Analysis of Gene Expression Profiles and Drug Activity Patterns for the Molecular Pharmacology of Cancer

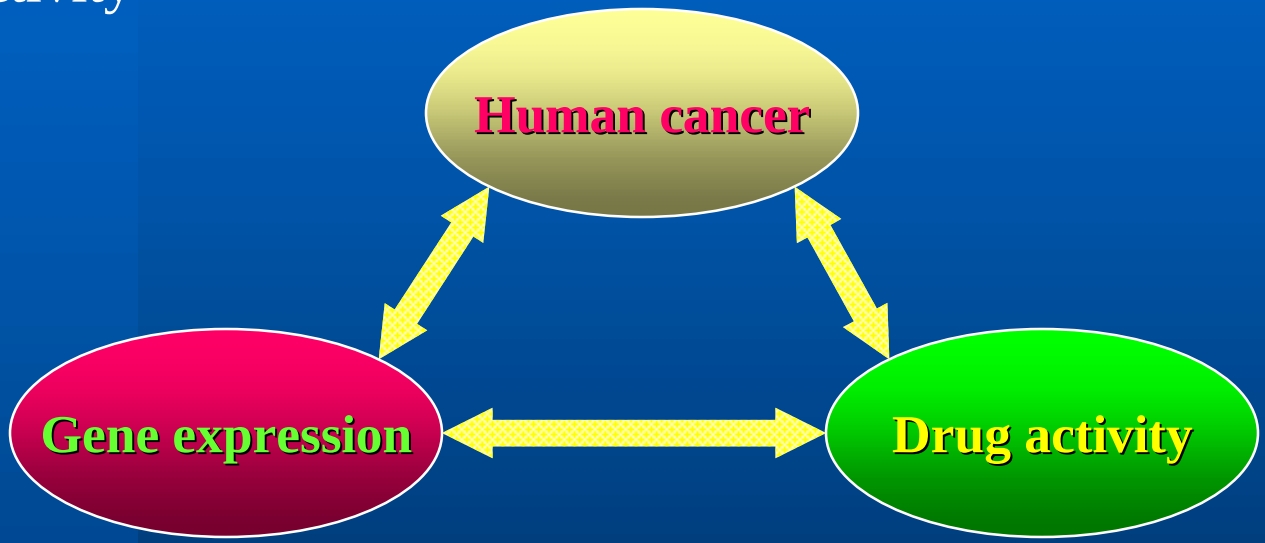
Jeong-Ho Chang, Kyu-Baek Hwang,
and Byoung-Tak Zhang
School of Computer Science and Engineering
Seoul National University
151-742 Seoul, Korea
http://bi.snu.ac.kr

Outline

- Introduction
- Analyzing Cell-Cell Relations through Clustering
  - Experimental Results
- Analyzing Gene-Drug Relations Using Bayesian Networks
  - Experimental Results
- Concluding Remarks

Mining on Gene Expression and Drug Activity Data

- Relationships among human cancer, gene expression, and drug activity



- Revealing these relationships →
  - Cause and mechanisms of the cancer development
  - New molecular targets for anti-cancer drugs

NCI60 Cell Lines Data Set

- From 60 human cancer cell lines [Scherf 00]
  - Colorectal, renal, ovarian, breast, prostate, lung, and central nervous system origin cancers, as well as leukemias and melanomas
- Gene expression patterns
  - cDNA microarray
- Individual targets
  - Analysis of molecular characteristics other than mRNA expressions
- Drug activity patterns
  - Sulphorhodamine B assay → changes in total cellular protein after 48 hours of drug treatment

Analytical Effort

- Analysis of cell-cell relationships using cluster analysis
  - Clustering of cell lines based on
    - Gene expression patterns only.
    - Drug activity patterns only.
    - Both patterns combined with weighted similarity.
- Analysis of gene-drug correlations using Bayesian networks
  - Analysis of gene expression-drug activity dependencies
    - Each cell line is represented by its gene expression profiles and drug activity patterns.
    - Bayesian networks are constructed and analyzed for the discovery of dependencies between gene expressions and drug activities.

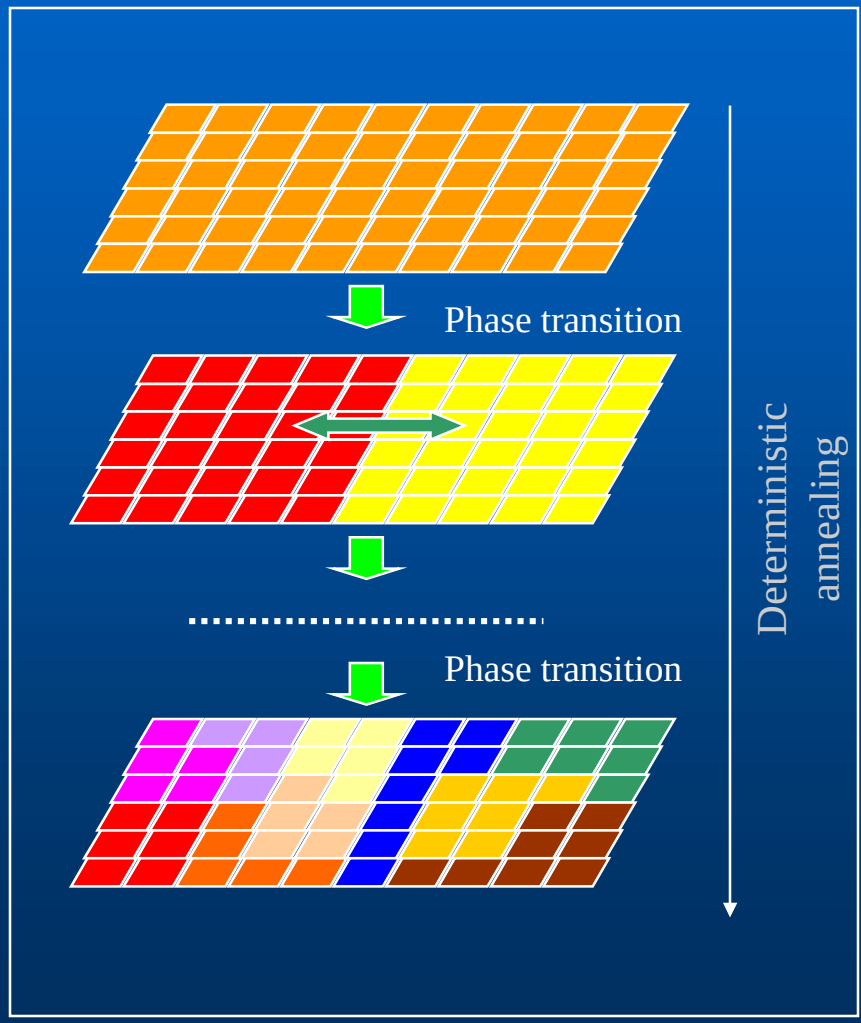
Analyzing Cell-Cell Relations through Clustering

