Gene expression analysis on biochemical networks using the Potts spin model

Rainer König* and Roland Eils

Division Intelligent Bioinformatics Systems, German Cancer Research Center (DKFZ), Im Neuenheimer Feld 580, 69120 Heidelberg, Germany

Received on October 12, 2003; revised on November 28, 2003; accepted on February 3, 2004

ABSTRACT

Motivation: Microarray technology allows us to profile the expression of a large subset or all genes of a cell. Biochemical research over the last three decades has elucidated an increasingly complete image of the metabolic architecture. For less complex organisms, such as Escherichia coli, the biochemical network has been described in much detail. Here, we investigate the clustering of such networks by applying gene expression data that define edge lengths in the network.

Results: The Potts spin model is used as a nearest neighbour based clustering algorithm to discover fragmentation of the network in mutants or in biological samples when treated with drugs. As an example, we tested our method with gene expression data from E.coli treated with tryptophan excess, starvation and tryptophan repressor mutants. We observed fragmentation of the tryptophan biosynthesis pathway, which corresponds well to the commonly known regulatory response of the cells.

Contact: rkoenig@dkfz.de

INTRODUCTION

DNA microarrays allow us to explore a major subset or all genes of an organism. In multi-conditional experiments, a variety of conditions such as samples of several treatments, mutants, developmental stages and time points are examined. The technique allows to classify tumour samples (Golub et al., 1999; Van’t Veer et al., 2002) and enables the discovery of regulatory mechanisms (Spellman et al., 1998; Gasch et al., 2000). The analysis for these studies was based on supervised or unsupervised learning schemes without applying prior knowledge of relationships between the genes, as e.g. revealed by biochemical pathways. Knowledge of protein–protein interactions was tremendously increased by high throughput techniques (e.g. Uetz et al., 2000) and integrated into the gene expression analysis, revealing crucial subnets (Ideker et al., 2002).

Over the last 30 years, biochemical investigations discovered a more and more consistent image of the metabolism of a cell (see, e.g. Stryer, 1995). For less complex organisms, like Escherichia coli, the metabolic network has been almost completely described (Karp et al., 2002). For yeast, knowledge derived from such biochemical networks was used to support the clustering procedure of gene expression data (Zien et al., 2000; Hanisch et al., 2002).

In our approach, we applied the potts spin clustering technique to investigate the network of E.coli and discover its fragmentations when the organism is treated, e.g. with drugs or starvation or if it is mutated. We represented a metabolic network as a bipartite graph, consisting of alternating enzymes and metabolites. The network was investigated by a nearest neighbour based clustering algorithm. This algorithm uses the potts spin model, coming from statistical physics, where it is used to describe the magnetic behaviour of glassy systems (Blatt et al., 1997; Domany, 1999).

Getz et al. (2000) applied a modification of this algorithm and clustered high-dimensional gene expression profiles without taking networks into account. They classified cancer patients and obtained relevant genes for this. Tornow and Mewes (2003) used the potts spin technique to relate protein–protein interaction data of yeast-two-hybrid screens with gene expression data of a substantial set of mutants and treated cells (Rosetta Compendium; see Hughes et al., 2000). Tornow and Mewes obtained not only commonly known modules but also novel functional modules of yeast. Spirin and Mirny (2003) investigated the connectivity of these protein–protein interaction networks and applied the Potts spin clustering technique to discover high connectivity regions in the network also yielding functional modules.

We applied the Potts spin clustering technique to cluster gene expression data by using the well-established knowledge about nearest neighbour relations of biochemical networks. Only clusters of neighbouring genes with similar expression profiles could occur. The prior knowledge of a metabolic network was applied to reduce noise and fluctuations induced by the gene expression experiments: clusters of genes are found only if they are relevant in a common biochemical context. In a case study, our method was tested with gene expression data of E.coli, provided by Khodursky et al. (2000). They treated E.coli cells with tryptophan excess and starvation

*To whom correspondence should be addressed.
and used tryptophan repressor mutants. The network showed fragmentation in the biosynthesis pathway of tryptophan, reflecting the expected regulatory response of the system.

SYSTEMS AND METHODS

Biochemical reactions were extracted from the EcoCyc database (Karp et al., 2002). The network was established by defining neighbours of enzymes: two enzymes are neighbours if and only if there exists a metabolite that is the product of the former and the educt of the latter one. We compared two networks: a metabolic network with incorporated gene expression data from a case study and a reference net. These networks differed only by the edge lengths: for the reference network, the edge length between each vertex pair was set to 1.

