafa G*

  $I(G) = \{(X \perp Y \mid Z): \text{d-sep}_G(X; Y \mid Z)\}$

- Ispostavlja se da je skup globalnih nezavisnosti upravo traženi skup nezavisnosti koji zadovoljavaju sve raspodele P koje se faktorišu pomoću grafa *G*

Pregled

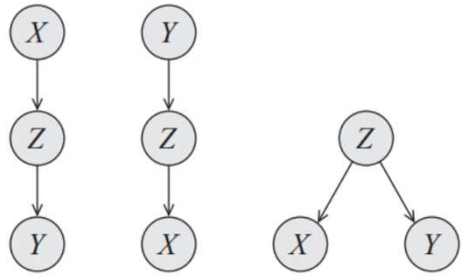
- Probabilistički grafovski modeli
- Naivni Bajesov klasifikator
- Definicija Bajesovske mreže i lokalne nezavisnosti
- Globalne nezavisnosti Bajesovske mreže
- Relacija I-ekvivalentnosti
- Primene Bajesovske mreže u rešavanju biomedicinskih problema – primeri

Relacija I-ekvivalentnosti

- Mada skup globalnih nezavisnosti *I(G)* zadovoljavaju sve raspodele P koje se faktorišu pomoću grafa *G, o*brat ne važi
- Ne postoji, za svaku raspodelu P, graf G takav da važi *I(*G*)* = *I(P)*

(*I(P)* je skup nezavisnosti oblika (X ⊥ Y | Z) koje važe u P)

- Čak i kada postoji, on ne mora biti jedinstven.
- PRIMER: za sva tri grafa sa slike važi *I(*G*)* = { (X ⊥ Y | Z) }



- Definisaćemo relaciju ekvivalencije na skupu grafova koja će nam omogućiti da apstrahujemo pojedinosti vezane za strukturu grafova i gledamo ih samo kao specifikacije nezavisnosti koje indukuju.

Relacija I-ekvivalentnosti

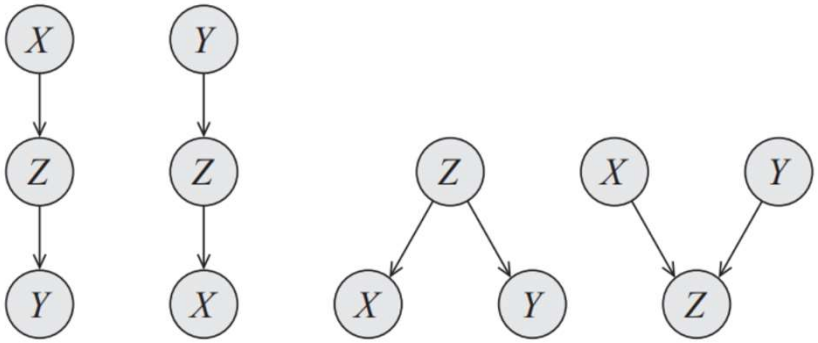
- *Definicija*: Dva grafa G1 i G2 sa skupom čvorova X su *I-ekvivalentna* ako važi

$$I(\mathrm{G1}) = I(\mathrm{G2})$$

- I-ekvivalentnost dva grafa implicira da se svaka raspodela P koja se faktoriše pomoću jednog grafa može faktorisati i pomoću drugog
- *Definicija*: Skelet usmerenog grafa G nad skupom čvorova X je neusmereni graf nad istim skupom čvorova takav da sadrži neusmerenu granu X–Y za svaku usmerenu granu X ⇋ Y ∈ G.
- Primetimo da sva tri grafa iz prethodnog primera imaju isti skelet.
- Ispostavlja se da svaka dva I-ekvivalentna grafa moraju da imaju isti skelet.

Relacija I-ekvivalentnosti

- Da li važi obrnuto – da su grafovi koji imaju isti skelet I-ekvivalentni?
- PRIMER: graf koji odgovara v-strukturi X → Z ← Y ima isti skelet kao i prva tri grafa, ali indukuje skup nezavisnosti *I(*G*)* = { (X ⊥ Y) }



- G1, G2 I-ekvivalentni ⇒ G1, G2 imaju isti skelet
- G1, G2 I-ekvivalentni ⇏ G1, G2 imaju isti skelet

Relacija I-ekvivalentnosti

- *Teorema*: Neka su G1 i G2 grafovi sa skupom čvorova X. Ako imaju isti skelet i skup v-struktura, onda su I-ekvivalentni.
- Obrnuti smer ovog tvrđenja ne važi, tj.