Clustering Methods

- Soft Topographic Vector Quantization [Graepel 98]
  - Based on statistical physics
  - Soft clustering + Topographic mapping
  - Clustering as an optimization
  - Learned by *deterministic annealing*

$$P(\mathbf{x}_i \in C_j) = \frac{\exp\left(-\beta \sum_k h_{jk} e_{ik}(\mathbf{x}_i, \mathbf{c}_k)\right)}{\sum_j \exp\left(-\beta \sum_k h_{jk} e_{ik}(\mathbf{x}_i, \mathbf{c}_k)\right)}$$

h_{jk} : neighborhood function between cluster j and k



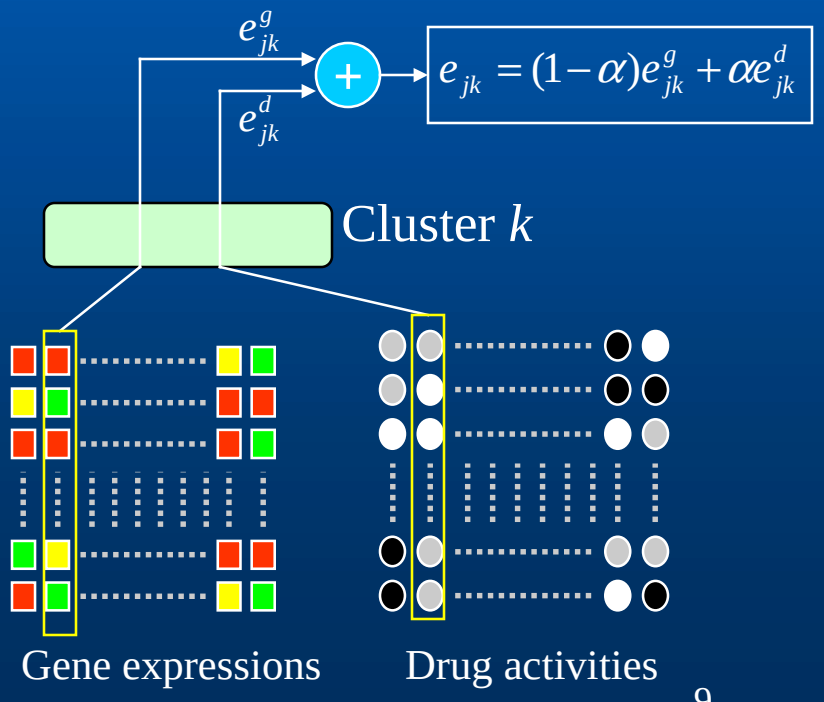
Clustering of Cell Lines based on Gene Expression Profiles

- Among ten runs, result with the best cost value is shown here.
- Neighbor clusters show similar patterns as in the SOM.
- Formed clusters tend to reflect the tissue of origin.
  - CNS, RE, ME, LE, and CO

Using Drug Activity Information in the Analysis of Cell-Cell Relations (1/3)

- Questions
  - Are drug activity patterns in cell lines also related with the tissue of origin?
  - Is this relationship similar to that of gene expression profiles?

- Cluster analysis based on gene-drug information
  - A linear interpolation of distances based on gene expression and drug activity.
  - If both patterns depend on the tissue of origin, the cluster structure will not differ strongly.



Using Drug Activity Information in the Analysis of Cell-Cell Relations (2/3)

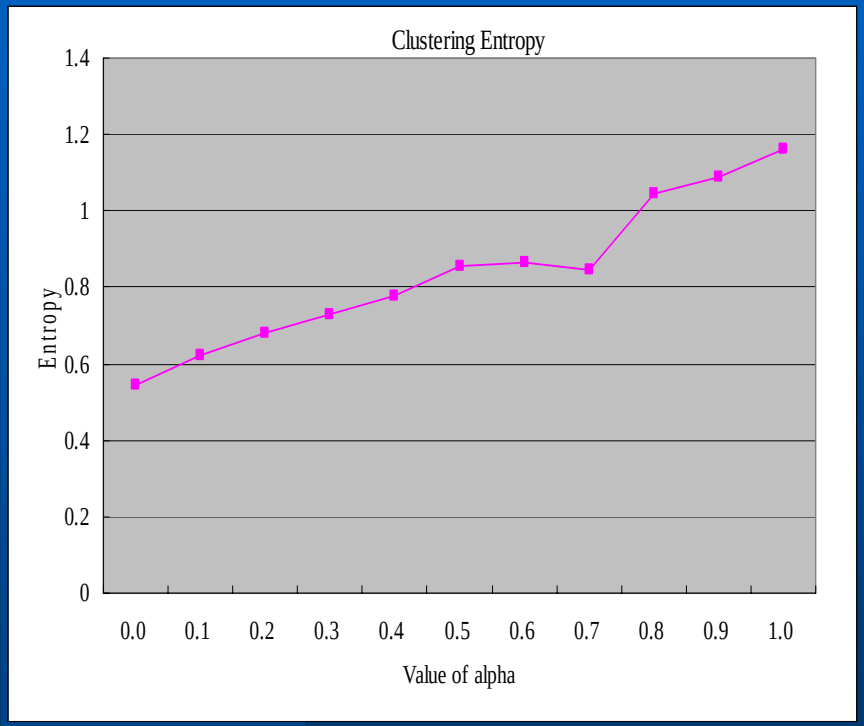
- Quantitative comparison between the clustering analyses
  - Entropy

$$E = \sum_{j=1}^{m} \frac{n_j E_j}{n} \qquad E_j = -\sum_i p_{ij} \log p_{ij} \qquad (0 \le E_j \le \log n_j)$$

