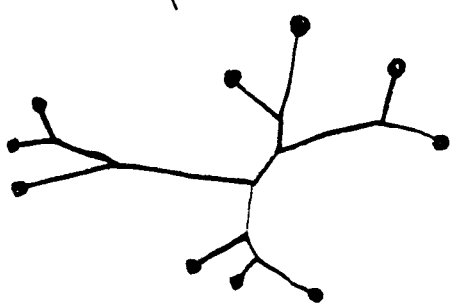
Clustering verfahren

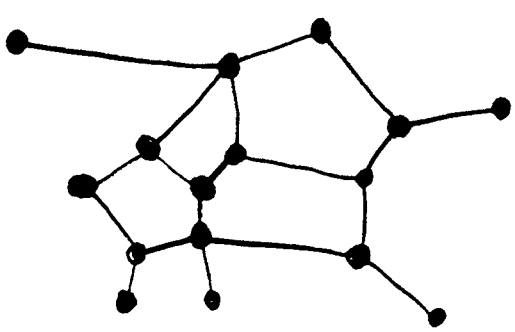
- Cluster analysis of gene expression dynamics
  Marco Ramoni, Paola Sebastiani, Isaac Kohane
  PNAS 14, 2002.

- Cluster inference methods and graphical models evaluated on NCI60 microarray gene expression data
  Peter Waddell, Hirohisha Kishino
  Genome Informatics 11, 2000.

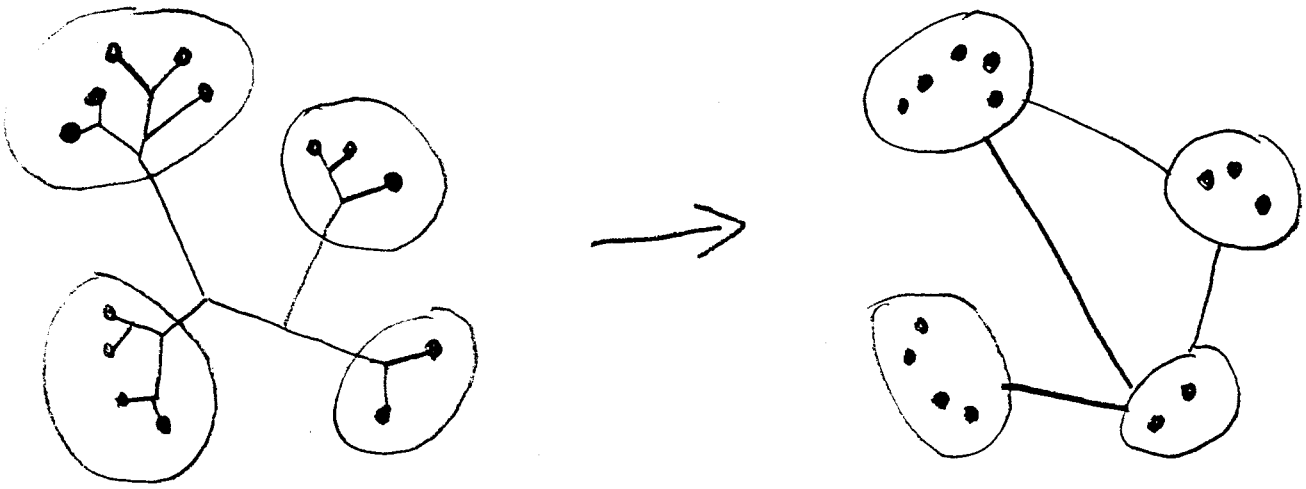
- Clusterverfahren erzeugen Bäume,



keine Netze im eigentlichen Sinne,



- Clusterverfahren erzeugen Gruppen ähnlicher Gene, die zur weiteren Analyse als Einheit gehandhabt werden können.



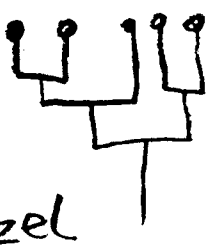
Dadurch vereinfacht sich die weitere Analyse erheblich.