  G1, G2 I-ekvivalentni $\Leftarrow$ G1, G2 imaju isti skelet i skup v-struktura

  G1, G2 I-ekvivalentni $\not\Rightarrow$ G1, G2 imaju isti skelet i skup v-struktura

- Imamo da je uslov istog skeleta neophodan, ali ne i dovoljan da bi dva grafa bila I-ekvivalentna. Sa druge strane, uslov istog skeleta i skupa v-struktura je dovoljan ali ne i neophodan da bi dva grafa bila I-ekvivalentna

Pregled

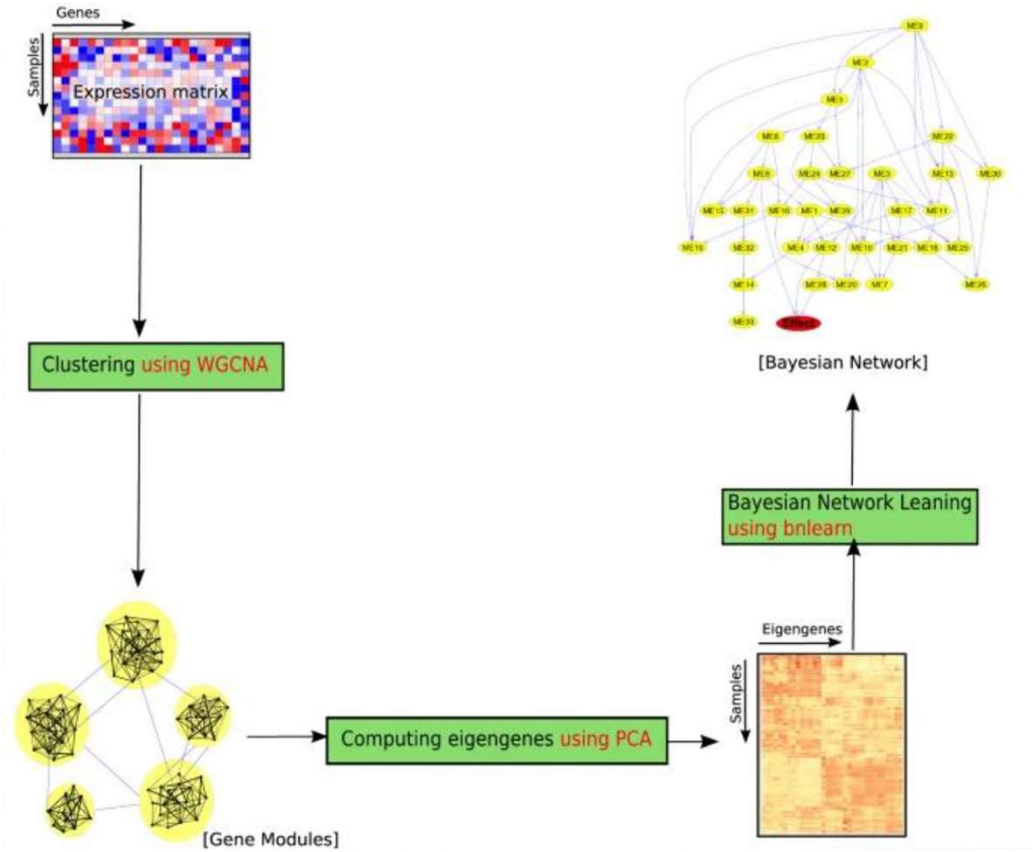
- Probabilistički grafovski modeli
- Naivni Bajesov klasifikator
- Definicija Bajesovske mreže i lokalne nezavisnosti
- Globalne nezavisnosti Bajesovske mreže
- Relacija I-ekvivalentnosti
- Primene Bajesovske mreže u rešavanju biomedicinskih problema – primeri

Primena Bajesovske mreže u rešavanju biomedicinskih problema - primeri

- **Applications of Bayesian network models in predicting types of hematological malignancies**
- Rupesh Agrahari, Amir Foroushani, T. Roderick Docking, Linda Chang, Gerben Duns, Monika Hudoba, Aly Karsan & Habil Zare
- *Sci Rep* 8, 6951 (2018)