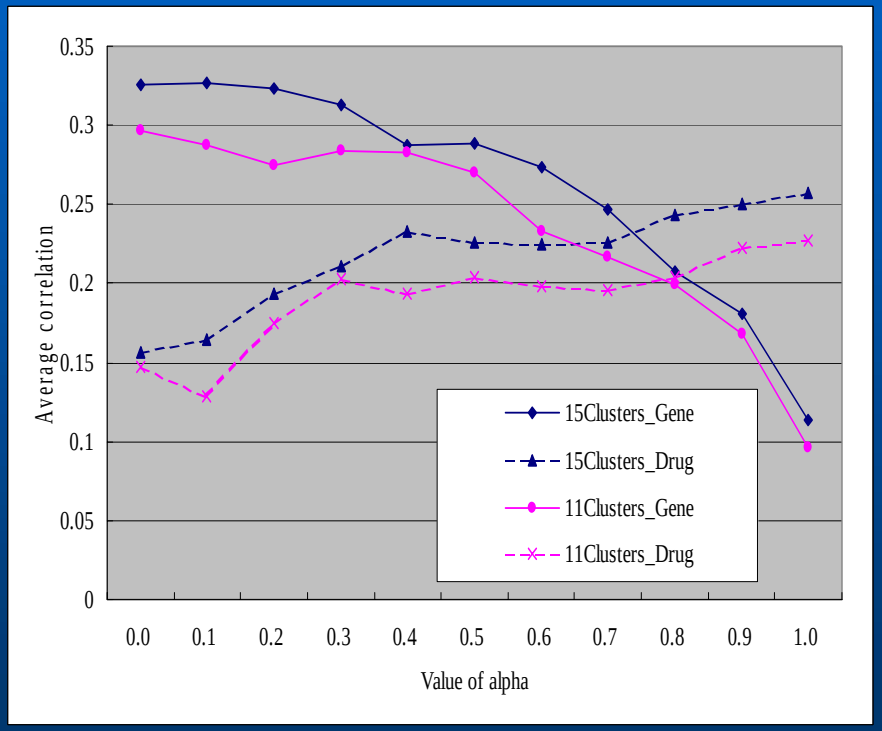
    - p_{ij} : the ratio of members in cluster j which belong to class i
    - n_j : the number of members in cluster j
    - If the number of clusters is fixed,
      - The higher value of entropy → lower reflection of the original class structure.
  - Averaged Pearson correlation

$$R = \sum_{j=1}^{m} \frac{n_j R_j}{n} \qquad R_j = \frac{2}{n_j (n_j - 1)} \sum_{i<k} r(\mathbf{x}_i, \mathbf{x}_k)$$

Using Drug Activity Information in the Analysis of Cell-Cell Relations (3/3)



Entropy with varying α



Average Pearson correlation with varying α

Clustering of Cell Lines based on Drug Activity Patterns

- Among ten runs, result with the best cost value is shown here.
- The clusters does not reflect the tissue of origin, compared to the result based on gene expression profiles.

Analyzing Gene-Drug Relations Using Bayesian Networks

Bayesian Networks

- The joint probability distribution over all the variables in the Bayesian network. [Heckerman 96]

$$P(X_1, X_2, ..., X_n) = \prod_{i=1}^{n} P(X_i \mid \mathbf{Pa}_i)$$

Local probability distribution for X_i

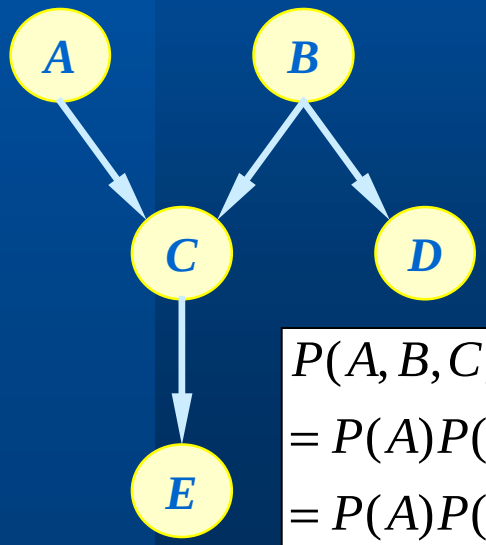
$\mathbf{Pa}_i$: the set of parents of X_i

$\Theta_i = (\theta_{i1}, ..., \theta_{iq_i})$ ~ parameter for $P(X_i \mid \mathbf{Pa}_i)$

$P(\theta_{ij}) = \mathrm{Dir}(\theta_{ij} \mid \alpha_{ij1}, ..., \alpha_{ijr_i})$

q_i : # of configurations for $\mathbf{Pa}_i$

r_i : # of states for X_i



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

Bayesian Network Learning

- Learning for the local probability distribution

$$P(\theta_{ij}) = \mathrm{Dir}(\theta_{ij} \mid \alpha_{ij1},\ldots,\alpha_{ijr_i})$$
$$P(\theta_{ij} \mid D) = \mathrm{Dir}(\theta_{ij} \mid \alpha_{ij1}+N_{ij1},\ldots,\alpha_{ijr_i}+N_{ijr_i})$$

- Learning for the network structure [Friedman and Goldszmidt 99]
  - Search for the best-scoring network structure (greedy search)
  - BD (Bayesian Dirichlet) score [Heckerman et al. 95]

$$p(D,S) = p(S)\cdot p(D \mid S)$$
$$= p(S)\cdot \prod_{i=1}^{n}\prod_{j=1}^{q_i}\frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij}+N_{ij})}\prod_{k=1}^{r_i}\frac{\Gamma(\alpha_{ijk}+N_{ijk})}{\Gamma(\alpha_{ijk})}$$

D : training data

S : network structure

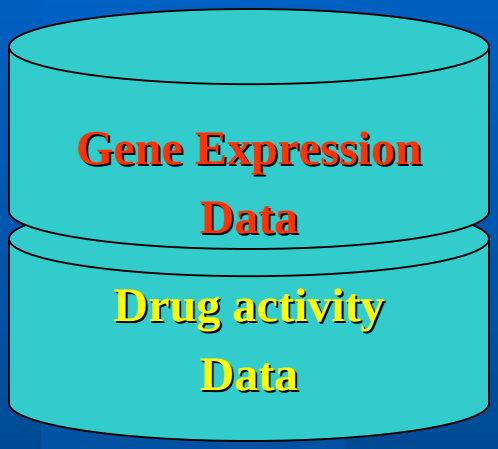
$\alpha_{ij} = \sum_k \alpha_{ijk}, \quad N_{ij} = \sum_k N_{ijk}$

