Interactive Molecular Networks Obtained by Computer-aided Conversion of Microarray Data from Brains of Alcohol-drinking Rats

Authors	F. Matthäus[1], V.A. Smith[2], A. Fogtman[3, 4], W.H. Sommer[5], F. Leonardi-Essmann[5], A. Lourdusamy[6], M.A. Reimers[7], R. Spanagel[5], P.J. Gebicke-Haerter[5]
Affiliations	Affiliation addresses are listed at the end of the article

Correspondence

Prof. Dr. P. J. Gebicke-Haerter
Department of
Psychopharmacology
Central Institute for Mental
Health
J5
68159 Mannheim
Tel: +49/621/170 362 56
Fax: +49/621/170 362 55
peter.gebicke@zi-mannheim.de

Abstract
▼

Lists of differentially expressed genes in a disease have become increasingly more comprehensive with improvements on all technical levels. Despite statistical cutoffs of 99% or 95% confidence intervals, the number of genes can rise to several hundreds or even thousands, which is barely amenable to a researcher's understanding. This report describes some ways of processing those data by mathematical algorithms. Gene lists obtained from 53 microarrays (two brain regions (amygdala and caudate putamen), three rat strains drinking alcohol or being abstinent) have been used. They resulted from analyses on Affymetrix chips and encompassed approximately 6 000 genes that passed our quality filters. They have been subjected to four mathematical ways of processing: (a) basic statistics, (b) principal component analysis, (c) hierarchical cluster-ing, and (d) introduction into Bayesian networks. It turns out, by using the p-values or the log-ratios, that they best subdivide into brain areas, followed by a fairly good discrimination into the rat strains and the least good discrimination into alcohol-drinking vs. abstinent. Neverthe-less, despite the fact that the relation to alco-hol-drinking was the weakest signal, attempts have been made to integrate the genes related to alcohol-drinking into Bayesian networks to learn more about their inter-relationships. The study shows, that the tools employed here are extremely useful for (a) quality control of data-sets, (b) for constructing interactive (molecular) networks, but (c) have limitations in integration of larger numbers into the networks. The study also shows that it is often pivotal to balance out the number of experimental conditions with the number of animals.

Introduction
▼

In recent years, moving together with techno-logical progress, modern molecular biology has evolved from focusing on single cell components to the analysis of whole biological systems. Now-adays scientists are able to perform global quali-tative and quantitative analysis of whole networks of molecular interactions within a cell. This has spurred scientists to generate a new branch of biological sciences, whose aim is to understand the functional aspects of entire sys-tems, namely, systems biology. An important task here is to explore the field of gene and gene prod-uct relationships. Understanding mechanisms and dependencies within gene regulatory net-works (GRNs) is crucial for obtaining more detailed insights into pathological processes, and for further drug target identification. DNA micro-arrays, which evolved in the middle 1990s [45],

play today an important role in obtaining expres-sion data of many genes measured simultane-ously. After ample experience with some generations of expression profiling platforms from deRisi&Brown and Affymetrix to Agilent's and Illumina's [7, 8, 37], we are facing substantial improvements both in terms of tissue prepara-tion and reliability and precision of microarrays. This has been accompanied by increasing num-bers of genes with statistically significant p-val-ues identified as differentially expressed in those conditions. Whilst only a few dozens of genes ful-filled those criteria previously, very often several hundreds are detected more recently. Because also statistical tools of processing chip data underwent considerable amendments, occur-rences of false positives or false negatives in those datasets have been reduced as well. This increased trustworthiness into the data has come along with a great deal of confusion to understand and

interpret the biological meaning of all those genes. It turns out to be simply impossible even for an experienced molecular biologist to devise a 2D- or even 3-D molecular network that displays the positions and tentative interactions of those genes (or better gene products). Therefore, a great deal of hope to tackle that problem rests on computer-assisted approaches.

Establishing and manipulating gene regulatory networks *in silico* can help to understand how expression profiles of single genes influence each other or influence the whole network. For those reasons, modelling of GRNs has become a valued objective in modern biotechnology and bioinformatics [14]. The field of modelling of GRNs evolves very quickly and results in new algorithms and approaches published almost every day. Usually, GRNs are represented by directed graphs, with nodes corresponding to genes, and edges indicating relations between the genes. There are several computational approaches for modelling gene regulatory networks, for example: *Boolean networks* [28], *Revelance Networks* [3], *Bayesian networks* [12, 39] and *differential equation models* [5].

Because the majority of all published microarray data do not entail time-course studies [46], we focus here on applying the Bayesian network approach to a time-independent microarray derived data set. A Bayesian network is a probabilistic model that analyses conditional independence structure between genes. Edges connecting two nodes represent probabilistic dependence relations between them, described by conditional probability distributions [27]. Distributions used here can be discrete or continuous, and Bayesian networks can be used to compute likely successor states for a given system in a known state. Bayesian networks have a great advantage of dealing well with noisy measurements and can be easily extended to deal with missing data [20]. Bayesian networks can be also extended in order to capture the dynamic aspect of regulatory networks by assuming that the system evolves in time. Other extensions try to deal with the typical settings related to microarray data (many genes and few time points).

In this report, we have used gene lists obtained from two brain regions (amygdala and caudate putamen) from three rat strains drinking alcohol or being abstinent. RNA extracted from those tissues was analysed on Affymetrix chips and approximatetly 6 000 genes with intensity values >100 in at least 25% of the samples were subjected to (a) basic statistics, (b) principal component analysis, (c) hierarchical clustering, and introduced into (d) Bayesian networks. The questions we aimed to answer were the following:

(a) Which genes exhibit the largest changes in their expression under alcohol intake, and could therefore be used as markers for alcohol addiction?
(b) What is the relation between the expression patterns of the genes identified in (a)?