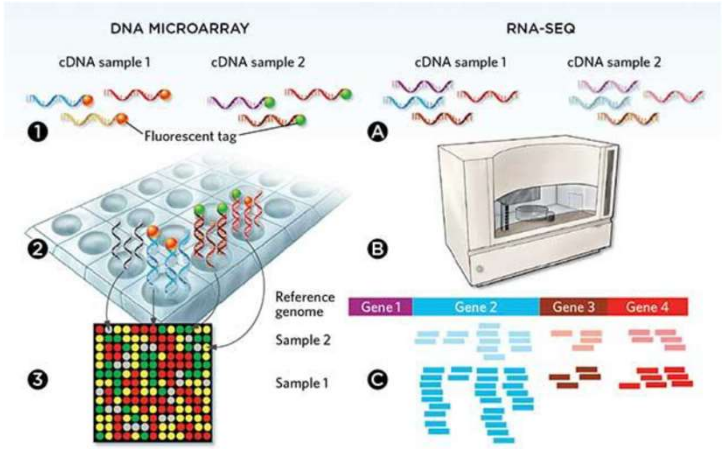
- **Abstract**
- Network analysis is the preferred approach for the detection of subtle but coordinated changes in expression of an interacting and related set of genes
- We introduce a novel method based on the analyses of coexpression networks and Bayesian networks, and we use this new method to classify two types of hematological malignancies:
  - –acute myeloid leukemia (AML)
  - –myelodysplastic syndrome (MDS)

Schematic overview of our methodology



Expression matrix

- Expression matrix is a matrix where each row represents a gene and each column represents a sample. Each entry in the matrix represents the expression level of a particular gene in a given sample (cell).



Coexpression networks

- It is correlation network based on co-expression of genes
  - A gene co-expression network is an undirected graph, where each node corresponds to a gene, and a pair of nodes is connected with an edge if there is a significant co-expression relationship between them (if it passes a set threshold of co-expression strength).
  - There is a number of methods that have been developed for constructing gene co-expression networks. In principle, they all follow a two step approach:
    - calculating co-expression measure
    - selecting significance threshold

Coexpression networks

- The input data for constructing a gene co-expression network often is expression matrix.
- For instance, in a microarray experiment the expression values of ***m*** genes are measured for ***n*** samples. The resulting matrix is an ***m×n*** expression matrix.
- In the first step, a similarity score (co-expression measure) is calculated between each pair of rows in expression matrix. The resulting matrix is a symmetrical ***m×m*** matrix called the *similarity matrix*.
- In the second step, the elements in the similarity matrix which are above a certain threshold (i.e. indicate significant co-expression) are replaced by 1 and the remaining elements are replaced by 0. The resulting matrix, called the *adjacency matrix*, represents the graph of the constructed gene co-expression network.

WGCNA

- WGCNA (Weighted Gene Coexpression Network Analysis) is used to group related genes into gene modules (clusters) based on their coexpression patterns.

- This method constructs a weighted network which means all possible edges appear in the network, but each edge has a weight which shows how significant is the co-expression relationship corresponding to that edge.

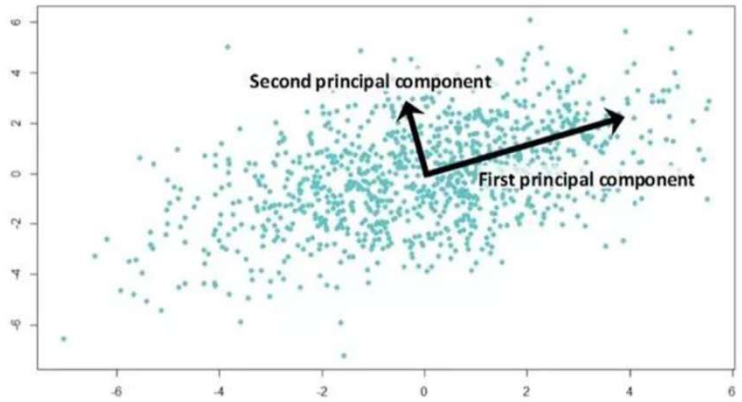
- Next step is to cluster genes into network **modules** using a network proximity measure (metrics, a suitably defined measure of interconnectedness). Thus, a network module is a set of nodes (genes) which are closely connected.

WGCNA - PCA

- For each gene module, WGCNA computes one **eigengene -** a representative of a gene modul.
- In this paper, eigengene of a given module is defined as the first principal component (PCA) of the standardized expression profiles.

PCA

- PCA (Principal Component Analysis) is a dimensionality reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set.