The edge lengths for the network with gene expression data were defined by the differential expressions of treated versus reference cells. For our case study, we collected raw intensity values of gene expression data from the work of Khodursky et al. (2000). They determined relative mRNA levels by two colour hybridization to DNA microarrays. Images were scanned and analysed, yielding signal and local background intensities for each spot (gene representative) and colour. We selected the data for 16 hybridizations of the following sample pairs (protocols and details, see Khodursky et al., 2000):

- reference medium (Vodel and Bonner minimal medium + 0.2% glucose) versus excess tryptophan (reference medium + 50 µg/ml l-tryptophan). Excess tryptophan was added and the samples were collected after 5, 15, 30 min and 1 h (time series);
- reference medium versus tryptophan-starved medium (reference medium + 10 µg/ml indole acrylate); indole acrylate was added and the samples were collected after 5, 15, 30 min and 1 h. Note, that indole acrylate is a tryptophan analogue that prevents the tryptophan repressor from acting and inhibits the tryptophanyl-tRNA synthetase;
- reference medium versus tryptophan-starved medium (reference medium + 15 µg/ml indole acrylate); indole acrylate was added and the samples were collected after 5, 15, 30 min and 1 h;
- mutant tnaA2 in excess tryptophan medium versus mutant tnaA2 trpR2 in excess tryptophan medium;
- wild-type in reference medium versus mutant trpR2 in reference medium;
- wild-type in excess tryptophan medium versus mutant trpR2 in excess tryptophan medium;
- mutant trpEA2 in excess tryptophan medium versus mutant trpEA2 trpR2 in excess tryptophan medium;

where trpR2 denotes repressor-minus, tnaA2 denotes tryptophanase-minus and trpEA2 denotes trp-operon-deleted. Except for two hybridizations, all reference samples were untreated and wild-type. In the other cases, two mutants were compared. One of them was less affected (no mutation of the tryptophan repressor) and therefore also regarded as reference. The expression data were normalized using a commonly used method described elsewhere (Beissbarth et al., 2000). Ratios were taken to define edge lengths (distances) in the graph. The edge lengths, $d_{ij}$, of neighbouring enzymatic reactions $E_i$ and $E_j$ were calculated from the Euclidean distances of the expression profiles for their corresponding genes. These edge lengths were embedded in the network of the treated case (see the following section).

ALGORITHM

The energy term for the clustering algorithm

Each vertex $E_i$ was assigned to an integer $s_i \in [1, 10]$ and is denoted as ‘spin’ in the following (the selection criteria for $s_i$ is described below). Nearest neighbour interactions were defined by the energy term

$$e_{ij} := -J_{ij} \delta(s_i, s_j),$$

where $\delta(s_i, s_j)$ is the Kroneker symbol, and

$$J_{ij} := 1 \left( 1 + \frac{\exp(d_{ij}/0.3)}{1 + \exp(d_{ij}/0.3)} \right),$$

is the interaction strength. $\langle d_{ij} \rangle$ denotes the mean of all edge lengths of the graph. We chose a sigmoidal function of the edge length to better distinguish between low and high similarity of the expression profiles. It was low for high edge lengths and high for low edge lengths. The total energy was defined by

$$E := \sum_{all \ neighbours \ (i, j)} e_{i, j}.$$  \hspace{1cm} (3)

The optimization procedure and clustering

We wanted to obtain clusters in the network that show high interactions. In particular, we were interested in fragmentations of the network that were caused by low interactions between parts of it. We applied the super-paramagnetic Potts spin clustering technique as described in detail elsewhere (Blatt et al., 1997). Briefly, the Monte Carlo Markov chain was as follows:

1. The initial configuration was generated by assigning a random value (spin) to each vertex.
2. Frozen bonds were assigned to neighbours $E_i$ and $E_j$ with the probability

$$p(T) = 1 - e^{-e_{ij}/T}.$$  \hspace{1cm} (4)

3. Subgraphs were connected by frozen bonds. A new configuration was created: spins of each subgraph were assigned to a new, randomly chosen spin value. Spins that belong to the same subgraph were assigned to the same value.
Steps 2–4 were repeated until the maximal number of iterations was reached (100,000).

Frozen bond averages, $c_{ij}$, were calculated for each edge. The first 500 iterations were not taken into account for that (energy, $E$, of the iterations reached a stable regime within the first 500 iterations).