Materials and Methods
▼
Animals
Three groups of 2–3 months old alcohol preferring rats were used for long-term alcohol consumption and gene expression profiling: male P rats (n = 15; Indiana University, Indianapolis), male HAD rats (n = 13; Indiana University, Indianapolis) and male AA rats (n = 14; National Public Health Institute, Helsinki). The rats were kindly provided by T. K. Li (Department of Psychiatry, Institute of Psychiatric Research, Indiana University School

of Medicine, Indianapolis; [35]) and D. Sinclair (Department of Mental Health and Alcohol Research, National Public Health Institute, Helsinki; [9]). Each rat strain shows alcohol preference, as indicated in abbreviations P (preference), HAD (high alcohol drinking), AA (alcohol accepting). All experimental procedures were approved by the Committee on Animal Care and Use (Regierungspräsidium Karlsruhe), and carried out in accordance with the local Animal Welfare Act and the European Communities Council Directive of 24 November 1986 (86/609/EEC).

According to the protocol of Vengeliene et al. [49], 8 P rats, 7 HAD rats, and 7 AA rats were given *ad libitum* access to tap water and to 5%, and 20% ethanol solution (v/v). All rats underwent a two-week deprivation cycle after 8 weeks of continuous alcohol availability. After the deprivation period, rats were given access to alcohol again and 3 more two-week deprivation periods were introduced in a random manner (the duration between deprivation periods varied between 4 and 16 weeks). The long-term voluntary alcohol drinking procedure, including all deprivation phases, lasted for a total of 52 weeks. Ethanol intake (4.0–5.3 g/kg of body weight/day for HAD-P-AA rats, respectively) was calculated as the daily average across 7 measuring days. For comparison, 3 age– and weight-matched control groups, consisting of 7 P rats, 6 HAD rats, and 7 AA rats, experienced identical handling procedures for the entire duration of the experiment, but did not receive alcohol.

Gene expression profiling in alcohol-preferring rat strains
Preparation of brain samples and RNA isolation have been performed as described [50]. Target preparation was done for individual samples from caudate putamen (cpu) and amygdala (amy) using 5 µg of total RNA. Hybridization to RG U34A arrays, staining, washing and scanning of the chips were performed according to the manufacturer's technical manual (Affymetrix, Santa Clara, CA).

Data mining
Micro Array Suite 5.0 (Affymetrix) derived cell intensity files (CEL) were processed in R 2.1.1 language and environment (http://www.R-project.org) using Bioconductor 1.6 Packages [16]. Each array was inspected for regional hybridization bias and quality control parameters as recently described [41]. Fifty-three arrays (27 from caudate putamen and 26 from amygdala) passed the quality filter and were included in the statistical analysis. Of the 8 799 probe sets on the RG_U34A array, only those with intensity values >100 in at least 25% of the samples were retained (6 344 probe sets). A 3-way analysis of variance (ANOVA) was used to identify differentially expressed genes across strain, brain region and treatment. ANOVA tests for significant differences in the mean of the two treatment groups by comparing the between-groups variability to the within-group variability. The difference in the mean values is not significant if the variability within the groups does not differ from the variability of the whole set. A measure of the significance of the difference between two groups is the p-value with $0 < p < 1$, which is given as the result of an F-test with the variance values as input.

Dimensionality reduction
We use two common approaches to reduce the dimensionality of the data, principal component analysis (PCA) and clustering. Principal component analysis is a linear transformation of the

data which is based on computing the eigenvalues and eigenvectors of the covariance matrix [26]. The eigenvectors (called principal components) to the largest eigenvalues have the feature that, if the data is projected onto the subspace spanned by these vectors, then this projection captures a very high percentage of the variation in the data. This means that the so obtained dimensionality reduction can be performed without loosing much information given by the data.

If the data includes a set of n variables and m observations, each variable can be represented as a point in an m-dimensional space. If the expression patterns of genes differ under alcohol self-administration from the expression patterns in control animals, then the data points (with the experiments as variables and the genes as observations) would form two distinct clusters in this m-dimensional space. We have used PCA to visualize the data by projecting it onto the plane spanned by the first two principal components.

In a further step, we apply a multiple-link nearest neighbour clustering algorithm, a bottom-up clustering method, where iteratively the elements or clusters with the smallest distance are joined. The so obtained hierarchical clustering yields a distance relationship between all elements that can be represented in the form of a tree (or dendrogramme). We use the open access software Cluster 3.0 by Michael Eisen and Michiel de Hoon and Treeview (http://jtreeview.sourceforge.net).

The objective of the clustering is to partition the data into groups, i.e. sets of genes with similar expression patterns for the given experiments, or sets of experiments that yield similar expression patterns for all genes. The advantage of hierarchical clustering over other clustering methods is that the data is not partitioned into a fixed number of groups. Instead, the obtained hierarchy provides information about clustering of the data at all scales, from fine to coarse.

Bayesian networks

A Bayesian network (BN) is a method to graphically display statistical dependencies among a number of variables. It is drawn with the variables as "nodes" connected by "links": a link connecting two nodes indicates that there is a statistical dependence between them. The statistical dependencies in a BN are the minimal set of such dependencies required to explain the data. In this way, BNs can distinguish direct influence among measured variables from indirect influence [11,19].

The direction of the links in a BN does not necessarily correspond to causality; it is only a representation of statistical dependence. All variables that have links directed to another variable are known as the "parents" of the latter variable, known as a "child". The relationship between a variable and its parents can be conceptualized, as the value of the parents are useful for predicting the value of the child. This can be either independent, for example, one parent being elevated means a higher probability the child is elevated, or in a combinatoric manner, for example, two parents being increased means a higher probability the child is increased (with no obligatory relationship to each individual parent).