- The first principal component is a direction that maximizes the variance of the projected data.

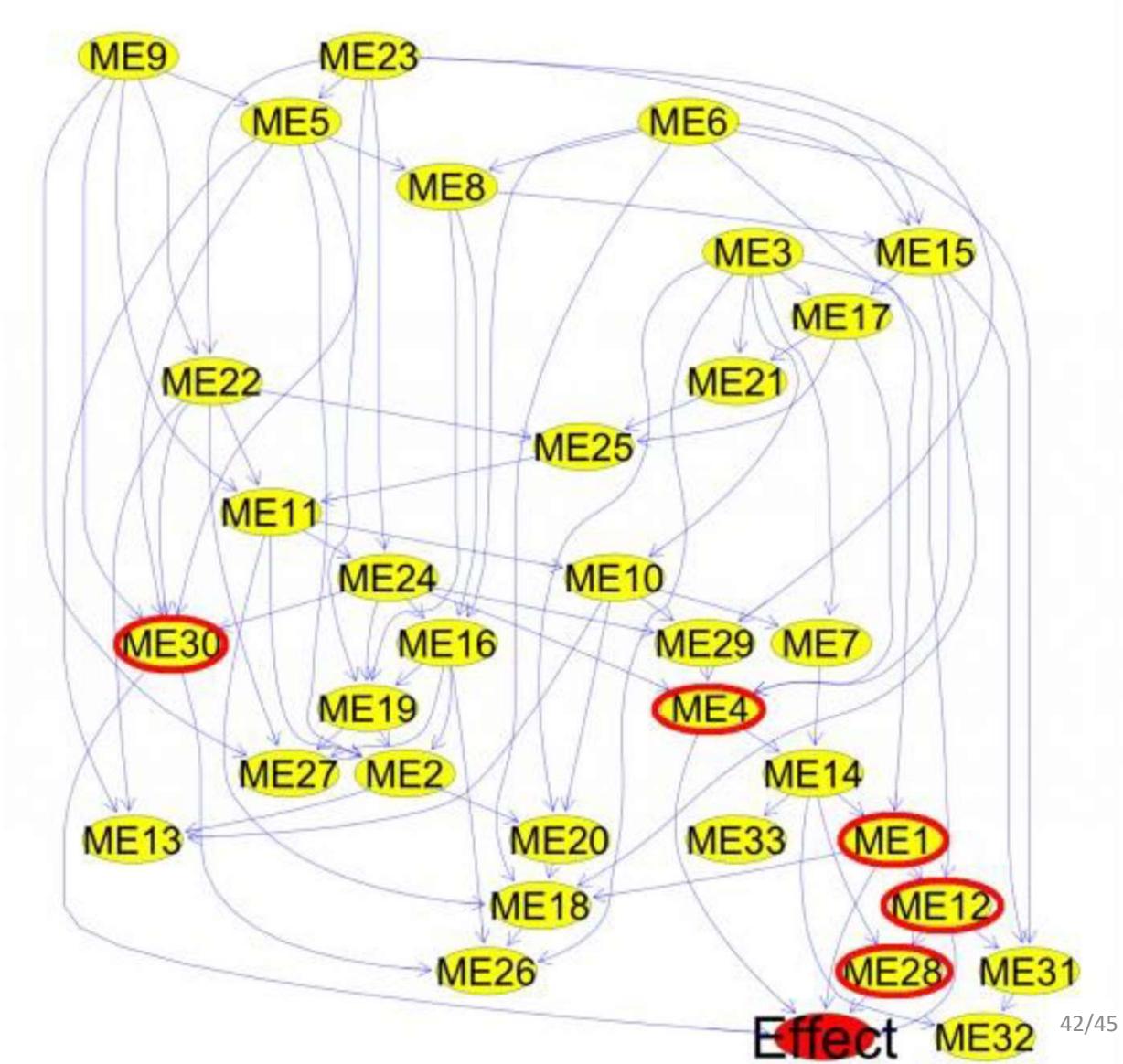
Bayessian network

- WGCNA identified 33 modules i.e. eigengenes
- We used these eigengenes to train a Bayesian network (BN) in which nodes represent gene modules (random variables that models the expression value of an eigengene), and the directed edges represent the probabilistic dependencies between the eigengenes.
- Besides nodes that represent gene moduls, we add ***Effect*** node - a binary variable that models the disease type.