$\Gamma(1) = 1, \quad \Gamma(x+1) = x\Gamma(x)$

Prior

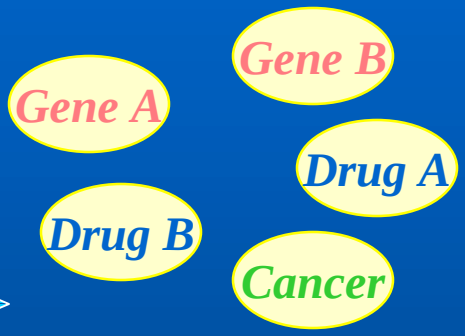
Sufficient statistics calculated from D

Schematic View of the Modeling Approach

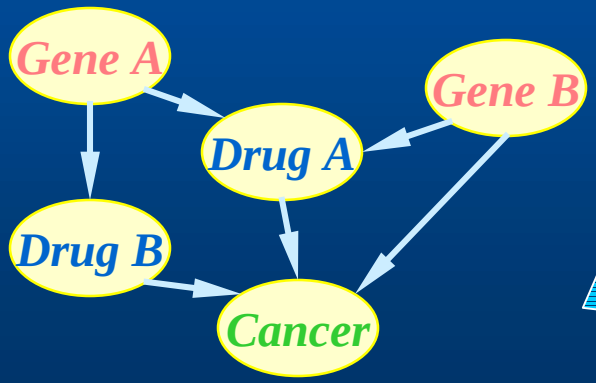


Preprocessing

- Thresholding
- Clustering
- Discretization



- Selected genes, drugs and cancer type node



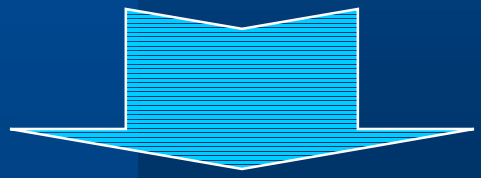
Bayesian network learning

< Learned Bayesian network >

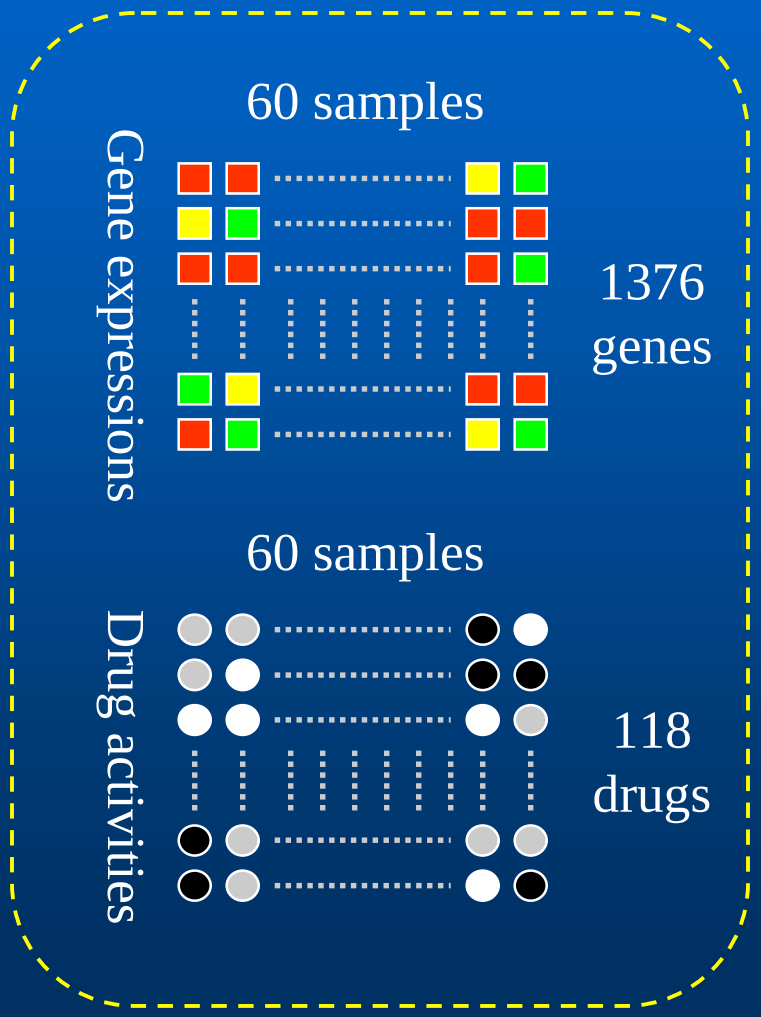
- Dependency analysis
- Probabilistic inference

Data Preparation

- cDNA microarray data
  - Gene expression profiles on 60 cell lines
  - 1376×60 matrix
- Drug activity data
  - Drug activity patterns on 60 cell lines
  - 118×60 matrix

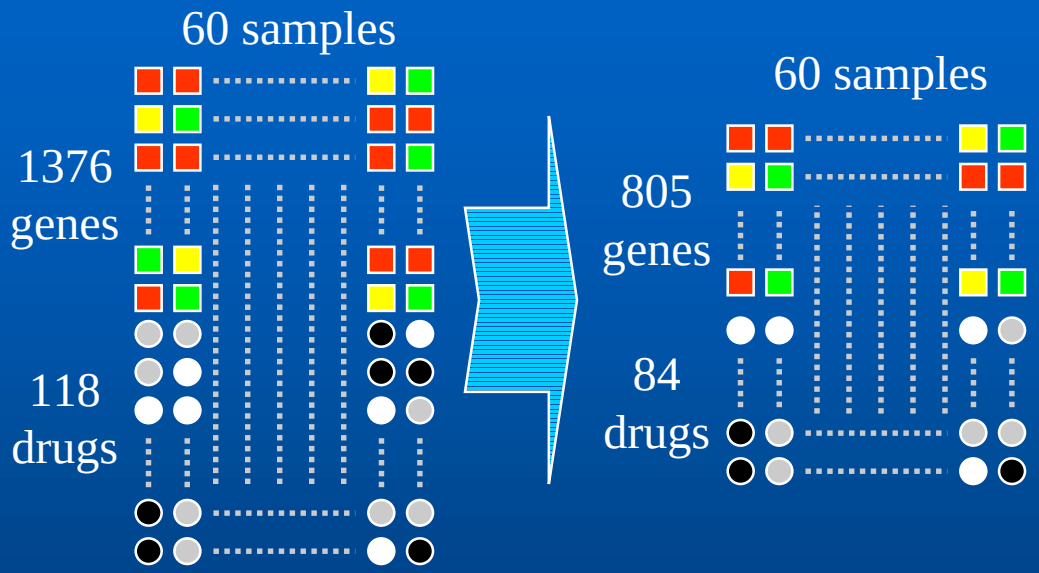


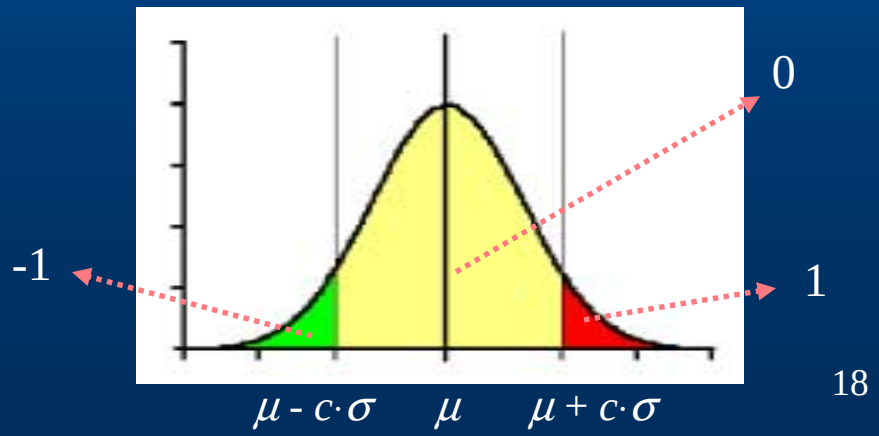
$(1376 + 118) \times 60$ data matrix



Preprocessing

- Thresholding
  - Elimination of unknown ESTs → 805 genes
  - Elimination of drugs which have more than 4 missing values → 84 drugs
- Discretization
  - Local probability model for Bayesian networks: multinomial distribution





Bayesian Network Learning for Gene-Drug Analysis

- Large-scale Bayesian network
  - Several hundreds nodes (up to 890)
  - General greedy search is inapplicable because of time and space complexity.