- Clusterverfahren benötigen ein Distanzmass (4) zwischen den Genen, eine sog. Metrik. Diese kann auf vielfältige Weise definiert werden.

- Clustering erfordert also zwei Schritte:

  1. Berechnung aller paarweisen Distanzen
  2. Erzeugung des Baums

- Die Menge der Distanzen kann bestimmte Eigenschaften haben, z.B. ultrametrisch sein:



- Oder auch nicht.

Partielle Korrelation (4.5)

$$pr_{xy.z} = (r_{xy} - r_{xz} r_{yz}) / \left((1 - r_{xz}^2)(1 - r_{yz}^2)\right)^{1/2}$$

Für mehr als 3 Größen:

$$pr_{xy.q} = -w_{xy} / (w_{yx} w_{xy})^{1/2}$$

$$W = V^{-1} \quad \text{(inverse Kovarianz matrix)}$$

Distanzen (4.6)

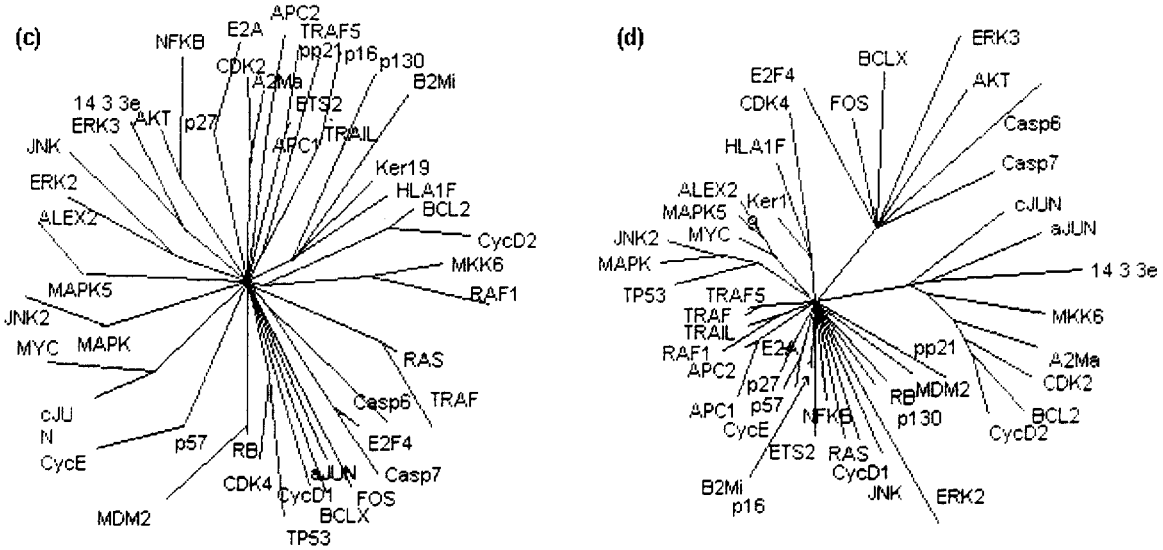
1. $\delta_r = 1 - r$ } $0 \ldots 2$
2. $\delta_{pr} = 1 - pr$ }
3. $\delta_{|r|} = 1 - |r|$ } $0 \ldots 1$
4. $\delta_{|pr|} = 1 - |pr|$ }

- Waddell & Kishino haben verschiedene Clusteringverfahren und verschiedene Metriken ausprobiert.

