Comparison and Translation Between Information Obtained from Microarray Experiments

Christine Nardini, DEIS Università di Bologna, Italy

EPFL Workshop 11-12 March 2005,
Lausanne, CH

Background

- Microarrays allow global view of biological processes in almost entire genomes
- this provide important insights into our understanding of biological processes
- Because of the large amounts of information generated in each experiment, dealing with such data is challenging on both the biological and computational side

Computational side

- Ongoing research on supervised and unsupervised data mining techniques applied to genomics
- Comparisons of different algorithms, not only in terms of efficiency, but also in terms of results is needed

Biological side

- Use of computational methods to extract and then define meaningful sets of genes
- *Signatures:* genes sets whose pattern of expression is concurrently shared by classes of samples. Used, for example, to identify subtypes of disease, to predict survival and disease free status.

Commonalities

- Both approaches share the generation and definition of new gene sets: *genomic knowledge base.*
- When exploring new biological hypotheses or devising new approaches for data mining, it is useful to interrogate this broad population of gene sets.

FIT
(In collaboration with UCSD and PoliTo)

- Definition of a Reference population of genes set, whose characterization is known
- Definition of a Test population of genes set, whose characterization is unknown
- FIT calculates a measure of similarity between all Test sets against all Reference sets.
- All sets are represented as distributions of genes among the Reference sets (categories of the distribution)

Significance of Enrichment

Evaluated with the hypergeometric distribution:

$$P = 1 - \sum_{i=0}^{k-1} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{i}}$$

N #genes in Background Test population

k (i) #genes in Test falling in the Reference category

M size of Reference set

n size of the Test set

Specificity of the Enrichment

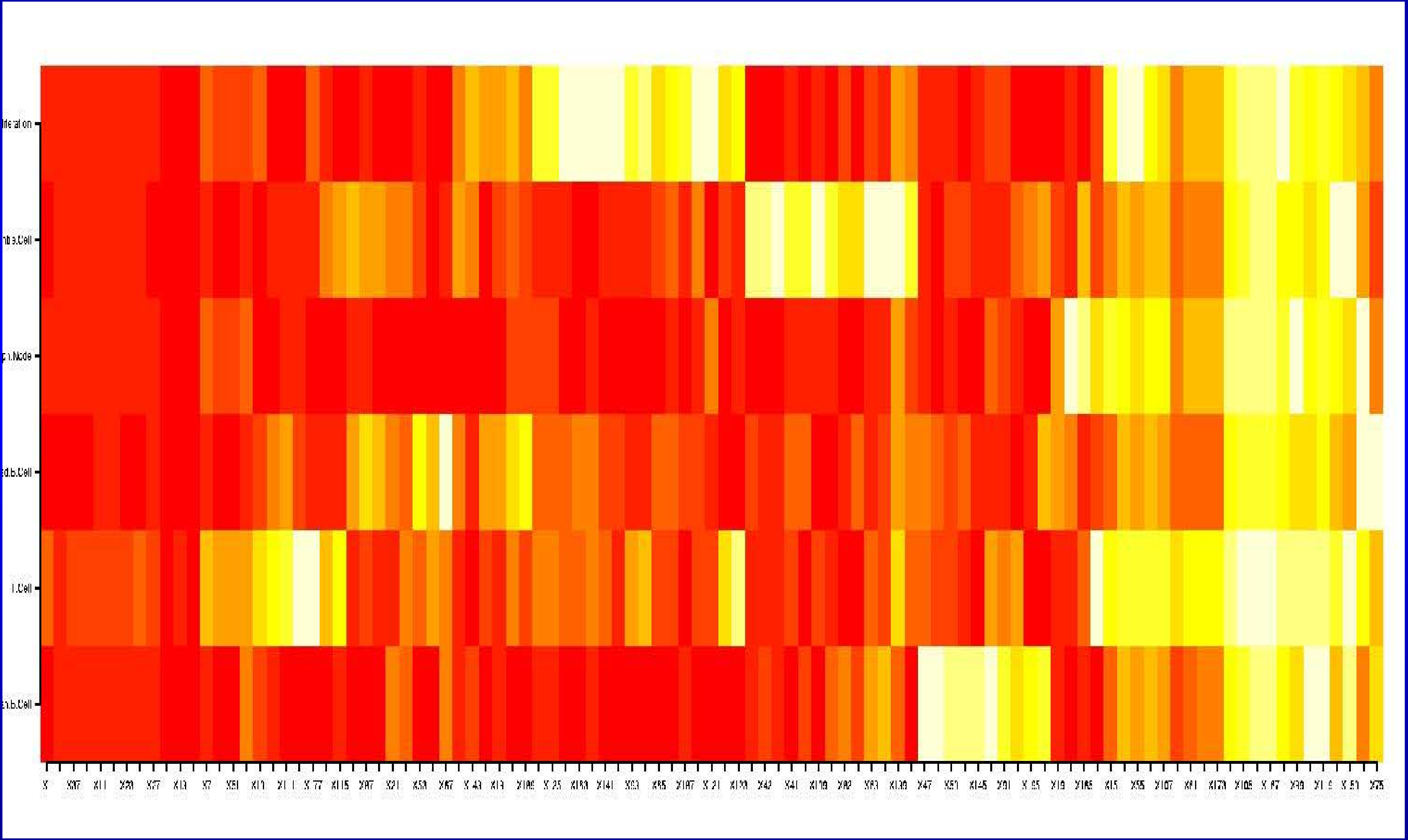
- Measure of correlation between the enrichment of a Test set and the enrichment of a Reference set