Discrete BNs require variables to come in a number of discrete states, i.e. no/yes, low/medium/high, etc., and are capable of representing many types of statistical dependencies including linear, nonlinear, stochastic, and arbitrary combinatoric [11,19]. Multiple types of variables can be combined in a BN, for example gene expression data and experimental manipulations can both be present as nodes in a BN. When working with data that is originally continuous, discrete BNs require intelligent discretisation of the values [13,54]. For example, some data sets have clear modes, which can suggest the number of division of discrete states. In other cases, a method of quantile discretisation is used frequently (i.e., lower 33.3% = low, middle 33.3% = medium, etc.). When presented to the BN algorithm, discrete states are coded as sequential integer values, starting from 0. The sequential nature of these states enables the BN algorithm to provide an "influence score" for each link produced [54]. This varies between −1 and 1, showing both direction and magnitude of influence between two variables. When the influence score is precisely equal to zero, this indicates that the influence is non-monotonic, i.e., neither clearly positive nor negative (for example, a U-shaped or combinatoric relationship; [54]).

A BN algorithm is based on the calculation of a score, known as a Bayesian Scoring Metric (BSM), which represents the fit of a given network to a given dataset [11,19]. The BSM used by Banjo, which is applied here, includes an inherent penalty for complexity which avoids overfitting networks to a dataset. Because the problem of finding the best BN to describe a set of data is NP-complete (computationally intractable), heuristic search techniques are applied to find a high-scoring network [11,19]. These techniques move through the space of possible BNs using a set of intelligent rules, which guides them to high-scoring networks. Often, some method of combining a number of the highest-scoring networks, or results of multiple searches, or both, is used to provide links with the highest degree of confidence [18].

Here, Banjo v.2.0.1 (http://www.cs.duke.edu/~amink/software/banjo/) was applied to each set of genes, plus three variables representing the drinking/nondrinking conditions, brain regions, and rat strains. We used a simulated annealing search with an equivalent sample size of 1, and a max parent count of 2 (to avoid potential false positives; [55]). Each search was run until it had visited 200 million networks and calculated a consensus network from the top 1000 scoring networks found [18]. 10 separate searches for each set of genes have been carried out.

Naïve Bayes Classifier

The naïve Bayes classifier is related to Bayesian networks. It is a Bayesian network where only one type of link is present: those from a variable representing some condition of interest to variables that are related to this condition [10,34]. In contrast to a BN, links in a naïve Bayes classifier can represent either direct or indirect influences. A naïve Bayes classifier can be used to select variables that have some sort of relation to a condition of interest, by comparing the BSM when the condition is linked to the variable to when it is not. The difference between these two values is the log of the ratio of their probabilities ("log ratio"), representing a measure of how probable it is that the variable is related to the condition of interest. With the BSM used by Banjo, each unit of the log ratio represents how many multiples of e more probable it is that the variable is related versus not related (e.g. log ratio = 1 is e times more probable; log ratio = 5.6 is $e^{5.6}$ = 270 times more probable).

Results
▼
Preprocessing and quality checks

The performed 3-way ANOVA shows that gene expression differs with high significance in the two considered brain areas (amygdala and caudate putamen). Smaller changes in the expression
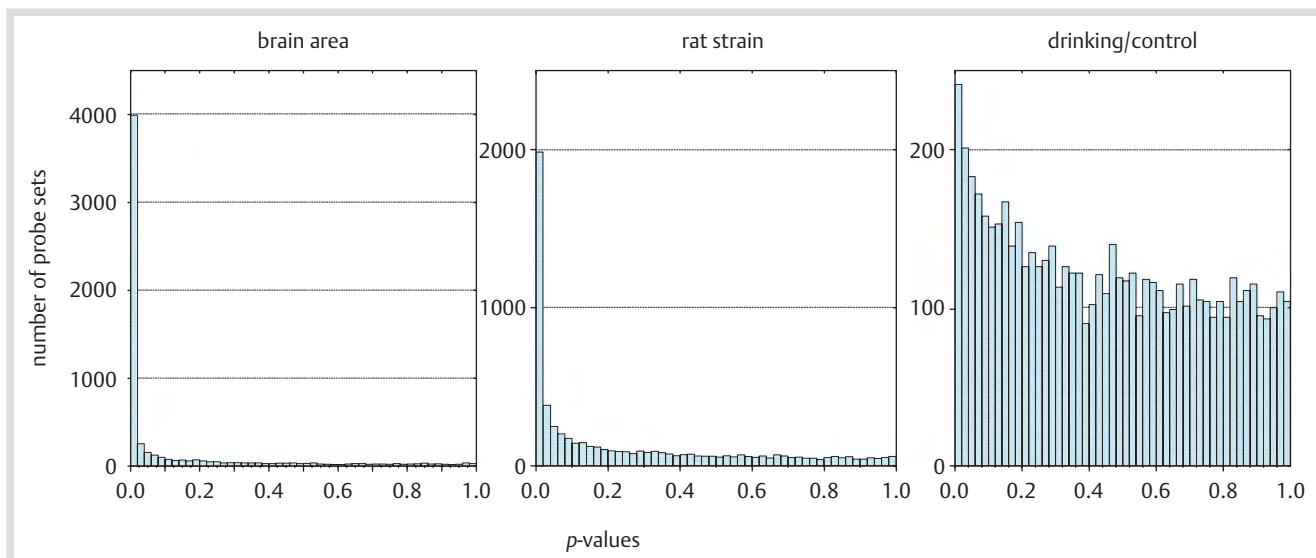
Fig. 1 p-value histogram for main effects of brain region, strain and ethanol treatment from 3-way ANOVA. Each panel shows the number of probe sets (y-axis) within a p-value bin of 0.02 width.
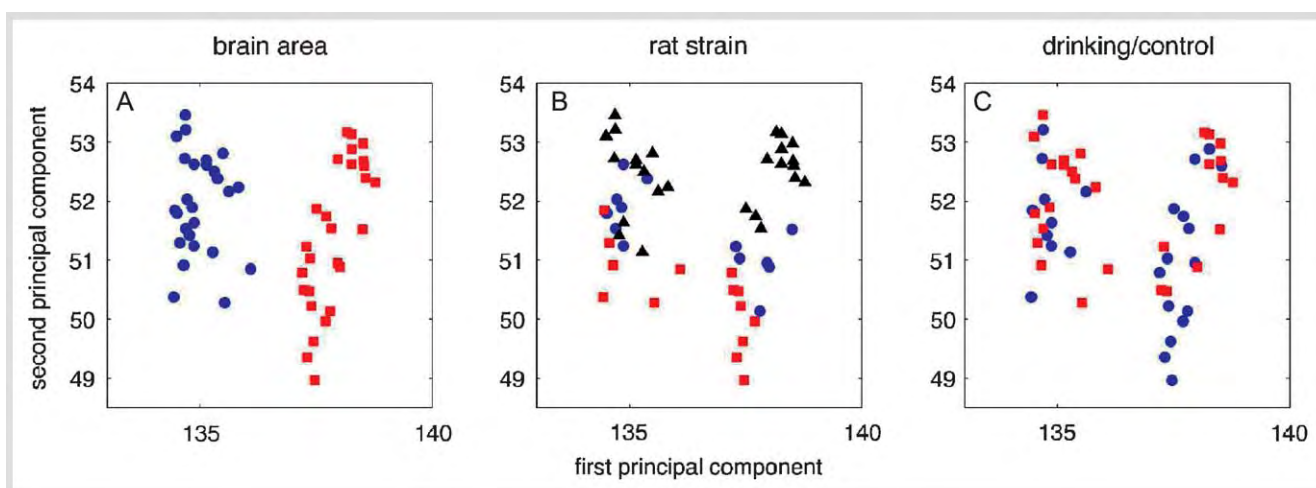