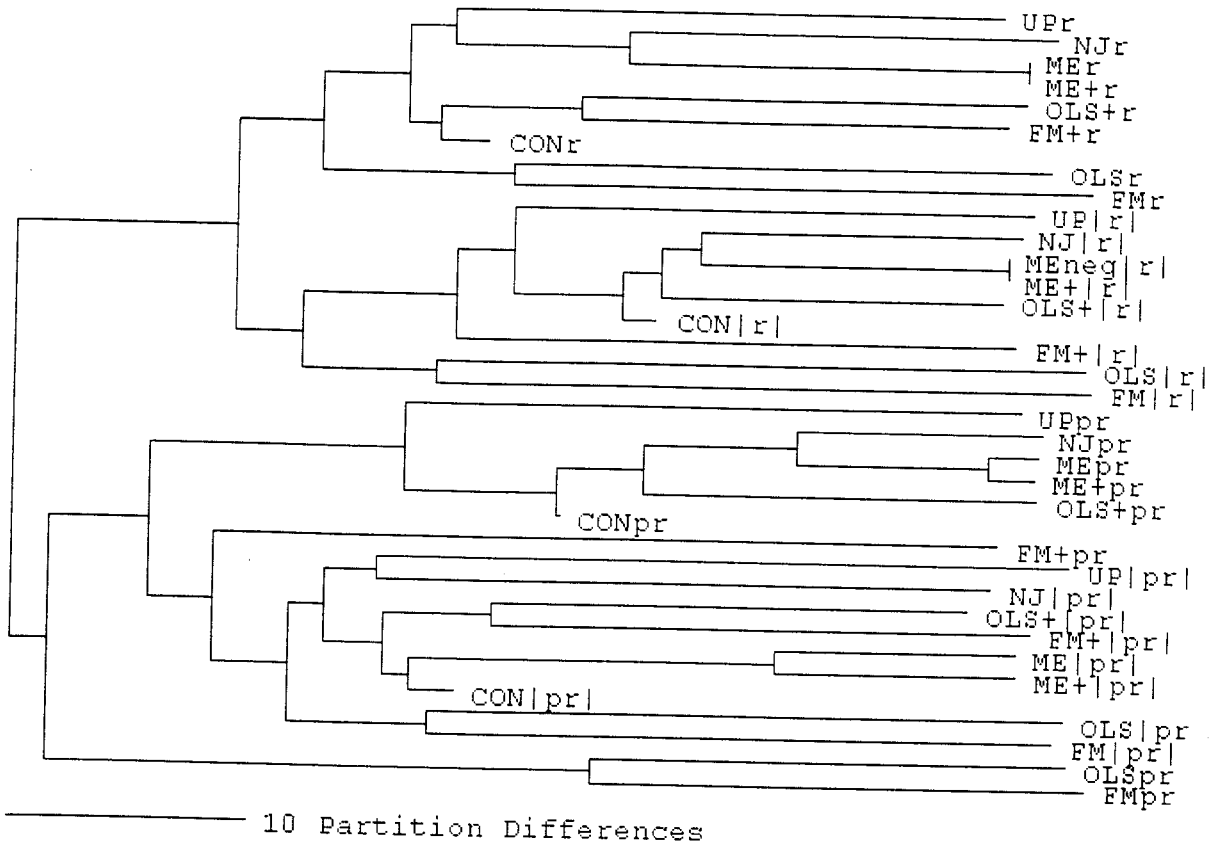
Table 1: Fit of trees to the data.

Correlation Distances					Partial Correlation Distances				
Method	Native Fit[a]	%s.d.[b]	OLS+ %s.d./δ_r[c]	OLS+ %s.d./$\delta_{|pr|}$	Method	Native Fit*	%s.d.	OLS+ %s.d./δ_r	OLS+ %s.d./$\delta_{|pr|}$
UPδ_r	-	-	14.456	17.031	UPδ_{pr}	-	-	17.612	16.435
NJδ_r	-	-	15.007	17.001	NJδ_{pr}	-	-	17.437	16.451
OLS-δ_r	18.041	13.209	15.145	17.006	OLS-δ_{pr}	36.462	18.778	17.881	15.205
OLS+δ_r	21.238	14.331	14.331	17.000	OLS+δ_{pr}	184.932	42.291	17.408	16.499
FM-δ_r	19.364	13.685	15.142	16.932	FM-δ_{pr}	62.538	24.593	17.931	15.260
FM+δ_r	22.853	14.867	14.363	17.032	FM+δ_{pr}	165.514	40.009	17.640	12.679
ME-δ_r	16.322	-	14.528	17.027	ME-δ_{pr}	15.627	-	17.447	16.453
ME+δ_r	16.322	-	14.528	17.027	ME+δ_{pr}	15.894	-	17.440	16.458
UP$\delta_{|pr|}$	-	-	16.768	16.926	UP$\delta_{|pr|}$	-	-	17.896	12.750
NJ$\delta_{|pr|}$	-	-	16.619	16.951	NJ$\delta_{|pr|}$	-	-	17.783	13.948
OLS-$\delta_{|pr|}$	7.524	8.530	16.809	17.011	OLS-$\delta_{|pr|}$	13.650	11.490	17.710	12.285
OLS+$\delta_{|pr|}$	8.159	8.883	16.544	16.940	OLS+$\delta_{|pr|}$	15.612	12.288	17.746	12.904
FM-$\delta_{|pr|}$	12.010	10.777	16.920	17.010	FM-$\delta_{|pr|}$	38.230	19.228	17.688	12.901
FM+$\delta_{|pr|}$	13.399	11.384	16.502	16.963	FM+$\delta_{|pr|}$	44.795	20.814	17.655	12.432
ME-$\delta_{|pr|}$	15.236	-	16.390	16.931	ME-$\delta_{|pr|}$	8.607	-	17.703	12.774
ME+$\delta_{|pr|}$	15.236	-	16.390	16.931	ME+$\delta_{|pr|}$	8.646	-	17.705	12.702

