

# Bayesianische Netzwerke I

Christiane Belitz

2.5.2003

# Überblick

- Der Vortrag basiert auf “Using Bayesian Networks to Analyze Expression Data“ von Friedman et al. (2000)
- Definition: Bayesianisches Netzwerk
- Anwendung bei Genexpressionsdaten
- Anwendungsbeispiel: Zellzyklus-Expressionsdaten
- Ausblick

## Definition: Bayesianische Netzwerke

- Sie sind eine Möglichkeit, die gemeinsame Verteilung der Variablen  $X_1, \dots, X_n$  darzustellen
- Sie bestehen aus zwei Komponenten:
  1. einem gerichteten azyklischen Graph (=DAG)  $G$
  2. bedingten Verteilungen der einzelnen Variablen  $X_i$

# 1. Gerichteter azyklischer Graph

- Besteht aus einer Menge von Knoten  $X_1, \dots, X_n$ , die durch Pfeile miteinander verbunden sind
- 'Gerichtet' bedeutet, dass die Verbindung zwischen zwei Variablen eine Richtung besitzt:

Eltern  $\longrightarrow$  Kinder

- 'Azyklisch' bedeutet, daß es keine Verbindungen zurück gibt, d.h. es gibt keine Variablen  $A_1, \dots, A_n$ , die miteinander verbunden sind und für die gilt:  $A_1 = A_n$

## Beispiel

E: Einbruch (ja/nein)

A: Erdbeben (ja/nein)

B: Alarm (ja/nein)

D: Radiobericht über Erd-  
beben (ja/nein)

C: Herr Müller hört den  
Alarm (ja/nein)

## Weitere wichtige Eigenschaften

- $G$  stellt bedingte Unabhängigkeits-Annahmen dar
- Grundlage bildet die Markov-Bedingung:

Jede Variable  $X_i$  ist unabhängig von ihren Nicht-Kindern gegeben ihre Eltern aus  $G$ .

- Definition: Bedingte Unabhängigkeit

$X$  ist bedingt unabhängig von  $Y$  gegeben  $Z$ , wenn gilt:  
 $P(X, Y|Z) = P(X|Z) \cdot P(Y|Z)$  oder  $P(X|Y, Z) = P(X|Z)$ .

- Dann gilt die Kettenregel:

$$P(X_1, \dots, X_n) = \prod_{i=1}^n P(X_i | Pa(X_i)),$$

wobei  $Pa(X_i) = \text{Eltern von } X_i \text{ aus } G$

## Beispiel

Unabhängigkeits-  
Strukturen:

- $I(A, E)$
- $I(E, (A, D))$
- $I(D, (E, B, C)|A)$
- $I(B, D|(E, A))$
- $I(C, (D, E, A)|B)$

Produktform der gemeinsamen Verteilung:

$$P(B, E, A, C, D) = P(A) \cdot P(E) \cdot P(D|A) \cdot P(B|E, A) \cdot P(C|B)$$

## 2. Bedingte Verteilungen

- Müssen für alle Variablen  $X_1, \dots, X_n$  spezifiziert werden
- Unterschiedliche Darstellungsformen je nach Variablentyp:  
diskret, stetig oder Mischung aus beiden Typen

## Diskrete Variablen

- Die bedingten Verteilungen sind Multinomialverteilungen
- Darstellung in Form von Tabellen
- Die Anzahl freier Parameter ist exponentiell zur Anzahl der Eltern

## Stetige Variablen

- Meistens Wahl der bedingten Normalverteilung
- Der Erwartungswert hängt im einfachsten Fall linear von den Werten der Eltern ab
- Die Varianz ist von den Eltern unabhängig

- Die bedingten Verteilungen haben die Form:

$$P(X|u_1, \dots, u_k) \sim N(a_0 + \sum_{i=1}^k a_i u_i, \sigma^2)$$

mit  $\{U_1, \dots, U_k\} = Pa(X_i)$

- Die gemeinsame Verteilung von  $X_1, \dots, X_n$  ist dann eine multivariate Normalverteilung

## Mischung aus beiden Variablentypen

- Nur der Fall stetiger Variablen mit diskreten Eltern ist erlaubt
- Zu jeder Wertekombination der diskreten Eltern wird eine bedingte Normalverteilung, gegeben die stetigen Eltern, bestimmt  
→ konditionale Gauß-Verteilung

# Äquivalenzklassen

- Graph  $G$  enthält eine Menge von Unabhängigkeits-Annahmen  $Ind(G)$
- Es gibt mehrere Graphen, die genau dieselben Unabhängigkeits-Annahmen erfüllen
- Beispiel: Zwei Variablen  $X$  und  $Y$  mit  $Ind(G) = \emptyset$

paßt zu  $X \longrightarrow Y$  und  $X \longleftarrow Y$

- Definition: Zwei Graphen  $G$  und  $G'$  sind genau dann äquivalent, wenn  $Ind(G) = Ind(G')$
- Äquivalente Graphen besitzen denselben zugrundeliegenden ungerichteten Graphen
- Sie unterscheiden sich in der Richtung mancher Kanten