Fig. 2 Data visualization after dimensionality reduction through PCA. Each picture shows the same data points, however, with different colouring denoting the two brain areas (**A**; blue – amygdala, red – caudate putamen), the three rat strains (**B**; blue – AA, red – P, black– H), and the treatment (**C**; blue – drinking, red– control).

levels are found for the three different rat strains, and almost no difference between the drinking and control animals (⊙ **Fig. 1**). The highly significant effect between the two brain regions has been found previously where it turned out that about 2/3rd of the interrogated probe sets had been affected. An important consideration for such a strong effect has been whether or not it was due to a technical bias. We therefore gathered additional information for the 20 top-ranked, region specific probe sets. These show a 2–8 fold difference in intensity values between caudate putamen and amygdala. The 20 probe sets represent 17 genes, 15 of which are represented in the Alan Brain Atlas for mouse (http://mouse.brain-map.org). In each case the difference between the regions was concordant with our microarray data (i.e. *Drd1a, Tnni3 Itpr1, Pcp4l1* and *Tac1* showing higher expression in *CPu, while Gnas, Nnat, Pnck, Hpcal1, Calb2, Crhb2, Gabbr2, Cacna1g, Oprl1* and *Camk2d* are more strongly expressed in amygdala). This speaks against a processing error as the main source for the variance between regions. Furthermore, major

expression differences between brain regions are very reliable phenomena in microarray applications [47].

Similar results are obtained through PCA and hierarchical clustering. For both approaches we used the genes with expression values that differ most with regard to treatment (542 genes with p-values below 0.05). PCA was performed, where the conditions (animal and brain area) were considered as variables and the corresponding values of gene expression as observations. In this way, it is possible to evaluate whether conditions can be separated using the genes; for example, if the gene expression would differ substantially for drinking animals compared to the control animals, the projection of the data onto the plane spanned by the first and second principal component would reveal clearly separated clusters, one containing the data for drinking animals and the other for the controls.

In ⊙ **Fig. 2**, data has been visualized in two dimensions after applying PCA. In ⊙ **Fig. 2A** the colouring distinguishes the two brain areas, the colouring in ⊙ **Fig. 2B** shows the three rat
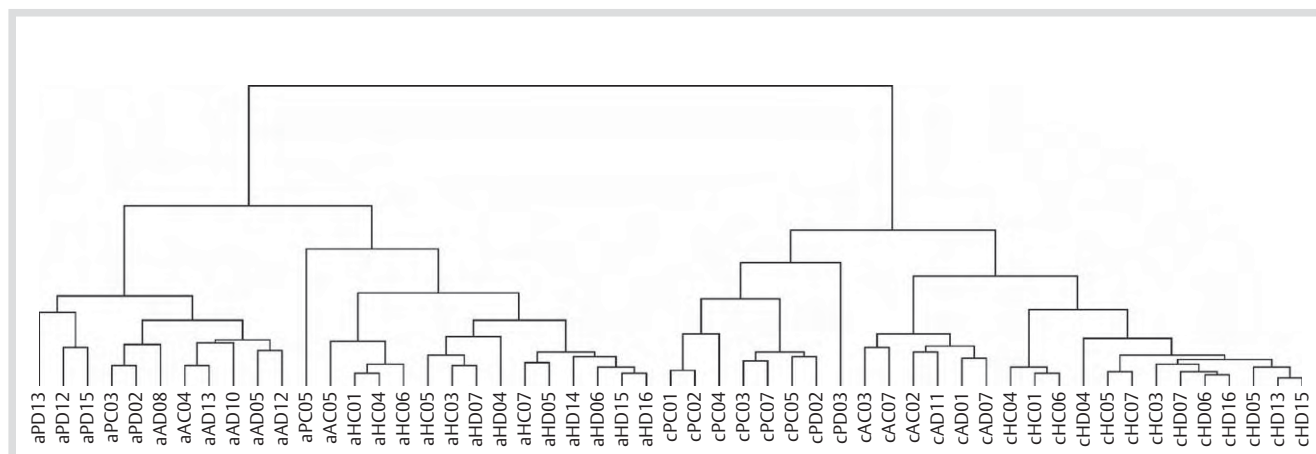
Fig. 3   Dendrogramme representing the results of nearest neighbors clustering. The first letter of the array label denotes the brain area (a = amygdala, c = caudate putamen), the second letter classifies the rat strain (A = AA rats, P = P rats and H = HAD rats), and the third letter the treatment (C = control, D = drinking). The final digits designate individual animals within that group.

strains, and the coloring in ⊙ **Fig. 2C** displays drinking and control conditions.