- Die erzeugten Cluster unterschieden sich zum Teil erheblich.



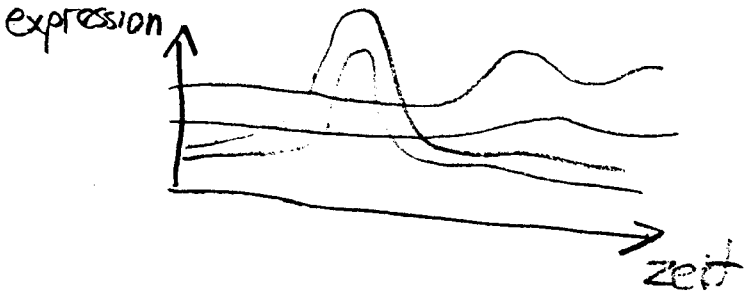
- Das Verhalten der Methoden kann in einem Baum dargestellt werden.



Ergebnis:

1. Gleiche Distanzmasse ergeben "ähnlichere" Bäume als verschiedene bei gleichem Clusterverfahren
2. ME - Methoden sind untereinander "ähnlicher" als gegenüber NJ.
3. Das Spektrum der erhaltenen Bäume ist so gross, dass wohl für jede Anwendung ein passender dabei ist.

- Ramoni et. al. clustern gene aufgrund ihrer expressionsdynamik.



- Ein Bayesianischer Ansatz liefert ein Distanzmass welches die Ähnlichkeit der zeitlichen Expressionsverläufe beschreibt
- Aufgrund dieser Distanzen erzeugt ein Programm namens CAGED Baumstrukturen

- Expressionsverlauf wird beschrieben durch einen autoregressiven Ansatz.

$$\vec{x}_t = \underline{\underline{A}}_1 \vec{x}_{t-1} + \underline{\underline{A}}_2 \vec{x}_{t-2} + \ldots + \underline{\underline{A}}_p \vec{x}_{t-p} + \vec{c} + \vec{\varepsilon}$$

Expression zum Zeitpunkt t hängt von den vorhergehenden Expressionsschritten ab. Oft verwendet man nur den letzten. p ist die Ordnung.

- Der zufällige Fehler $\vec{\varepsilon}$ erlaubt nach Annahme einer geeigneten Verteilung die Berechnung der Wahrscheinlichkeit, das zwei Expressionsverläufe in Wirklichkeit identisch sind.