The clustering: the spin–spin correlation, $g_{ij}$, was calculated as

$$g_{ij} = \frac{(q - 1)c_{ij} + 1}{q},$$

where $q = 10$ was the number of possible spin orientations. The clusters were built by a thresholding procedure: if $g_{ij} > 0.6$, a link was set between neighbours $E_i$ and $E_j$. In contrast to Blatt et al. (1997), we did not capture solitary vertices into larger clusters to better trace the extension of the granulation levels. Interestingly, we observed that solitary vertices split from the large clusters at these phase shifts. However, these solitary vertices resulted in a slightly more diffuse granulation. To address this problem, we tuned the clustering threshold $[g_{ij} > 0.6]$ instead of $g_{ij} > 0.5$ as Blatt et al. (1997) proposed. Note that the resulting connected graph depends weakly on this value as long as it is larger than $1/q$ and less than $1 - 1/q$ (Blatt et al., 1997).

The thermodynamic quantity $\chi$ (susceptibility or variance of the magnetization) was calculated as

$$\chi := \frac{N}{T}(m^2 - \langle m \rangle^2),$$

where $N$ is the number of vertices and $m$ is the magnetization,

$$m := \frac{qN_{\text{max}} - N}{(q - 1)N}.$$

$N_{\text{max}}$ defines the largest number of spins with the same orientation. $\chi$ was used to select the temperature for which granulation took place. Peaks of $\chi$ indicate phase shifts: the shift between ordered (magnetic state) and partially ordered (granular, super-paramagnetic state) as well as partially ordered and unordered (non-magnetic state). Starting with low temperature and increasing the temperature, $\chi$ increases quickly when clusters begin to split (granulation). We chose the $T$ of the first peak of $\chi$ to obtain the principle degradation of the system (see Results section).

**RESULTS**

The reference network (all edge lengths $= 1$) was compared with the network for treated cells (edge lengths = Euclidean distances of differential expressions between treated and reference cells). We collected the optimal temperatures for each net by determining the super-paramagnetic phases where the first granulations of the networks took place (Fig. 1). For the reference network, we determined the temperature, $T = 0.3$, for the network of the treated cells at a lower temperature ($T = 0.05$). The network of the treated cells had one large cluster with 495 vertices, one cluster with five vertices, one cluster with two vertices (genes $b2810$, glpE) and four with one vertex (genes $aroH$, $aroF$, $nrdE$, $trpS$). The cluster with five vertices contained genes trpE, trpD, trpC, trpB, and trpA.
Gene expression analysis on biochemical networks

Fig. 2. Tryptophan biosynthesis, taken from the Web sites of the EcoCyc database (http://biocyc.org/). Genes found in the cluster of the network for treated cells encode the enzymes that metabolize chorismate into tryptophan: 4.1.3.27 (trpD, trpE), 2.4.2.18 (trpD), 5.3.1.24 (trpC), 4.1.1.48 (trpC) and 4.2.1.20 (trpA, trpB).

These genes express enzymes that catalyse the metabolic reactions from chorismate to tryptophan (Fig. 2). Note that the expression of these genes depends on the concentration of tryptophan: transcription of the tryptophan operon is reduced by tryptophan excess and induced by tryptophan starvation and when the tryptophan repressor is diminished (see, e.g. Lewin, 2000, chapter on Attenuation).

The observed cluster reflects the adaptation of *E. coli* to excess and starvation of tryptophan and mutation of the tryptophan repressor. When increasing the temperature, the tryptophan cluster remained for $T \leq 0.35$. To validate our result, we performed a clustering of the reference network. $T$ was 0.3 for the first peak of the variance of magnetization. However, the reference network did not show clusters for $T \leq 0.35$ and formed four clusters at $T = 0.4$ (neglecting clusters with one or two vertices). The four clusters contained genes for the transformation from BCCP–biotin to BCCP–biotin–CO₂ (BirA, BioB, BisC, accA), the transformation between 2,3-Dihydroxyphenyl-propionate and carboxyethyl-3,5,cyclohexadiene-1,2-diol (mhpA, mhpB, hcaB, hcaE, hcaF) and genes needed for biosynthesis of farnesyl (uppS, ispB, ispA). Note that these clusters were formed due to the connectivity of the reference net: the reference network shows fragments if parts of the network are separated by the large cluster because of a high number of inner contacts and a low number of contacts to the large cluster. Notably, the tryptophan cluster was not obtained at any temperature in the reference network.

**DISCUSSION**

We presented an approach that allows the analysis of gene expression data with a well-established biological network by applying the super-paramagnetic clustering algorithm as an unsupervised-clustering technique. It was used to reveal partitions or sub-graphs of the net if genes for these parts are
expressed differently from the rest of the net. As a case study, we applied gene expression data of tryptophan treated cells of *E.coli*. The algorithm revealed the expected fragmentation of the net. We observed a subgraph of the net that represented genes of the tryptophan biosynthesis pathway. This cluster could be readily obtained by our clustering technique.