One can see, that the expression values differ significantly between the two brain areas (⊙ **Fig. 2A**). The separation between the three rat strains is less clear, unless strain "blue" is ignored (⊙ **Fig. 2B**). A comparison of the data points in terms of drinking vs. abstinence does not show any clear separation at all (⊙ **Fig. 2C**).

Hierarchical clustering results in similar findings. ⊙ **Fig. 3** shows the resulting dendrogramme, which represents the similarity of conditions with respect to their corresponding gene expression patterns. Similar conditions are joined early in the process, like cH13 and cH15 on the far right position in ⊙ **Fig. 3**. The largest dissimilarity is found between the two brain areas (a = amygdala; c = caudate putamen), as seen in the dendrogramme where the two groups a* and c* join last. Within each of the two groups, the next split separates the three groups corresponding to the three rat strains (*P* corresponding to P, *A* to AA, and *H* to HAD rats). But then, within each group where brain area and rat strain agree, there is no further clear separation into control and drinking (C/D).

The three approaches confirm similar findings, i. e. that changes in gene expression due to treatment are much less pronounced than expression differences related to brain area or rat strain. Therefore, **the effect of treatment can be compared only within the relatively small groups of animals of the same strain, and within the same brain area**. Basically, ANOVA is very useful to tease apart effects of multiple factors, and to identify those genes having the strongest difference related to our particular factor of interest (low p-values for drinking). This leaves us with a large list of genes and no further information. Any further analysis attempting to reveal function or relationships to drinking of those genes is confounded by the different conditions. For example, there is no *a priori* reason to assume that a particular gene will be affected in the same way by drinking behaviour in different brain regions. In fact, because different functions of brain regions can often be delineated by differences in gene expression [51] and the same gene expressed in different brain regions can result in different behaviour [17], a more reasonable assumption is that drinking behaviour would have a unique influence on gene expression in each brain region, thus requiring separate analysis of brain regions. For each of the

experimental groups in this study, expression patterns in maximally 7 animals can be compared to the patterns of the same gene in no more than 7 other animals. Since alcohol drinking causes only very small changes in gene expression levels, the statistics are not very reliable. Thus, the identification of target genes involved in the development of alcohol addiction is not well supported by this kind of data. Much better results would have been obtained if only animals of one strain had been used, but increasing the number of animals to the *total* number used here.

Identification of target genes for alcoholism

Nevertheless, we aimed at identifying the set of genes showing expression patterns that differ most for the two types of treatment (drinking/control). First we used the 3-way ANOVA performed for the three groups (brain area, rat strain, treatment), and chose the genes with the lowest p-values (all below 0.0015) with respect to treatment. Second, we applied a naïve Bayes classifier, using drinking as the condition of interest, to all 6 344 genes, and selected those genes that were related to drinking with a log ratio greater than 1 (⊙ **Table 1**).

To examine the relationship between these two methods, we checked whether a high significance of the changes in gene expression rates in drinking conditions, (low p-values), corresponded to a high significance in terms of the log ratio. Interestingly, we found no correlation between p-values and log ratios. Furthermore, many of the genes characterized by a high log ratio have very high (non-significant) p-values.

Reasons for the missing correlation of the two measures are 1) the different importance of the data variance for the two approaches, 2) the discretisation needed for the Bayesian approach, and 3) the non-monotonic relationship found by the naïve Bayesian approach. To perform a Bayesian network analysis, the data need to be discretised. We have chosen a discretisation into three states (low, medium, high), whereby the low, medium and high level is defined for each gene separately. Thus, the value of a low expression level in gene *a* can correspond to a high expression level in gene *b*, if the average expression level of *b* is smaller than of *a*. Through the discretisation of the data, we loose information about the variance in the expression levels. A comparison of the variance of the two groups (drinking/control) is, however, the central point of the ANOVA.

Table 1   Genes selected as related to drinking by the naïve Bayes classifier (left) and analysis of variance (right) In the left table all 30 genes selected by the classifier are listed, along with influence scores showing their relationship to drinking (a positive influence (red) means that drinking is related to higher expression of the gene; negative (blue) that drinking is related to lower expression; non-monotonic (NM) means that the relationship is neither positive nor negative), and a log ratio representing a measure of the significance of the relationship (higher ratio means more significance). The 16 genes with log ratios greater than one are shaded in grey: they are considered the most significantly related to drinking, and were used to form a network (for "log-ratio" see: Mat.&Meth.: Naïve Bayes Classifier). The right panel shows the 30 genes with the lowest p-values.