- Sie besitzen dieselben v-Strukturen, d.h. hier ist die Richtung festgelegt

$$(v\text{-Struktur: } X \longrightarrow Y \longleftarrow Z)$$

- Alle äquivalenten Graphen bilden zusammen eine Äquivalenzklasse
- Äquivalenzklassen werden durch einen teilweise gerichteten Graphen (*PDAG*) eindeutig festgelegt

# Anwendung bei Genexpressionsdaten

Ziel: Herausfinden, welche Gene sich gegenseitig beeinflussen

- Mit Bayesianischen Netzwerken lassen sich Prozesse gut beschreiben, bei denen jede Komponente direkt von einer relativ geringen Zahl anderer Komponenten abhängt  
→ trifft auf Genexpressionsdaten zu
- Bayesianische Netzwerke liefern ein Modell für mögliche kausale Zusammenhänge

## Durchführung

- Die Expressionsniveaus der Gene werden durch jeweils eine Zufallsvariable dargestellt
- Andere Zufallsvariable z.B. für experimentelle Bedingungen oder Zellstadium können hinzugenommen werden
- Damit soll ein Bayesianisches Netzwerk geschätzt werden

# Probleme

- Genexpressionsdaten enthalten ein paar Tausend Variablen
- Datensätze bestehen nur aus ein paar Dutzend Beobachtungen

$\implies$  Schätzungen werden unsicher

## Lösung

- Schätzen mehrerer Äquivalenzklassen von Netzwerken mit Hilfe des Bootstrap-Verfahrens
- Beschränken auf charakteristische Eigenschaften  $f$ , die in den meisten Äquivalenzklassen vorkommen

# Bootstrap-Verfahren

- Es werden  $m = 200$  neue Datensätze  $D_i$  gebildet
- Dies geschieht durch das Ziehen mit Zurücklegen von  $N$  Beobachtungen aus dem ursprünglichen Datensatz
- Aus  $D_i$  ( $i = 1, \dots, m$ ) wird eine Netzwerkstruktur  $G_i$  gelernt
- Für alle interessierenden Eigenschaften  $f$  werden Konfidenzwerte

$$\text{conf}(f) = \frac{1}{m} \sum_{i=1}^m f(G_i) \text{ mit } f(G_i) = \begin{cases} 1, & f \in G_i \\ 0, & f \notin G_i \end{cases}$$

berechnet

## Charakteristische Merkmale

- Beschränkung auf Merkmale, die jeweils nur zwei Variablen betreffen
- zwei Arten von Merkmalen:
  1. Markov-Relationen
  2. Ordnungs-Relationen

# 1. Markov-Relation

- Betrifft direkte Nachbarschaft im Modell
- Gibt an, ob die Variable  $Y$  in der Markov-Umgebung von  $X$  liegt
- Zur Markov-Umgebung von  $X$  gehören alle Variablen, die mit  $X$  direkt verbunden sind oder die mit  $X$  gemeinsame Kinder besitzen
- Die Beziehung ist symmetrisch
- Variable  $X$  gegeben ihre Markov-Umgebung ist unabhängig von allen anderen Variablen im Modell

## 2. Ordnungs-Relation

- Gibt an, ob Variable  $X$  in der gelernten Äquivalenzklasse ein Vorfahr von  $Y$  ist
- Ist ein Hinweis darauf, dass  $X$  einen Einfluss auf  $Y$  besitzt

# Lokale Wahrscheinlichkeitsmodelle

Für die Wahl der bedingten Wahrscheinlichkeitsverteilungen gibt es folgende Möglichkeiten:

1. Multinomiales Modell
2. Lineares Gauß-Modell

# 1. Multinomialmodell

- Die eigentlich stetigen Variablen werden kategorisiert, z.B.
  - 1 unterexprimiert
  - 0 normal
  - 1 überexprimiert
- Die bedingten Verteilungen der Variablen gegeben die Werte der Eltern folgen dann einer Multinomialverteilung
- Nachteil: Informationsverlust durch Kategorisierung
- Vorteil: Modell ist sehr flexibel und entdeckt viele Arten von Abhängigkeiten

## 2. Lineares Gauß-Modell

- Die bedingten Verteilungen sind bedingte Normalverteilungen der Form:

$$P(X|u_1, \dots, u_k) \sim N\left(a_0 + \sum_{i=1}^k a_i u_i, \sigma^2\right),$$

d.h.

$$E(X|u_1, \dots, u_k) = a_0 + \sum_{i=1}^k a_i u_i$$

- Die Koeffizienten  $a_i$  können also durch eine lineare Regression geschätzt werden
- Nachteil: Das Modell erfasst nur lineare Abhängigkeiten
- Vorteil: Es gibt keinen Informationsverlust