- Search heuristics
  - Local to global search heuristics
  - Exploit the locality of Bayesian networks to reduce the entire search space.
    - The local structure: Markov blanket [Pearl 88]
    - Find the candidate Markov blanket (of pre-determined size *k*) of each node → reduce the global search space

Local to Global Search Heuristics

Input:

- A data set D.
- An initial Bayesian network structure B_0.
- A decomposable scoring metric,

$$Score(B, D) = \sum_i Score(X_i \mid Pa^B(X_i), D).$$

Output: A Bayesian network structure B.

Loop for $n = 1, 2, \ldots$, until convergence.

- Local Search Step:

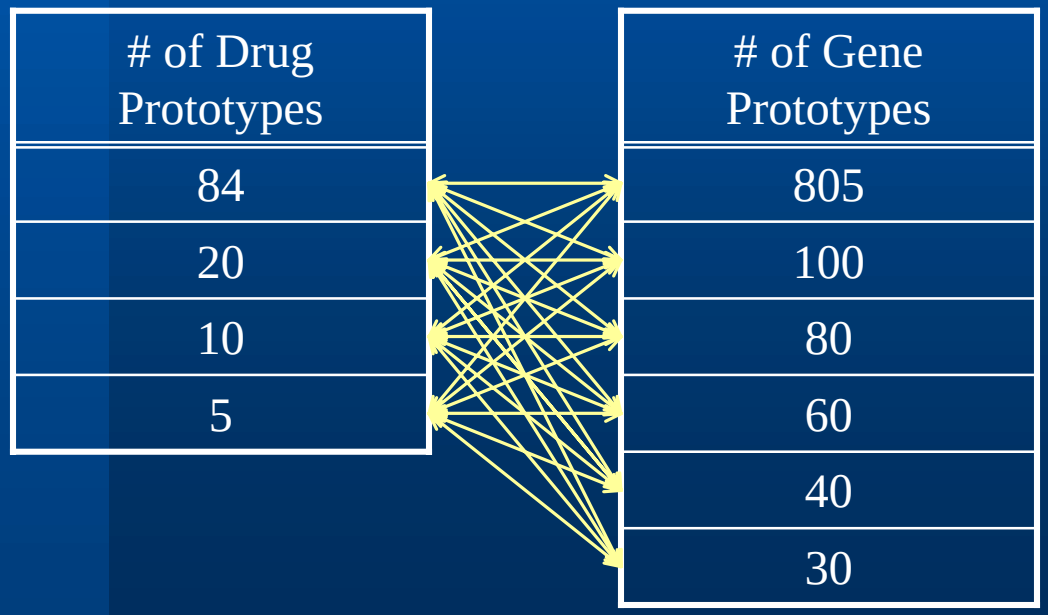
* Based on D and B_{n-1}, select for X_i, a set CB_i^n ($|CB_i^n| \leq k$) of candidate Markov blanket of X_i.
* For each set $\{X_i, CB_i^n\}$, learn the local structure and determine the Markov blanket of X_i, $BL^n(X_i)$, from this local structure.
* Merge all Markov blanket structures $G(\{X_i, BL^n(X_i)\}, E_i)$ into a global network structure H_n (could be cyclic).

- Global Search Step:

* Find the Bayesian network structure $B_n \subset H_n$, which maximizes $Score(B_n, D)$ and retains all non-cyclic edges in H_n.

Dimensionality Problem

- The number of attributes (nodes) >> sample size
  - Unreliable structure of the learned Bayesian network
  - Probabilistic inference is nearly impossible.
- Downsize the number of attributes by clustering
  - Prototype: mean of all members in a cluster



In the preprocessing step

Experimental Results of Gene-Drug Analysis

- Bayesian network learning: implemented by C code
- Network visualization and probabilistic inference: MSBN software [MSBN 96]

Full Size Bayesian Network

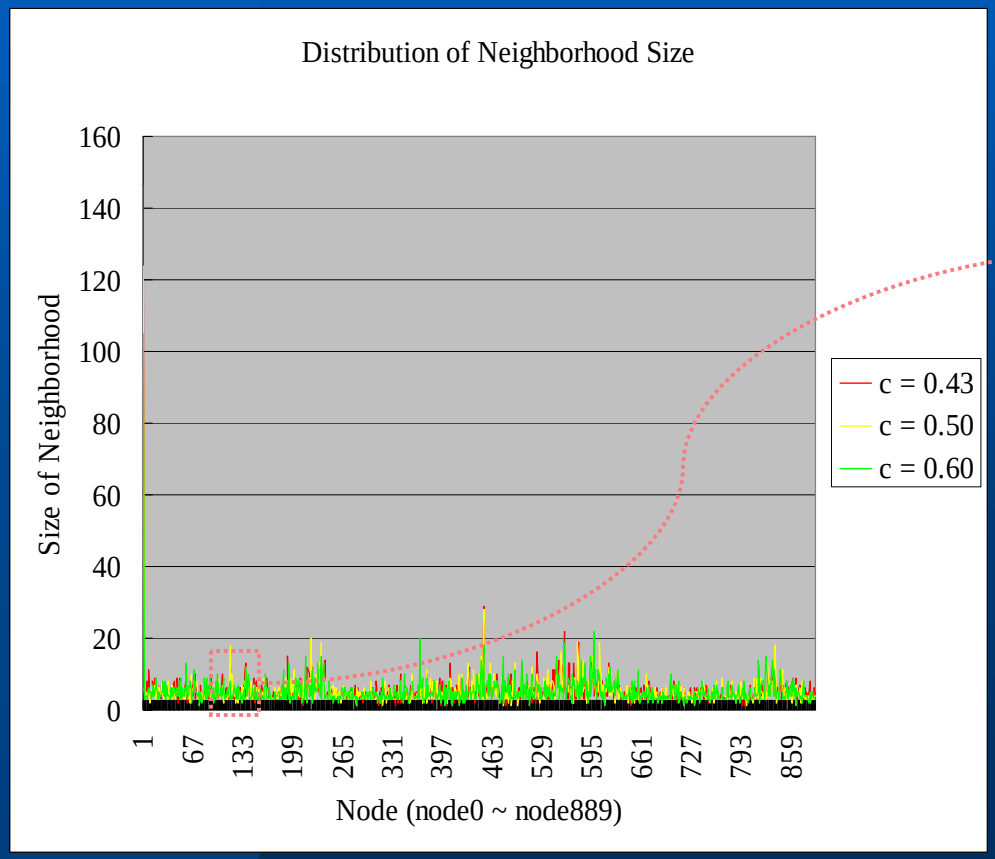
- Node types (890 nodes in all)
  - 805 genes
  - 84 drugs
  - Cancer label
- Discretization boundary
  - $\mu - c \cdot \sigma, \mu + c \cdot \sigma$

c	Distribution Ratio		
	-1	0	1
0.43	33.3%	33.3%	33.3%
0.50	30.8%	38.3%	30.8%
0.60	27.4%	45.1%	27.4%