$$Spec_{i,j} = 1 - (0.5 \bullet corr(\mathrm{Test}_i, \mathrm{Ref}_j) + 0.5)$$

Lymphoma data set

- Example of comparison between populations of gene sets obtained with different data mining techniques
- **Reference population:** signatures found by Alizadeh et al. [1] (6 sets)
- **Test population:** biclusters found by Cheng et al. [2] (100 sets)

Results Lymphoma data set



EPFL Workshop 11-12 March 2005,
Lausanne, CH

Microarrays Experiments, Pathways, Networks

Possibilities to integrate dynamics in genes expression data

EPFL Workshop 11-12 March 2005,
Lausanne, CH

Assumption

To possibly prevent, interrupt or reverse a disease process, the understanding of the dynamics in the activation pathways and in the biochemistry underlying the relationships in the network of genes and genes products is critically needed

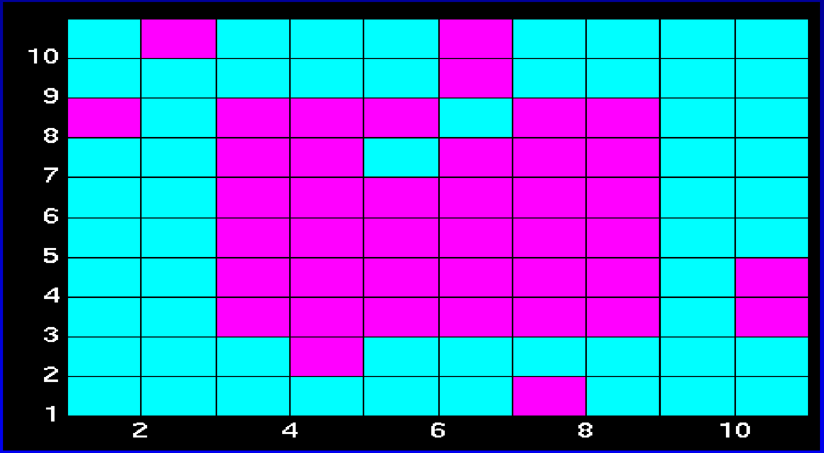
Actors

- **Microarrays:** can give a meaningful photo of the (stable) state of a disease. Microarray data repositories (ArrayExpress, SMD) contain numerous characterized, well described photos (signatures).
- **Pathways:** some of the interactions among genes and gene products are well known and collected in public data bases (KEGG, MAPPFinder), many pathways are incomplete or undefined.
- **Networks:** allow to consider the interaction among numerous "active units". Hopfield neural networks can be used to memorize stable patterns

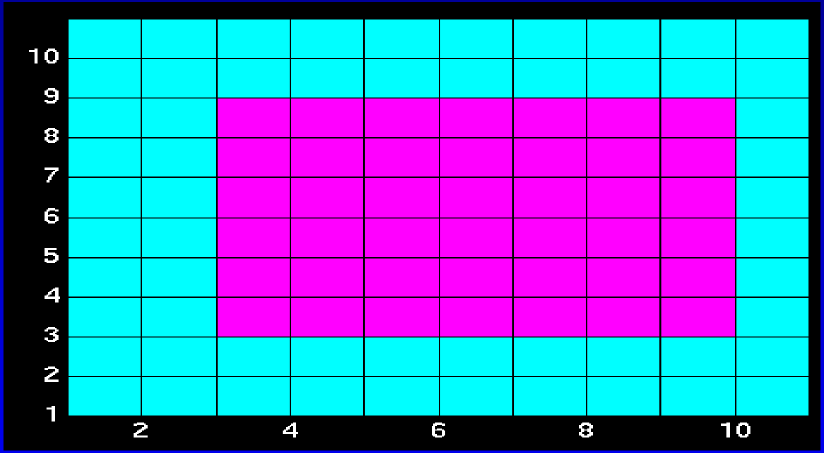
Exploring a possibility

- Representing genes interactions with a neural network, where neurons -> genes and synapses -> activation pathways.
- Considering microarrays signatures as stable patterns in the gene network.
- Training Hopfield networks to memorize such stable patterns.
- Stable patterns are ATTRACTORS (minimums of energy) for any unstable input pattern to the network