Bayessian network

- There were 5 consensus networks obtained from the top third of 500 fitted networks. They all have fairly similar structures.
- Because of the local Markov property of the BNs, the parents of the *Effect* node are the modules most related to, and predictive of, the disease type
- Nine modules were among the parents of the *Effect* node in at least one BN. The most frequent were Modules 4 and 12. These modules were the parents of the *Effect* node in four BNs. Therefore, they should be enriched with the genes that are associated with AML or MDS.

Bayessian network

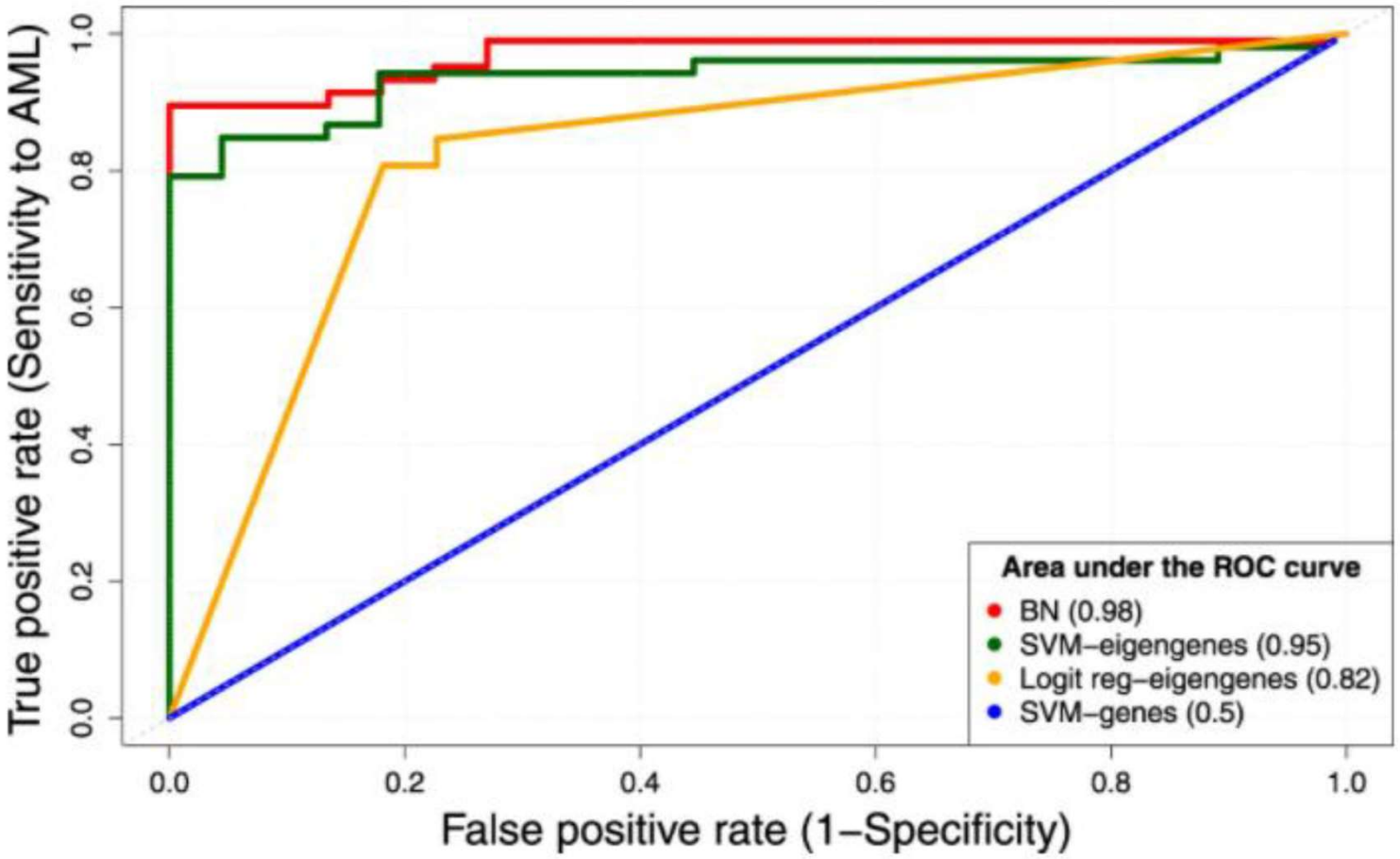


Results

Model	Training partition		
	Accuracy (%)	Precision (%)	Recall (%)
Model 1	96.9	95.7	98.7
Model 2	92.8	89.6	97.3
Model 3	91.5	88.8	87.4
Model 4	88.0	83.1	94.3
Model 5	96.6	96.9	96.9
Mean	**93.2**	**90.8**	**94.9**