- Bayesian network learning
  - Varying candidate Markov blanket size (k = 5 ~ 8)
  - Select the best one
  - Three data sets (c = 0.43, 0.50, 0.60) → three Bayesian networks

Influential Nodes (1/2)

- The size of neighborhood of a node → its power of influence on other gene expressions and drug activities



Average correlation between the neighborhood size of all nodes in three Bayesian networks:

0.841

Influential Nodes (2/2)

- Ten influential nodes in average
  - From three Bayesian networks (average neighborhood size = 5.21)

Node Name	# of Neighbors
Origin (cancer type)	125
SID W 487878, SPARC/osteonectin [5':AA046533, 3':AA045463]	25
Homo sapiens Cyr61 mRNA, complete cds Chr.1 [486700, (DIW), 5':AA044451, 3':AA044574]	18.3
SID W 162479, Homo sapiens epithelial-specific transcription factor ESE-1b (ESE-1) mRNA, complete cds [5':H27938, 3':H27939]	16
CDH2 Cadherin 2, N-cadherin (neuronal) Chr. [325182, (DIRW), 5':W48793, 3':W49619]	13.7
H.sapiens mitogen inducible gene mig-2, complete CDS Chr.14 [488643, (IW), 5':AA045936, 3':AA045821]	13.3
SID W 429623, Homo sapiens clone 24659 mRNA sequence [5':AA011634, 3':AA011635]	13.3
SID W 290871, Integrin alpha-3 subunit [5':N99380, 3':N71998]	13
COL4A1 Collagen, type IV, alpha 1 Chr.13 [145292, (EW), 5':R78225, 3':R78226]	12. 7
COL4A1 Collagen, type IV, alpha 1 Chr.13 [489467, (IEW), 5':AA054624, 3':AA054564]	12. 7

Bayesian Network with 45 Prototypes

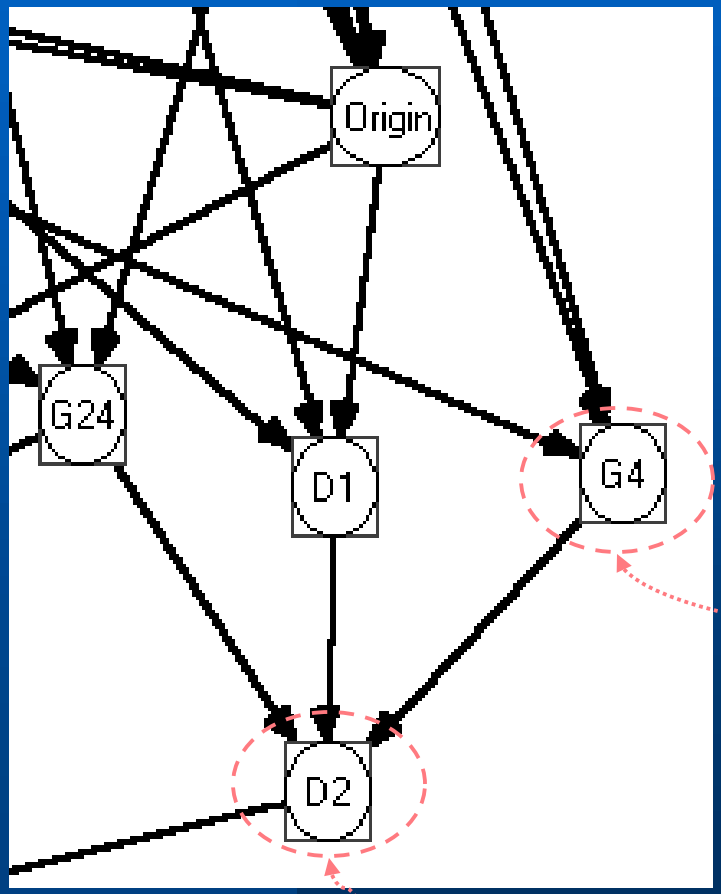
- Node types (46 nodes in all)
  - 40 gene prototypes
  - 5 drug prototypes
  - Cancer label
- Discretization boundary
  - $\mu - c \cdot \sigma, \mu + c \cdot \sigma$

c	Distribution Ratio		
	-1	0	1
0.43	33.3%	33.3%	33.3%
0.50	30.8%	38.3%	30.8%
0.60	27.4%	45.1%	27.4%

- Bayesian network learning
  - Varying candidate Markov blanket size (k = 5 ~ 15)
  - Select the best one
  - Three data sets (c = 0.43, 0.50, 0.60) → three Bayesian networks
  - Probabilistic inference

Correlations between ASNS and L-Asparaginase

- Part of the Bayesian network (c = 0.60)



< Conditional probability table >

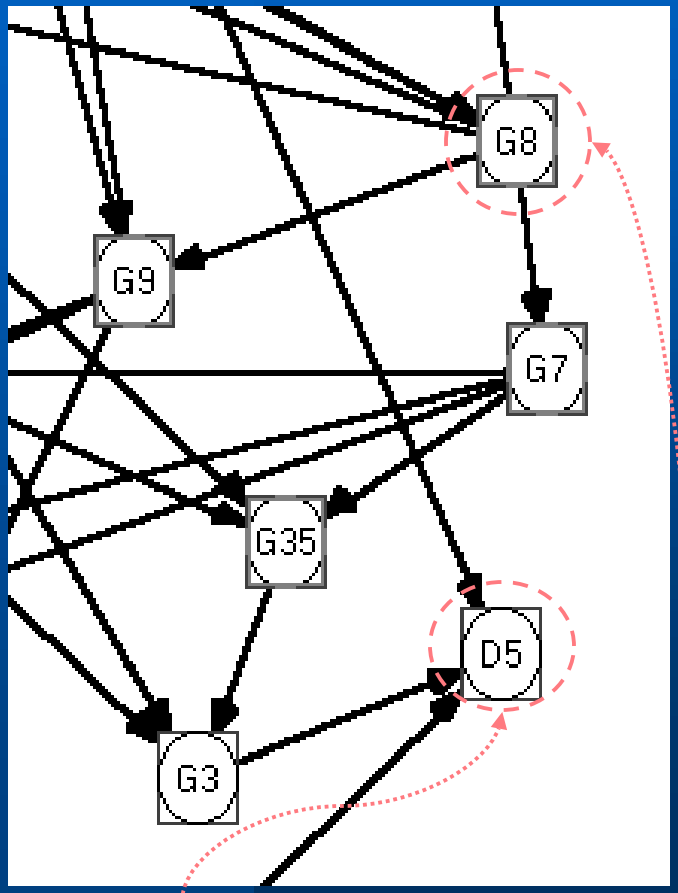
P(D2|G4)	D2 = -1	D2 = 0	D2 = 1
G4 = -1	0.32096	0.27086	0.40818
G4 = 0	0.31387	0.41247	0.27366
G4 = 1	0.32167	0.34920	0.32913