- Die Likelihood Funktion:

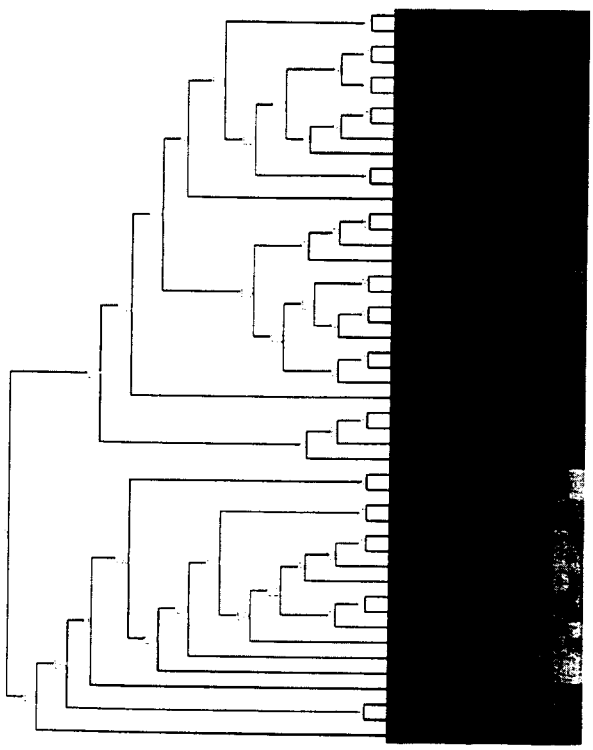
$$P(M_c|y) \propto P(M_c) f(y|M_c)$$

$$f(y|M_c) = \frac{\Gamma(1)}{\Gamma(1+m)} \times \prod_{k=1}^{c} \frac{\Gamma(m_k/m+m_k)}{\Gamma(m_k/m)} \frac{\left(\frac{RSS_k}{2}\right)^{(q-n_k)/2} \Gamma\left(\frac{n_k-}{2}\right.}{(2\pi)^{(n_k-q)/2}} \det(X_k^T X_k$$

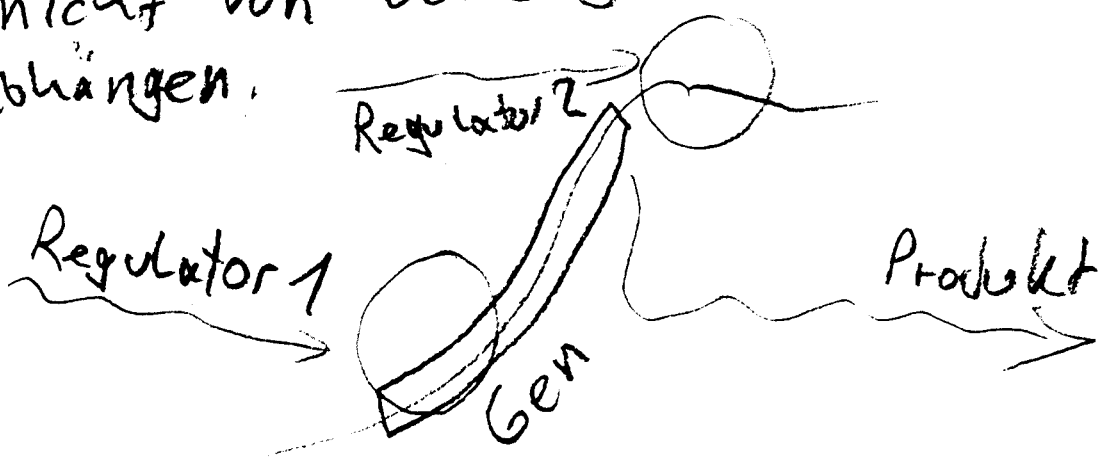
- Distanzmass:

$$D_e(S_i, S_j) = \sqrt{\sum_{t=1}^{n} (y_{it} - y_{jt})^2}$$

- Ergebnis ist ein Baum mit einer Darstellung ⑨ der Expressions verläufe an den Spitzen



- Ramoni et al. berücksichtigen erstmals die Abhängigkeiten in einer Zeitreihe. Andere Verfahren betrachteten die Messpunkte als unabhängig und machten Fehler.
- Der autoregressive Ansatz benötigt unötig viele unbekannte Parameter: Ein Gen kann nicht von beliebig vielen anderen Genen abhängen.



Zusammenfassung:

- Der Ansatz berücksichtigt die dynamische Struktur der Genexpression
- Er reduziert die Zahl der Dimensionen durch Zusammenfassen ähnlicher Gene
- Besser wäre das direkte Ableiten eines Netzwerks