

Inferring Differential Expression in cDNA Microarray Experiments: a Bayesian Network¹

Valutazione dell'Espressione Differenziale in Esperimenti cDNA Microarray: una Rete Bayesiana

Federico M. Stefanini

Dipartimento di Statistica 'G.Parenti', Università degli Studi di Firenze
Viale Morgagni 59, 50134 Firenze, stefanin@ds.unifi.it

Riassunto: Le attuali difficoltà nell'impiego dei microarray si possono ricondurre alla presenza di un ampio errore sperimentale la cui struttura è assai complessa. Inoltre, il basso numero di repliche per ogni sequenza genica (EST) e l'elevato numero di differenti sequenze simultaneamente considerate richiedono di operare con la massima efficienza computazionale. Il contributo originale di questo lavoro consiste nell'uso di reti Bayesiane miste discrete-Gaussiane per la valutazione dell'espressione differenziale e nella valutazione delle fluorescenze *background* e *foreground* come variabili latenti. L'approccio proposto, pertanto, può essere impiegato anche in presenza di pochi geni invarianti, qualora siano stati predisposti opportuni spot di controllo.

Keywords: Bayesian networks, Graphical models, DNA microarrays.

1. Introduction

It has been hypothesized that within the decade microarrays might be routinely used in the selection, assessment, and quality control of pharmaceutical drugs, as well as for disease diagnosis and for monitoring desired and adverse outcomes of therapeutic interventions. Despite the exciting results obtained in fields like breast cancer, in which gene-expression signature successfully predicts clinical outcome [van't Veer *et al.* (2002)] and acts as a good predictor of survival [van de Vijver *et al.* (2002)], much of the potential promises of microarrays have still to be realized [Petricoin III *et al.* (2002)].

The very complex structure of experimental noise (Nguyen *et al.* (2002)) is at the root of the substantial bias in raw fluorescence values and it typically determines high coefficients of variation. Further statistical challenges come from a typically huge number of genes (ESTs) on the array and a small number of per-gene replicates.

The inclusion of some features of the measurement process, even if simplified, in statistical models for the analysis of differential expression has been already shown to be successful. Computational efficiency may be obtained by casting the model in a framework allowing for exact local computation. Bayesian networks have been introduced in gene-expression analysis by the seminal paper of Friedman *et al.* (2000). Dynamic Bayesian networks have been used to characterize the genetic regulation of a well known box of co-regulated genes in *E.coli* Ong and Page (2001). Spirtes *et al.* (2000) considered discrete Bayesian networks in the analysis of causal relations at genetic level through microarray data.

(¹) This work is partially supported by COFIN 2003

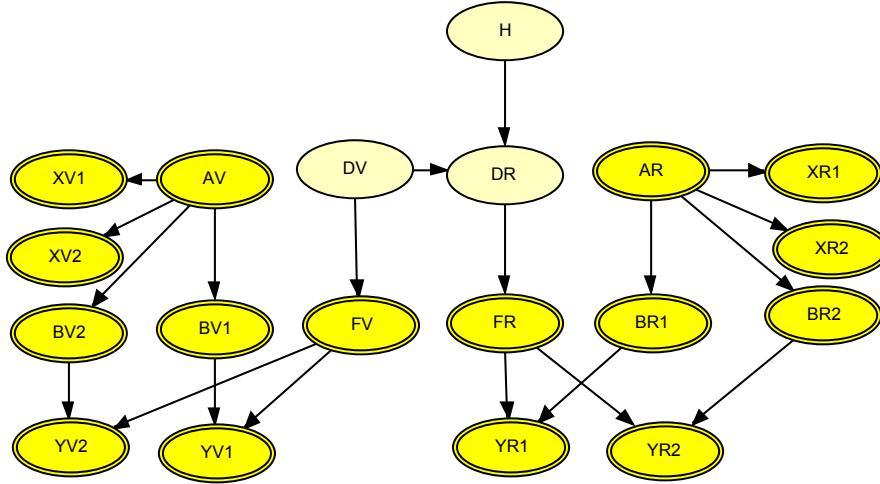
In this work conditionally-Gaussian networks (CGNs) are developed to avoid the loss of information due to the discretization of fluorescence values and to perform exact local computation. The improved formulation of controls relaxes the need of identifying an invariant set of ESTs for the normalization of raw data, which is a very important novelty for studying metabolic pathways instead of genome-wide gene expression.

