Bayesianische Netzwerke - Lernen und Inferenz

Manuela Hummel

9. Mai 2003

Gliederung

1. Allgemeines
2. Bayesianische Netzwerke zur Auswertung von Genexpressionsdaten
3. Automatische Modellselektion
4. Beispiel
5. Schwächen und Ausblick

1. Allgemeines

Analyse von Genexpressionsdaten

- Finden statistischer Abhängikeiten zwischen zellulären Regulationsfaktoren

- biologisch interpretierbare graphische Darstellung von Genregulations-Netzwerken

$\rightarrow$ Verwendung von Strukturlernen kausaler Netze

$\rightarrow$ gelernte Netze sind statistische Schätzer für Genregulations-Netzwerke

Probleme der Daten

- starkes 'Rauschen'
- noch wenig biologisches Wissen über Genregulation
- komplexe Mechanismen
- Dimensionsproblem: viele Variablen - wenige Beobachtungen

Andere Verfahren

- Clustern von Genen
- Boolsche Modelle
- System von Differentialgleichungen zur Modellierung von dynamischen Prozessen

2. Bayesianische Netzwerke zur Auswertung von Genexpressionsdaten

Vorteile

- Robustheit
- nicht eingeschränkt auf paarweisen Vergleich zwischen Variablen
- haben probabilistische Semantik
- Behandlung von latenten Variablen möglich
- Priori-Wissen kann mit der Information aus den Daten kombiniert werden
- biologisch interpretierbar

Grundlagen

- **Bayesianisches Netzwerk:** gerichteter azyklischer Graph (DAG) + bedingte Wahrscheinlichkeitsverteilungen $P(A_i|\text{Eltern}(A_i))$

$\rightarrow$ Berechnung einer Likelihood

- **Variablen:** Expressionsniveaus von Genen (z.B.)

- **Kanten:** bedingte Abhängigkeiten, bzw. kausale Zusammenhänge zwischen Genen

$\rightarrow$ Bayesianische Netzwerke beschreiben gegenseitige Kontrolle der Genexpressions-Faktoren

Bayesian Scoring Metric

- Modellvalidierung

- Definition:

$$
\begin{aligned}
S(G) &= logP(G|D) \\
&= \underbrace{logP(G)}_{Log-Priori} + \underbrace{logP(D|G)}_{Log-Likelihood} + \underbrace{c}_{Konstante} \\
\text{mit } P(D|G) &= \int_{\theta} \cdots \int_{\theta} P(D|\theta, G) \cdot P(\theta|G) d\theta
\end{aligned}
$$

- Differenz der Scores zweier Modelle bestimmt die Signifikanz

- Eigenschaften:
  - Scores werden besser, wenn das Modell gegen die richtige Darstellung des genetischen Netzwerkes konvergiert
  - Komplexität wird bestraft
  - Modelle dürfen unvollständig sein
  - Unsicherheit bezüglich der Abhängigkeiten erlaubt
  - latente Variablen können durch MCMC-Methoden bewertet werden

3. Automatische Modellselektion

Problem bisher: Modelle, die verglichen werden sollen, müssen erst formuliert werden

jetzt: automatische Modellsuche durch Fusionierung von biologischem Vorwissen und Genexpressionsdaten

→ Modellselektion durch heuristische Suchalgorithmen:

- greedy search
- simulated annealing

Greedy Search

- beginnt z.B. mit dem leeren Modell
- sucht in jedem Schritt diejenige Struktur aus der 'Nachbarschaft' mit dem höchsten Score
- hat diese Struktur einen höheren Score als der aktuelle Graph, so wird sie zum neuen Modell
- **Nachteil:** bleibt leicht in lokalen Maxima hängen

Simulated Annealing

- basiert auf Abkühlungsprozess:
  Wahl einer hohen Pseudo-Temperatur T_0

- aktuelle Temperatur T_n und Score-Differenz Δ bestimmen die Übergangswahrscheinlichkeit zum alternativen Modell:

  Übergang: wenn $\Delta > 0$ oder $e^{-\frac{\Delta}{T_n}} > random$

  Abkühlung: $T_{n+1} = \rho \cdot T_n, \qquad \rho < 1$

- bei genügend großem ρ findet man das globale Optimum

Erweiterungen des Algorithmus'