Accession number	Gene name	Influence	Log Ratio	Accession number	Gene name	p value
M10094_9at	RT1-aw2	0.31	5.59	M27217_at	Klk1b21	0.000014
X57169_i_at	Crygd	0.31	3.43	AF001898_at	Aldh1a1	0.000019
rc_AI639520_at	not defined	NM	3.34	X57169_i_at	Crygd	0.000024
D83348_at	Cdh22	0.26	2.38	rc_AI178971_at	globin, alpha	0.000038
rc_AA891695_i_at	Ly6al	−0.37	2.26	X56325mRNA_s_at	Hba-a2	0.000048
AA684537_at	not defined	0.28	2.21	S79304_s_at	Cox6	0.000101
rc_AA894099_at	Vps4a	−0.22	1.82	rc_AA875406_at	not defined	0.000106
AF038043_at	Treh	NM	1.49	rc_AA799801_at	RGD1306959	0.000132
rc_AA900582_at	A2m	NM	1.36	X62327cds_r_at	not defined	0.000135
AF037072_at	Car3	−0.19	1.33	rc_AA860044_at	Ccdcd117	0.000138
rc_AA799745_at	Cdk5rap3	−0.39	1.22	L01793_g_at	Gyg1	0.000142
D86297_at	Alas2	−0.37	1.15	M10094_g_at	RT1-Aw2	0.000387
X60769mRNA_at	Cepb	0.24	1.15	AB017140_g_at	Homer1	0.000389
AF001898_at	Aldh1a1	-0.37	1.02	AB003726_at	Homer1	0.000390
M98826mRNA_at	Phkg1	NM	1.02	D86297_at	Alas2	0.000487
X06150cds_g_at	Gnmt	NM	1.02	rc_AI070295_g_at	Gadd45a	0.000545
X56917_at	Itpka	NM	0.95	rc_AA893074_at	Adora2b	0.000574
rc_AA860044_at	Ccdc117	−0.37	0.9	rc_AA849038_at	Rpl31	0.000654
J04791_s_at	Odc1	NM	0.82	rc_AI232078_at	Ltbp1	0.000678
X54531mRNA_at	Dnm1	NM	0.82	X13722_at	Ldlr	0.000791
rc_AI229421_at	Mapkapk3	NM	0.79	S69383_at	Alox15	0.000891
rc_AA892561_at	RGD1309534	−0.35	0.7	AJ005023_at	RT1-A3	0.000899
X13722_at	Ldlr	−0.28	0.54	rc_AA892799_i_at	Grhpr	0.000959
X62146cds_g_at	Rpl11	NM	0.54	X02322Poly_A_Site.1_s_at	not defined	0.001046
rc_AA892863_at	Mtch2	−0.32	0.42	rc_AA799678_s_at	Egln3	0.001085
S55427_s_at	Pmp22	−0.28	0.39	rc_AA859645_at	Atrn	0.001189
rc_AA875563_g_at	Rcn1	NM	0.36	U53858_at	Capn1	0.001241
rc_AI104679_s_at	Ndufc1	0.2	0.36	rc_AA892551_i_at	not defined	0.001386
S83194_s_at	Camkk1	0.2	0.36	rc_AA892647_at	LOC684887	0.001394

NM = non-monotonic

In **Fig. 4**, we illustrate the different information gained from the two types of analysis by showing expression values of exemplarily selected genes for the control and drinking experiments. **Fig. 4A** shows the gene with the lowest p-value, while **Fig. 4B** shows the gene with the highest log ratio. It can be seen that the expression values of the gene in **Fig. 4B** scatter much stronger than the values of the gene in **Fig. 4A**. Thus the ANOVA favours data where the variance of the expression values in the two groups (control/drinking) is small, while the Bayesian approach has no relation to variance in the data. It is visible, however, that the gene expression differs for both cases between the two groups. For the p-value selected gene, expression is higher in control (non-drinking) animals; for the naïve Bayes selected gene, expression in drinking animals is less variable and tends to be high, while expression in control animals is more evenly spread across the range.
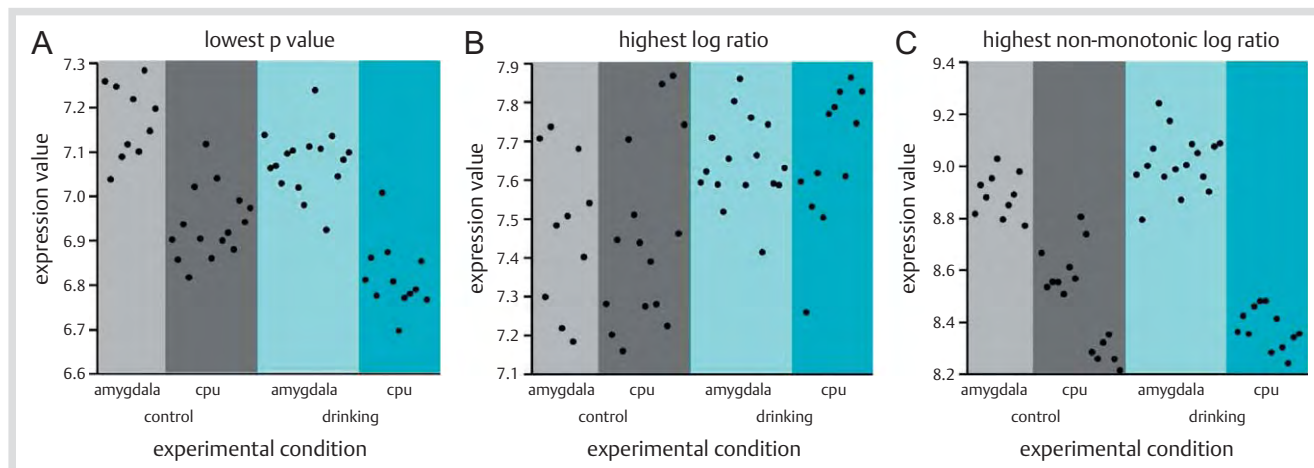
In **Fig. 4C**, gene expression values of the gene with the highest log ratio having a non-monotonic dependence on drinking are shown. Here, the expression values show low, medium and high values in the control group, while in the drinking group the medium expression values are lacking completely. The ability of finding such non-monotonic relationships, as well, is a great advantage of the Bayesian approach. In contrast, analysis of variance can detect only purely monotonic changes.