2. Methods

In this section we assume that raw data have been normalized with respect to artifacts and suitably transformed.

The proposed model defines the measured fluorescence in one channel (dye) Y at one spot as a sum of two latent variables: the within-spot background (B) and the spot foreground (F , sequence-specific fluorescence). The measured background (X) refers to a region closely located near the spot, but outside the spot. Thus we define $Y = B + F + \epsilon$, with ϵ a random variable which represents further noise. The random variables X and B are marginally correlated. Let A be the array average of background values. We assume the conditional independence of B and X given A , namely $B \perp\!\!\!\perp X \mid A$. The above definitions have to be applied to dye-channel specific variables (e.g.: X_R is the measured background on the red dye, X_V on the green one).

Figure 1: Bayesian network for one EST with two replicates.



Let D be the variable representing the DNA concentration of a given EST. Let T be the rate of incorporation of dyes into the sequence S of length L . Following results from the literature, we assume that a logistic function maps D in F . Moreover, while F is a function of T, S, L and D , in the considered experimental setup with L fixed, we can assume the conditional independence: $F \perp\!\!\!\perp (T, S) \mid (L, D)$. Finally, the concentrations on the two channels D_V and D_R are equal given the hypothesis of no differential expression ($H = 0$).

In Figure 1, the conditional independence relations described above are instantiated for a simple case study in which one EST is considered. The two replicates (spots) are indexed by numbers 1 and 2. Continuous variables are labelled by double circles.

Parameter learning is made feasible by the presence of synthetic controls spiked in the targets samples at known concentrations. We assume that the expected value of F_c, c in

$\{V, R\}$ is:

$$E[F_c | D_c] = \beta_{c,0} + \frac{\beta_{c,1} - \beta_{c,0}}{1 + \exp(\frac{\beta_{c,2} - D_c}{\beta_{c,3}})}, \quad (1)$$

with $\beta_{c,0}$ the horizontal asymptote for small D_c values, $\beta_{c,1}$ the horizontal asymptote for large D_c values, $\beta_{c,2}$ the D_c value at the inflection point of the curve and $\beta_{c,3}$ the scale parameter. Due to the definition of F_c as sequence-specific fluorescence. Moreover, we set $\beta_{c,0} = 0$ and $\beta_{c,1} = 65535$, a value which is determined by the sixteen bits used to acquire scanned array images. The discretization of D_c is needed to perform exact local propagation of evidences. The possible problem of saturation is not addressed here. A constant coefficient of variation is often suited to model the dispersion of F_c , although a more flexible model have been investigated in Stefanini (2003).

Inference on the node of interest, H , follows from propagating evidences according to the HUGIN architecture [Cowell *et al.* (1999)]: after finding cliques of the moralized graph (Table 1) and obtaining the junction tree (Table 1), evidences collected on measured spots are assigned to nodes $YV1, YV2, YR1, YR2, XV1, XV2, XR1, XR2$ and propagated.

Table 1: *Cliques of the graph (left) and junction tree.*

Clique	Members	Cliques link		Separator
15	$\{DV, DR, H\}$	14	15	$\{DV\}$
14	$\{DV, FV\}$	13	14	$\{FV\}$
13	$\{FV, AV, BV1\}$	12	13	$\{FV, BV1\}$
12	$\{FV, BV1, YV1\}$	11	13	$\{FV, AV\}$
11	$\{FV, AV, BV2\}$	10	11	$\{FV, BV2\}$
10	$\{FV, BV2, YV2\}$	9	13	$\{AV\}$
9	$\{AV, XV2\}$	8	9	$\{AV\}$
8	$\{AV, XV1\}$	7	15	$\{DR\}$
7	$\{DNAR, FR\}$	6	7	$\{FR\}$
6	$\{FR, AR, BR1\}$	5	6	$\{FR, BR1\}$
5	$\{FR, BR1, YR1\}$	4	6	$\{FR, AR\}$
4	$\{FR, AR, BR2\}$	3	4	$\{FR, BR2\}$
3	$\{FR, BR2, YR2\}$	2	6	$\{AR\}$
2	$\{AR, XR2\}$	1	2	$\{AR\}$
1	$\{AR, XR1\}$			