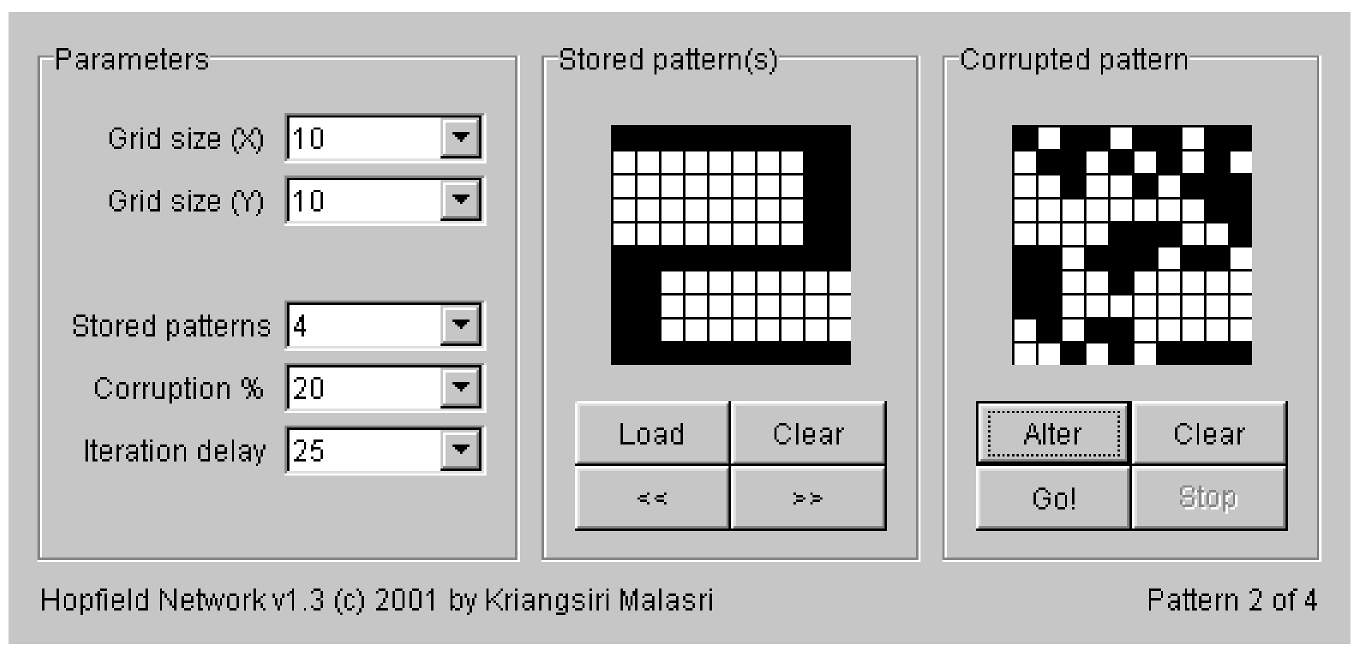
Recognition of stable states



Imperfect, corrupted or unstable pattern



Stable pattern



Parameters

Grid size (X) 10

Grid size (Y) 10

Stored patterns 4

Corruption % 20

Iteration delay 25

Stored pattern(s)

Corrupted pattern

Hopfield Network v1.3 (c) 2001 by Kriangsiri Malasri

Pattern 2 of 4

Key concept

- signatures sharing subsets of genes will be represented by a single stable state unless we shape the synapses with specific direction and weigth
- this gives a sequentiality in the activation of the genes of the signature
- signatures sharing subsets of genes represent an opportunity to describe causal effects
- This ‘intuition’ can be studied:
  - observing the evolution of the state of the network towards different attractors
  - using previous biological knowledge
  - using known pathways information stored in pathway repositories

Bibliography

1. A.A. Alizadeh, M.B. Eisen, R.E. Davis, C. Ma, I.S. Lossos, A. Rosenwald, J.C. Boldrick, H. Sabet, T. Tran, X. Yu, J.I. Powell, L. Yang, G.E. Marti, T. Moore, J. Hudson Jr, L. Lu, D.B. Lewis, R. Tibshirani, G. Sherlock, W.C. Chan, T.C. Greiner, D.D. Weisenburger, J.O. Armitage, R. Warnke, R. Levy, W. Wilson, M.R. Grever, J.C. Byrd, D. Botstein, P.O. Brown, L.M. Staudt, **Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling**, Nature 403(6769), 503-511 (2000).
2. Y. Cheng, G. M. Church, **Biclustering of expression data**, Proceedings of ISMB, (2000).
3. B. M. Fine, M. Stanulla, M. Shrappe, M. Ho, S Viehmann, J. Harbott, L. M. Boxer, **Gene expression patterns associated with recurrent chromosomal translocations in acute lymphoblastic leukemia**,Blood 103(3), 1043-1049 (2004).
4. B. M. Fine and G. J. L. Kaspers and M. Ho and A. H. Loonen and L. M. Boxer, **A genome-wide view of the *in vitro* response to L-Asparaginase in acute lymphoblastic leukemia**, Cancer Res. 65(1), 291-299 (2005).