- Bedingungen bezüglich der An- und Abwesenheit bestimmter Kanten

- Bedingungen bezüglich der Qualität der Zusammenhänge

$\rightarrow$ Priori-Wissen kann berücksichtigt werden

Model Averaging

- Mittelung über die besten Modelle
- Wahrscheinlichkeit der Anwesenheit einer Kante zwischen Variablen X und Y:

$$\begin{aligned} P(E_{XY}) &= \sum_G P(E_{XY}|D,G) \cdot P(G|D) \\ &= \sum_G \mathbf{1}_{XY}(G) \cdot e^{S(G)} \end{aligned}$$

- gewichtete Mittelwert-Approximation über die n besten Graphen:

$$P(E_{XY}) \approx \frac{\sum_{i=1}^{n} \mathbf{1}_{XY}(G_i) \cdot e^{S(G_i)}}{\sum_{i=1}^{n} e^{S(G_i)}}$$

4. Beispiel

- **Daten:** Biologisches Vorwissen und Expressionsdaten aus *Saccharomyces cerevisiae* (Bäckerhefe)

- **Ziel:** Entwicklung eines Bayesianischen Netzwerkes zur Darstellung der Expressions-Regulation

- **interessierende Gene:** kodieren für Proteine der Pheromon-Antwort

Datenaufbereitung

- **Biologisches Vorwissen**
  - Ziel der Pheromon-Antwort ist das Protein Ste12
  - Ste12 bindet an DNA und aktiviert so die Transkription weiterer Gene
  - man weiß a priori, an welche intergenischen Regionen Ste12 bindet

- **Expressionsdaten**
  - je gesamtes Genom von 320 *S. cerevisiae* Populationen
  - Normalisierung der gemessenen Werte
  - Auswahl von 32 (von 6135) für die Pheromon-Antwort oder andere Aspekte der Zellpaarung relevante Gene
  - Log-Transformierung und Diskretisierung
    $\rightarrow$ jedes Gen hat vier Expressions-Level
  - zusätzliche Variable *mating-type*
    (zwei Paarungstypen bei Hefe)

  $\rightarrow$ 33x320-Matrix

Erstellung der Graphen

- Modellsuche ohne und mit Bedingungen bezüglich bestimmter Kanten

- Wahrscheinlichkeit der Kanten wird mit der gewichteten Mittelwert-Approximation berechnet ($n = 500$)

- Netzwerk mit allen Kanten, die eine Wahrscheinlichkeit über 0.5 besitzen

Ergebnisse (beider Graphen)

- *mating-type* ist Wurzel
- Kanten haben meist recht hohe Wahrscheinlichkeiten
- zwei Subgraphen mit Genen, die jeweils nur in Zellen eines Paarungstyps exprimiert werden
- meist Kanten mit hohen Wahrscheinlichkeiten zwischen *mating-type* und diesen Genen
- TUP1 hat viele Nachkommen - ist Transkriptions-Repressor

Uneingeschränkte vs eingeschränkte Suche

- mit Expressionsdaten allein können oft nicht die richtigen Beziehungen gelernt werden

- bei eingeschränkter Suche können mit tatsächlich beobachteten Beziehungen übereinstimmende statistische Abhängigkeiten betrachtet werden

$\longrightarrow$ eingeschränkte Suche hier sinnvoller

5. Schwächen der Bayesianischen Netzwerke

- statistische Abhängigkeiten lassen nicht unbedingt auf tatsächliche schließen
- multiple biologische Mechanismen erscheinen als jeweils gleiche Abhängigkeitsstrukturen
- bei Daten mit starkem Rauschen geringe Score-Differenz zwischen guten und schlechten Modellen
- Problem der kleinen Stichprobenumfänge
- keine Modellierung von Zyklen möglich

Ausblick

- sinkende Preise von Microarray chips
- Theorie und Algorithmen der heuristischen Suche müssen verbessert werden
- Einbindung von Daten aus anderen Quellen
- Dynamische Bayesianische Netzwerke zur Modellierung zeitlicher Expressionsdaten
- keine Beschränkung auf Analyse von Beziehungen zwischen Genen