# Anwendungsbeispiel: Zellzyklus-Expressionsdaten

- Die Daten wurden an der Bäckerhefe (*Saccharomyces cerevisiae*) erhoben
- Der Datensatz enthält 76 Beobachtungen
- Gemessen wurden die mRNA-Level von 6177 ORF's ( $\approx$  Genen)
- Eine frühere Untersuchung (von Spellman et al. (1998)) lieferte den Datensatz, der aus den 800 Genen besteht, deren Expression sich im Laufe des Zellzyklus ändert

- Gemessen wurde zu unterschiedlichen Stadien des Zellzyklus
- Die Beobachtungen werden als unabhängig angesehen
- Der zeitliche Aspekt wird durch eine zusätzliche Variable berücksichtigt
- Biologisches Vorwissen wird nicht berücksichtigt

# Robustheitsanalyse

- Erstellen eines zufälligen Datensatzes durch unabhängiges Permutieren der Messungen bei den einzelnen Genen
- Vergleich der Konfidenzwerte beim echten und beim zufälligen Datensatz
- Ergebnis: Konfidenzwerte sind beim zufälligen Datensatz deutlich niedriger

Konfidenzwerte für multinomiales (oben) und Gauß-Modell (unten),  
jeweils für Markov- (links) und Ordnungs-Relationen (rechts)

## Vergleich der beiden Modelle

- Vergleich der Konfidenzwerte der Relationen im multinomialen und im Gauß-Modell
- Vergleiche getrennt für Markov- und Ordnungs-Relationen
- Ergebnis:
  - Markov-Relationen: keine Korrelation zwischen den Konfidenzwerten
  - Ordnungs-Relationen: schwache Korrelationen
  - Die beiden Modelle finden vor allem bei den Markov-Relationen unterschiedliche Arten von Zusammenhängen heraus

Vergleich der Relationen von Ordnungs- und Markov-Relationen

## **Biologische Analyse**

Die gelernten Relationen mit hohen Konfidenzwerten entsprechen bekannten biologischen Gegebenheiten

## Ordnungs-Relationen

- Existenz weniger dominanter Gene, d.h. nur diese wenigen erscheinen vor allen anderen Genen

—→ Hinweis darauf, dass diese Gene Initiatoren des Zellzyklus-Prozesses sind

- Dominante Gene werden mit Hilfe des Dominanz-Scores bestimmt:

$$DS(X) = \sum_{Y, C_o(X,Y) > t} C_o(X, Y)^k$$

wobei  $C_o(X, Y)$  die Konfidenz von „ $X$  ist Vorfahr von  $Y$ “ angibt

- Die dominantesten Gene sind bei beiden Modellen gleich und unterscheiden sich nur in der Reihenfolge (geordnet nach *DS*)
- Zu den dominanten Genen gehören:
  - $\left. \begin{array}{l} MSH6 \\ RFA2 \\ POL30 \end{array} \right\}$  beteiligt an DNA-Reperatur
  - CDC45: wird für die Replikation der Chromosomen benötigt

## Markov-Relationen

- Liefern Gen-Paare, deren Proteine an denselben Vorgängen beteiligt sind oder die durch denselben Mechanismus reguliert werden
- Die beiden Modelle liefern zum Teil verschiedene Paare
  - Grund: Das Gauß-Modell erkennt vor allem Beziehungen zwischen hochkorrelierten Genen

- Beispiele:
  - HHT1 - HTB1 (multinom. Modell;  $conf = 0.975$ ):  
beides sind Histone
  - MCD1 - MSH6 (multinom. Modell;  $conf = 0.985$ ):  
beide binden während der Mitose an die DNA
  - HTA1 - HTA2 (Gauß-Modell;  $conf = 1.0$ ):  
miteinander verbundene Histone
- Vorsicht, vor allem beim Gauß-Modell:  
Gene, die sich auf den komplementären Strängen überlappen,  
liefern oft fälschlicherweise hohe Konfidenzwerte

- Bedingte Unabhängigkeiten zwischen hochkorrelierten Genen
- Beispiel aus dem multinomialen Modell:
  - CLN2, RNR3, SVS1, SRO4 und RAD51
  - CLN2 ist eine zentrale und frühe Zellzyklus-Kontrolle
  - CLN2 ist mit den anderen Genen durch eine Markov-Relation verbunden
  - Zwischen den anderen Genen besteht kein funktioneller Zusammenhang

Bayesianisches Netzwerk für das Gen CLN2

# Ausblick

- Erweiterung auf nicht-parametrische Modelle
- Miteinbeziehung biologischen Wissens
- Betrachtung von komplizierteren Relationen, die z.B. mehrere Variablen betreffen
- Miteinbeziehen von experimentellen Manipulationen durch stärkere Unabhängigkeits-Annahmen (kausale Netzwerke)