3. Discussion

There are several differences between (Friedman *et al.*, 2000, FLNP) and what discussed here. First, FLNP dealt with the relation among expression levels of several genes, while here we build a setup for the assessment of differential expression on a one-to-one basis. This approach is yet widespread because microarrays are often used as a preliminary screening tool before applying less noisy techniques of mRNA assessment, like rt-PCR. Second, FLNP did not consider issues like the use of information from the background and the exploitation of control spots to normalize raw data, something to be handled before structural and parametric learning take place following FLNP. Third, FLNP did not propose mixed networks but investigated discrete and Gaussian networks, the first

type for the possibility of capturing non linear relations among genes, the second one for the more efficient use of information under the (approximately) assumption of normal distributions for expression levels.

Here the novelties range from the way the background is introduced into the analysis to the type of control probes to be used so that inferences on the DNA concentration is made possible. The approach leads to an easy interpretation of the output through the marginal distribution of the node H conditioned to the observed evidences (and the data from control spots). Nevertheless, it rests on the availability of control probes and target sequences of fixed (or known) length.

The current network allows exact propagation but still it is limited in several ways: it does not account for outlying observations, thus data must be filtered before inference takes place; the assumption of normality required to perform exact propagation may be too strong in some experimental setting, even after data normalization; the spatial structure of background noise is not yet considered; although originally on a continuous scale, the DNA concentrations D_V and D_R are originally defined as discrete nodes.

References

- Cowell R.G., Dawid A.P., Lauritzen S.L. and Spiegelhalter D.J. (1999) *Probabilistic Networks and Expert Systems*, Springer, New York.
- Friedman N., Linial M., Nachman I. and Pe'er D. (2000) Using bayesian networks to analyze expression data, Technical report, Proceedings of the Fourth Annual International Conference on Computational Molecular Biology.
- Nguyen D., Arpat A.B., Wang N. and Carroll R.J. (2002) DNA microarray experiments: biological and technological aspects, *Biometrics*, 58, 701–717.
- Ong I. and Page D. (2001) Inferring regulatory pathways in e. coli using dynamic bayesian networks, Proceedings of the 9th International Conference on Intelligent Systems for Molecular Biology, Copenhagen, Denmark.
- Petricoin III E.F., Hackett J.L., Lesko L.J., Puri R.K., Gutman S.I., Chumakov K., Woodcock J., Feigal Jr. D.W., Zoon K.C. and Sistare F.D. (2002) Medical applications of microarray technologies: a regulatory science perspective, *Nature Genetics*, 32 supplement, 474–479.
- Spirtes P., Glymour C., Scheines R., Kauffman S., Aimala V. and Wimberly F. (2000) Constructing bayesian network models of gene expression networks from microarray data, Proceedings of the Atlantic Symposium on Computational Biology, Genome Information Systems & Technology.
- Stefanini F. (2003) Background fluorescence in cDNA microarray experiments, *Submitted*.
- van de Vijver M., He Y., van't Veer L., Dai H., Hart A., Voskuil D., Schreiber G., Peterse J., Roberts C., Marton M., Parrish M., Atsma D., Witteveen A., Glas A., Delahaye L., van der Velde T., Bartelink H., Rodenhuis S., Rutgers E., Friend S. and Bernards R. (2002) A gene-expression signature as a predictor of survival in breast cancer, *New England Journal of Medicine*, 347, 1999–2009.
- van't Veer L.J., Dai H., van de Vijver M., He Y., Hart A., Mao M., Peterse J., van der Kooy K., Marton M., Witteveen A., Schreiber G., Kerkhoven R., Roberts C., Linsley P., Bernards R. and Friend S. (2002) Gene expression profiling predicts clinical outcome of breast cancer, *Nature*, 415, 530–536.