Prototype for ASNS and SID W 484773, PYRROLINE-5-CARBOXYLATE REDUCTASE [5':AA037688, 3':AA037689]

Prototype for L-Asparaginase

Correlations between DPYD and 5FU

- Part of the Bayesian network (c = 0.60)



< Conditional probability table >

P(D5|G8)	D5 = -1	D5 = 0	D5 = 1
G8 = -1	0.33011	0.34747	0.32242
G8 = 0	0.33085	0.34397	0.31799
G8 = 1	0.34048	0.34269	0.31683

Prototype for 5FU

Prototype for DPYD

Bayesian Networks on Subset of Genes and Drugs

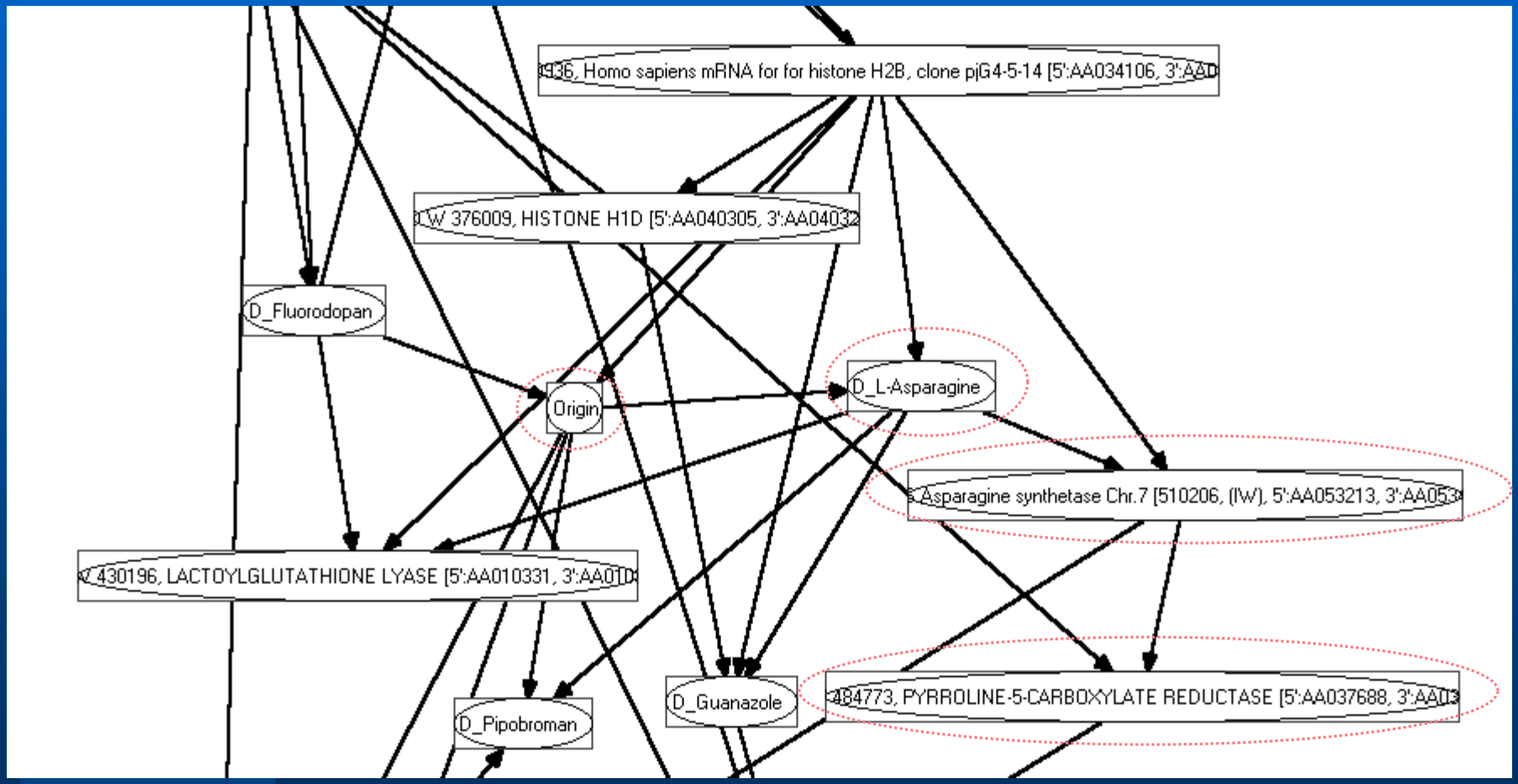
- Node types (17 nodes in all)
  - 12 genes
  - 4 drugs
  - Cancer label
- Discretization boundary
  - $\mu - c \cdot \sigma, \mu + c \cdot \sigma$

c	Distribution Ratio		
	-1	0	1
0.43	33.3%	33.3%	33.3%
0.50	30.8%	38.3%	30.8%
0.60	27.4%	45.1%	27.4%

Clustering of genes and drugs together
- From neighboring clusters

- Bayesian network learning
  - General greedy search with restart (100 times)
  - Select the best one
  - Three data sets (c = 0.43, 0.50, 0.60) → three Bayesian networks
  - Probabilistic inference

Around the L-Asparaginase



< Part of the Bayesian network (c = 0.6) >

Probabilistic Relationships Around the L-Asparaginase

- Cancer type unobserved
  - D1: L-Asparaginase
  - G1: ASNS gene
  - G2: PYRROLINE-5-CARBOXYLATE REDUCTASE

P(D1|G1)	D1 = -1	D1 = 0	D1 = 1
G1 = -1	0.19857	0.27471	0.52672
G1 = 0	0.31110	0.49795	0.19095
G1 = 1	0.42159	0.36279	0.21561

P(D1|G2)	D1 = -1	D1 = 0	D1 = 1
G2 = -1	0.27510	0.35226	0.37263
G2 = 0	0.31621	0.41072	0.27307
G2 = 1	0.33837	0.39664	0.26499