The clustering algorithm finds clusters not only because of similar expression profiles but also due to a high connectivity of a subset of genes in the net. To reveal such structural clusters, we compared our network with a reference network where each edge length was set to 1. Several clusters were obtained for the reference network, e.g. containing genes that are involved in the transformation of BCCP–biotin to BCCP–biotin–CO₂. However, the trp-cluster did not occur for any degree of granulation at the reference net. At very low temperatures these genes remain within the single large cluster and only a few single genes split from this cluster. By raising the temperature, the (reference) net splits into several smaller clusters that did not contain the trp-operon genes as a whole but rather split pieces of this operon (data not shown). Note that clusters found for the reference net are independent of the expression data. In contrast, the findings of the network for the treated cells depend strongly on the expression data, i.e. data from a different gene expression profiling experiment may show entirely different clusters of genes.

Blatt et al. (1997) used a Gaussian function for the interaction potential, gauged by a ‘local length scale’ to get a short-range interaction decay. They obtained this function after intense testing of various choices. In initial trials on simple square lattices with artificial data, we tested the Potts spin clustering technique with a variety of interaction potentials, including the Gaussian function (data not shown). We found that the clustering performed best if we applied a sigmoidal function. In contrast, physical dipole–dipole interaction potentials are proportional to \(1/r^3\) and \((1/r^5)\) for very short ranges if \(r\) denotes the distance of the dipoles. However, these potentials did not suit for our method as they do not include the spatial extension of the dipoles. Note, that we did not intend to reproduce a physical behaviour but rather to investigate a physical model to be used as a machine learning approach.

Like us, Hanisch et al. (2002) combined the metabolic network of yeast with expression data (diauxic shift). As in our study, they obtained sub-clusters of similar expression patterns that contain genes with a common functional context. However, they used the average linkage clustering concept as their machine learning technique, which does not explicitly take nearest neighbour relations into account. We argue that the nature and behaviour of biological networks depend strongly on nearest neighbour interactions. A thorough comparison of average linkage and nearest neighbour based methods may gain insight into the importance and understanding of nearest neighbour relations in biological systems.

Performing clustering by advanced methods like self-organizing maps without prior interaction knowledge (as given by the biochemical network) provides clusters of genes with similar expression profiles. These clusters form because of biological effects, such as common regulatory mechanisms, but also because of statistical fluctuations and noise. In contrast, our approach reveals clusters of genes that are expressed within a common biochemical context. Our method can be expected to cope better with statistical fluctuations of the expression data. Note that finding new clusters of genes using our method will not instantaneously imply biological correspondence to the treatment of the cells and also here a further (manual) refinement of the results needs to be done. As an advantage of our strategy, analysing the results is straightforward. In our case study, at the first granulation only a few clusters (one large cluster, one cluster with two vertices, four singletons and the trp-cluster) showed up. Hence, instead of having tens or hundreds of clusters when clustering gene expression data in a common way, here we can easily perform a thorough analysis of the clustered genes for their biological relevance.

We embedded biological knowledge by well-defined nearest-neighbour relations of a biochemical network and applied a clustering technique that explicitly uses such nearest neighbour relations. The presented strategy may also be used to select clusters yielded by common unsupervised learning techniques such as single linkage (e.g. Kawaji et al., 2001). Note that hierarchical clustering with single linkage addresses nearest-neighbour relations. The Potts spin technique averages over a sufficient by large set of Monte Carlo Markov chain steps (100 000 in our case). Hence, we expect that it performs better for our needs compared with more rigid hierarchical approaches. A precise comparison of the Potts spin method with these hierarchical methods remains an interesting question that is presently addressed in our laboratory.

The Potts spin clustering cannot only be used to cluster an already well-established network. Tornow and Mewes (2003) applied it on yeast two-hybrid data to discover functional modules. This approach may be combined with ours if the networks of functional modules found are explored with another data set of expression data, e.g. from a more specialized study of a focused medical treatment.

Our method could be refined by embedding other network based algorithms, such as normal cut (Shi and Malik, 1997), to provide further insights into the clustering of the network and could be used for different biological networks. Gene expression data have been applied on two-hybrid interaction data to obtain transcription networks (Steffen et al., 2002). These networks could be further refined by network clustering: parts of the network may be more valid if these clusters show common regulation mechanisms when gene expression data of studies with varying transcription are applied. In the future, the clustering principle described may be used for challenging aspects of cancer research, e.g. by mapping gene expression
data on signal transduction networks to obtain insights into the malignant behaviour on a system view.

ACKNOWLEDGEMENTS
We thank the organizing team of the German Bioinformatics Conference 2003. The work was funded by the German National Genome Research Network (NGFN 01GR0101) and the Deutsche Krebshilfe.

REFERENCES