The conclusions from this study are that both approaches have certain advantages and disadvantages. In contrast to the ANOVA, the Bayesian approach has the disadvantage that it looses information contained in the original (non-descretised) values such as level of variance. On the other hand, this approach can also handle non-monotonic changes, which is not possible through analysis of variance. Therefore, it makes a lot of sense to combine information from both to identify interesting genes.

Gene interactions

Beyond creating lists of genes that act differently across drinking versus non-drinking, building networks enables us to provide a context surrounding these differentially expressed genes. Because the size of the BN search space rises super-exponentially with the number of nodes, it is preferable to have a smaller set of nodes to work with. We selected genes to include in a BN in two ways, making two sets of BN analyses. First, we made use of the ANOVA p-values showing the influence of the drinking/ non-drinking conditions: we performed a network analysis of the 30 genes with the lowest p-values. Second, we used the genes selected by the naïve Bayes classifier with a log ratio greater than 1.

For each set of genes, we ran the BN search 10 times in order to evaluate the consistency of the search. For example, **Fig. 5** shows overlays of networks resulting from 10 different searches.

Fig. 4 Gene expression values for the genes with the lowest p-value (**A**), highest log ratio (**B**) and highest non-monotonic log ratio (**C**). The expression values for the control animals are plotted on the left, for the drinking animals on the right.

One can see that most links exist for all 10 searches, a few were found in a smaller number, and there are two links that were only found in one search.

○ **Fig. 5** also illustrates some general features of Bayesian networks relevant to finding gene interactions in a data set. First, because BNs work by using heuristic search, the same data set does not always produce the same answer. Therefore, it is important to run multiple searches as done here. Those links that reappear consistently across searches are the ones in which to have the most confidence. Hence, the success of the overall search can be assessed by appearance of similar answers across searches: in case of inconsistent links, it would be necessary to either collect a consensus over more networks, look at more networks in each search, and/or start with higher quality data. Moreover, because link direction in the network does not correspond to causation, the direction of links can be variable across networks. For example, all 10 searches found the interaction between RT1-Aw2 and crystallin, gamma D (Crygd). However, they were equally split on the direction of interaction. This does not mean that there is some confusion about the directionality of this link. It only means that the statistical dependence is balanced sufficiently so that the direction does not matter. Correspondingly, when all 10 searches reveal the same direction, this does not say anything about causation: link direction is generally forced by considerations of combinatoric influences (all links directed toward a child from two parents would suggest a combinatoric relationship, where both parents are needed to determine the child's value; all links directed away from a parent to multiple children would suggest, that there is no combinatoric relationship among the children relevant to their parent).

Thus, when assembling the network of interactions in form of a summary from a dataset, we only consider links found in all 10 searches, but disregard the direction of links. ○ **Fig. 6 and 7** show these summaries of the 10 searches for the p-value selected and naïve Bayes selected genes, respectively.

We see that both sets of genes revealed a number of interactions among them. However, the genes selected using the naïve Bayes method had more interactions with the drinking node. This makes sense, because we chose these genes specifically because they scored high with the log ratio for being related to drinking, and the BN search uses a scoring method based on the same basis to determine the links.

Both of these networks enhance our selected gene lists by providing further information, such as which genes may interact with each other and which serve as intermediaries between drinking and other genes.

Discussion
▼
Biology and Molecular Networks
#### Haemoglobin-related subnetwork
The present results confirm some previous findings, but also add new twists to our understanding of molecular mechanisms of alcohol dependence. We have described recently the tentative involvement of downregulated haemoglobin transcription [15]. Here, we see a direct connection of globin mRNA expression with alcohol drinking (○ **Fig. 6**). Globin mRNA has been shown to be expressed by neurons [44]. Moreover, haemoglobin-alpha and delta aminolevulinate synthase (Alas2) (○ **Fig. 7**), that were included in that hypothetical network, have been identified in this report. Therefore, networks composed of molecules associated with haeme synthesis, haemorphins and circadian rhythms appear to be affected by ethanol. Whilst this aspect is no longer surprising, it is good to find another "expected" transcript, as well: aldehyde dehydrogenase (○ **Fig. 7**). The role of its substrate acetaldehyde that may accumulate in brain upon occurrence of genetic variants or its inhibition by drugs has been reviewed just recently [21].