Model	Validation partition		
	Accuracy (%)	Precision (%)	Recall (%)
Model 1	78.4	71.8	84.8
Model 2	89.2	84.2	94.1
Model 3	86.5	80.1	94.5
Model 4	94.6	95.2	95.2
Model 5	91.5	95.1	90.1
Mean	**88.0**	**85.2**	**91.8**

Results


True positive rate (Sensitivity to AML)
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate (1−Specificity)
0.0
0.2
0.4
0.6
0.8
1.0
Area under the ROC curve
BN (0.98)
SVM−eigengenes (0.95)
Logit reg−eigengenes (0.82)
SVM−genes (0.5)

Identification of differentially expressed genes in SARS-Cov-2 infected cells using Bayesian network models

Nevena Ćirić[1], Aleksandar Veljković[1]

[1]Faculty of Mathematics, University of Belgrade, Studentski trg 16, 11000 Belgrade, Serbia

BELBI 2021

Introduction

Differential gene expression analysis is the best computational approach for identifying genes whose expressions are altered due to viral infection, such as SARS-CoV-2. Network analysis is the most convenient method for representation of a functionally related set of genes and detection of changes in their expression. This study builds upon the Bayesian network model and coexpression network analysis applied to identification of differentially expressed genes in SARS-CoV-2 infected cells.

Dataset

We used expression data of SARS-CoV-2 mock treated transformed lung alveolar (A549) cells and human lung epithelium cells. Dataset consisting of 21797 genes and 78 samples of MT_A549 cells (SRR11412249) and SI_A549 cells (SRR11412251) was obtained from the study with accession number GSE147507 [2].

The retrieved data was pre-processed and normalized using 70% non-zero expression cutoff, quantile normalization and log2 trainsformation.

Fig. 1: Schematic view of the methodology.

Methodology

Weighted gene coexpression network analysis (*WGCNA*) is used to group related genes into gene modules based on their coexpression patterns [1]. For each gene module, *WGCNA* computes one eigengene – weighted average of the expression of all the genes in that module, whereby weights are determined so that loss in the biological information is minimized.

Eigengenes are used to train a Bayesian network in which nodes (random variables) represent gene modules and directed edges represent the conditional dependencies between corresponding gene modules [1]. Besides random variables that model the expression value of each eigengene, the network has one additional binary variable which models type of sample – infected or non-infected.

Results

WGCNA identified 139 modules. We obtained five Bayesian network models from the subsampling approach – dataset was randomly partitioned into five subsets that were almost equal in size and each model is trained using 4 of the 5 subsets. Each model was tested on the subset that was left over by inferring the value of *Effect* node (Table 1).

According to the Markov property of the Bayesian networks, the parents of that node are the modules most related to, thus they should be enriched with genes that are associated with the disease. 14 modules were among the parents of the *Effect* node in at least one of the constructed models. One of those modules, which contains 74 genes, was parent of the *Effect* node in two models. Further pathway and functional analysis can determine the specific role of these genes in SARS-CoV-2 infection.

References

[1] Agrahari, R., Foroushani, A., Docking, T.R. *et al.* Applications of Bayesian network models in predicting types of hematological malignancies. *Sci Rep* 8, 6951 (2018).

[2] Opeyemi S. Soremekun, Kehinde F. Omolabi, Mahmoud E.S. Soliman, Identification and classification of differentially expressed genes reveal potential molecular signature associated with SARS-CoV-2 infection in lung adenocarcinomal cells, Infomatics in Medicine Unlocked, Volume 20, 2020, 100384, ISSN 2352-9148.

	Training partition		Test partition	
	Accuracy	F1 score	Accuracy	F1 score
Model 1	0.8947	0.8703	0.7262	0.7009
Model 2	0.9355	0.9535	0.6875	0.5455
Model 3	0.9032	0.9189	0.8125	0.8236
Model 4	0.9487	0.9608	0.8621	0.7946
Model 5	0.8889	0.8955	0.2000	0.2500
Mean	0.9142	0.9198	0.6577	0.6229

Table 1: Classification accuracy and F1 score in predicting the value of *Effect* node made by individual models on train and test partitions.