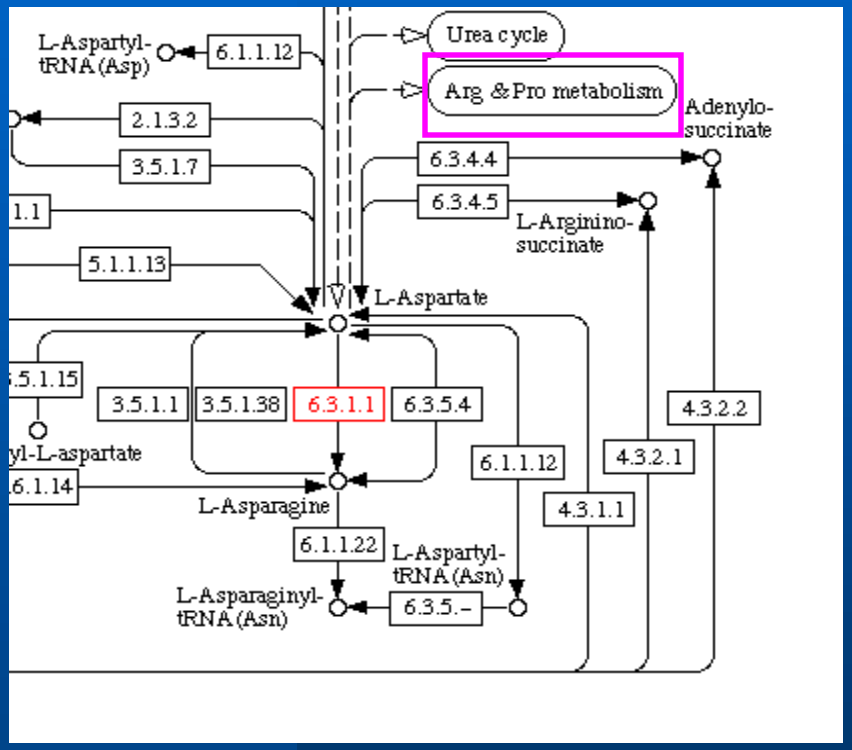
- Cancer type observed (= leukemia)
  - D1: L-Asparaginase
  - G1: ASNS gene
  - G2: PYRROLINE-5-CARBOXYLATE REDUCTASE

P(D1|G1,L)	D1 = -1	D1 = 0	D1 = 1
G1 = -1	0.17536	0.22838	0.59626
G1 = 0	0.27128	0.53790	0.19081
G1 = 1	0.38500	0.42437	0.19063

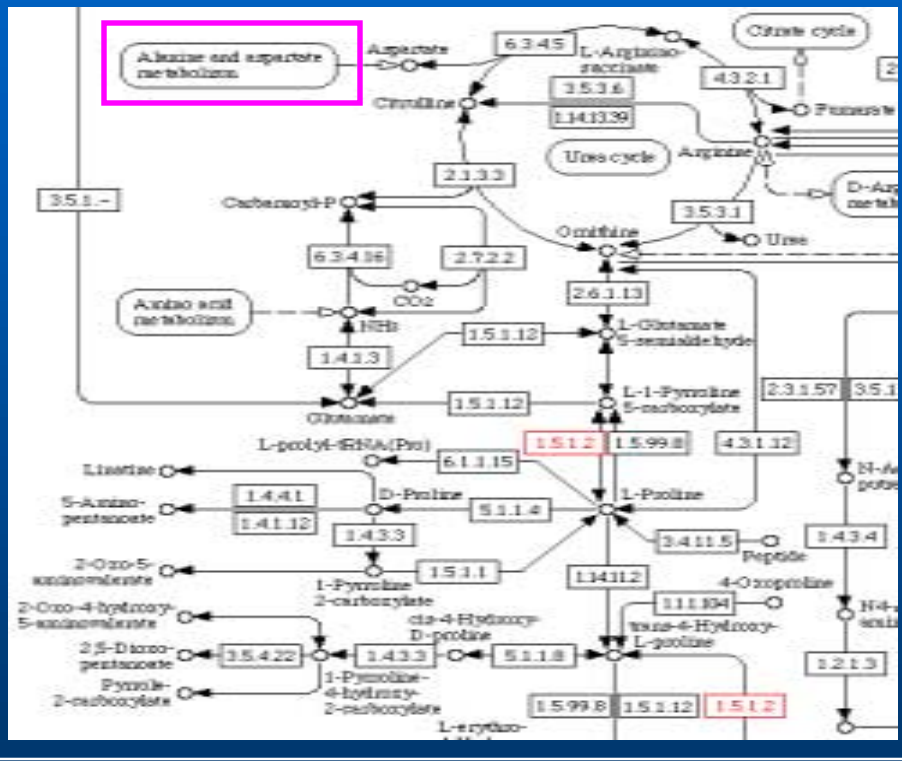
P(D1|G2,L)	D1 = -1	D1 = 0	D1 = 1
G2 = -1	0.23812	0.33853	0.42335
G2 = 0	0.27978	0.42666	0.29356
G2 = 1	0.30371	0.42108	0.27520

ASNS and P5CR in Metabolic Pathway

Alanine and aspartate metabolisim



Arginine and proline metabolism



EC (Enzyme Commission) Number:

6.3.1.1: ASNS, 1.5.1.2: P5CR

Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB)

Source: http://www.genome.ad.jp/kegg

Kyoto Encyclopedia of Genes and Genomes (KEGG), Metabolic and regulatory pathways

Concluding Remarks

- Gene expression profiles have closer relationships with cancer type than drug activity patterns.
- Among hundreds of genes and drugs, only few dozens of them are influential.
- Dimensionality problem
  - Reduction of experimental noise and redundant information ←→ hiding real characteristics of gene expressions and drug activities
- Bayesian network learning
  - Reveal probabilistic relationships between gene expressions, drug activities, and cancer type → biologically meaningful
  - Probabilistic inference in large-scale networks is difficult.

References

- [Friedman and Goldszmidt 99] Friedman, N. and Goldszmidt, M., Learning Bayesian networks with local structure, *Learning in Graphical Models*, pp. 421-460, MIT Press, 1999.
- [Graepel 98] Graepel, T., Statistical physics of clustering algorithms, Master thesis, Technical University of Berlin, 1998.
- [Heckerman et al. 95] Heckerman, D., Geiger, D., and Chickering, D.M., Learning Bayesian networks: the combination of knowledge and statistical data, Technical Report, MSR-TR-94-09, Microsoft Research, 1995.
- [Heckerman 96] Heckerman, D., A tutorial on learning with Bayesian networks, Technical Report, MSR-TR-95-06, Microsoft Research, 1996.
- [MSBN 96] Microsoft Bayes Networks software, Microsoft Corporation, 1996.
- [Pearl 88] Pearl, J., *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann, 1988.
- [Scherf 00] Sherf, U., et al., A gene expression database for the molecular pharmacology of cancer, *Nature Genetics*, vol. 24, pp. 236-244, 2000.