Immunoregulatory Subnetwork
By contrast, additional transcripts raise questions as to their contribution to addiction and encourage the development of new hypotheses. Some transcripts appear to constitute a connection to the immune system [33], such as RT1-Aw2, A2m, Alox15, Ldlr, and Ly6al. Rt1-Aw2 is the heavy chain of MHC class I, located on plasma membranes of antigen processing and presenting cells, whereas lymphocyte antigen 6 complex (Ly6al) are molecules expressed by lymphocytes. Arachidonate 15-lipoxygenase (Alox15) is involved in eicosanoid synthesis and other lipid metabolizing pathways, which the LDL-receptor (Ldlr) may be associated to, as well. LDL-R-related protein plays an important role in the clearance of plasma-activated alpha 2-macroglobulin [24]. Alpha 2-macroglobulin (A2m) has been
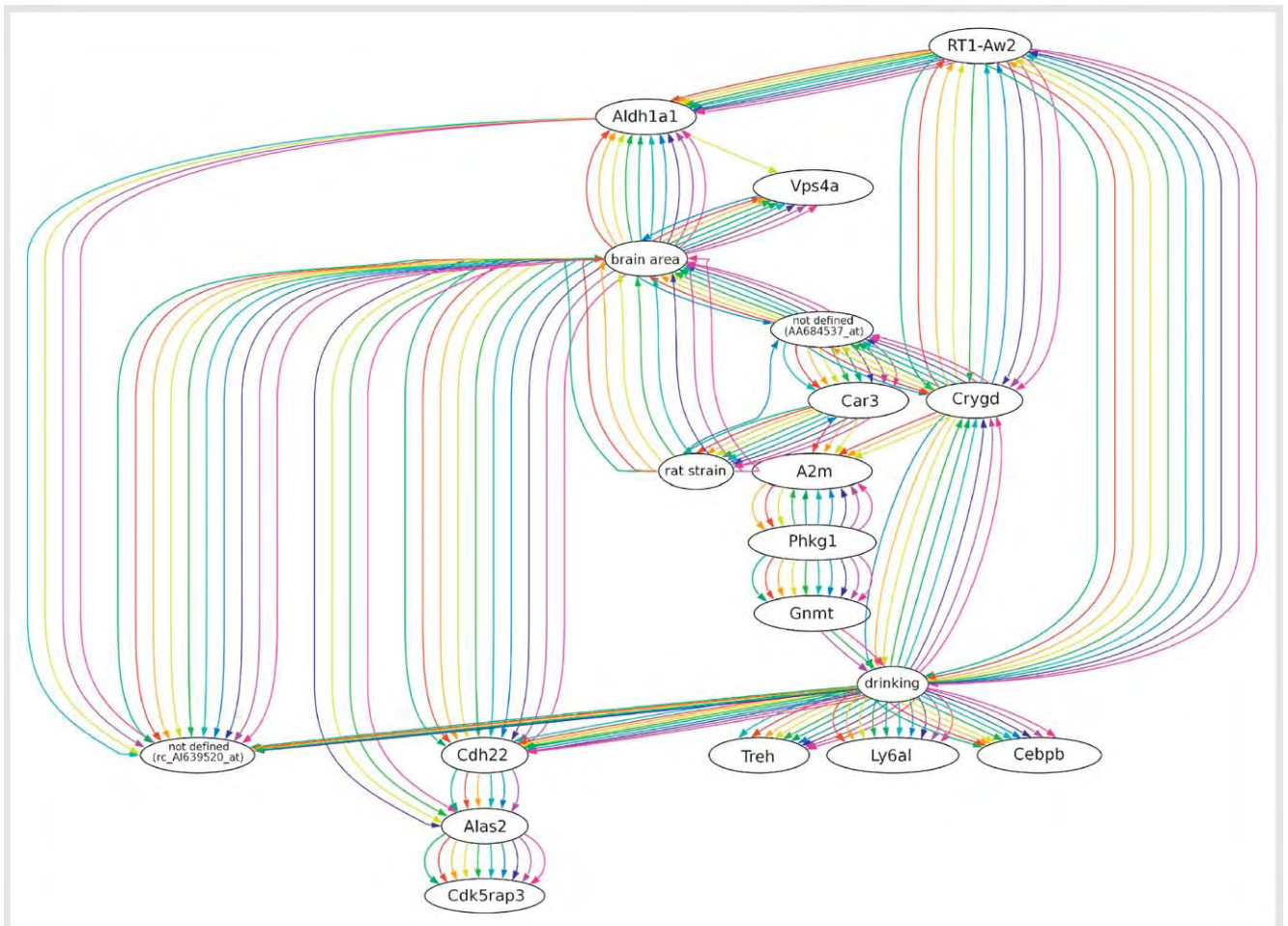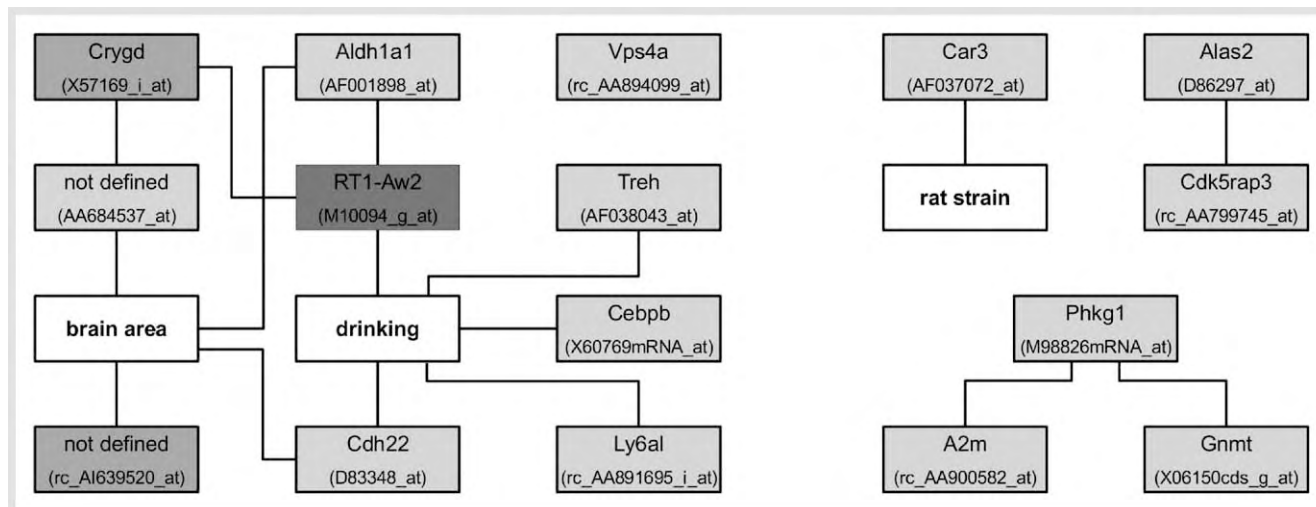
Fig. 5 Links found in 10 individual BN searches using naïve Bayes selected genes. Each search is represented by links of one colour.
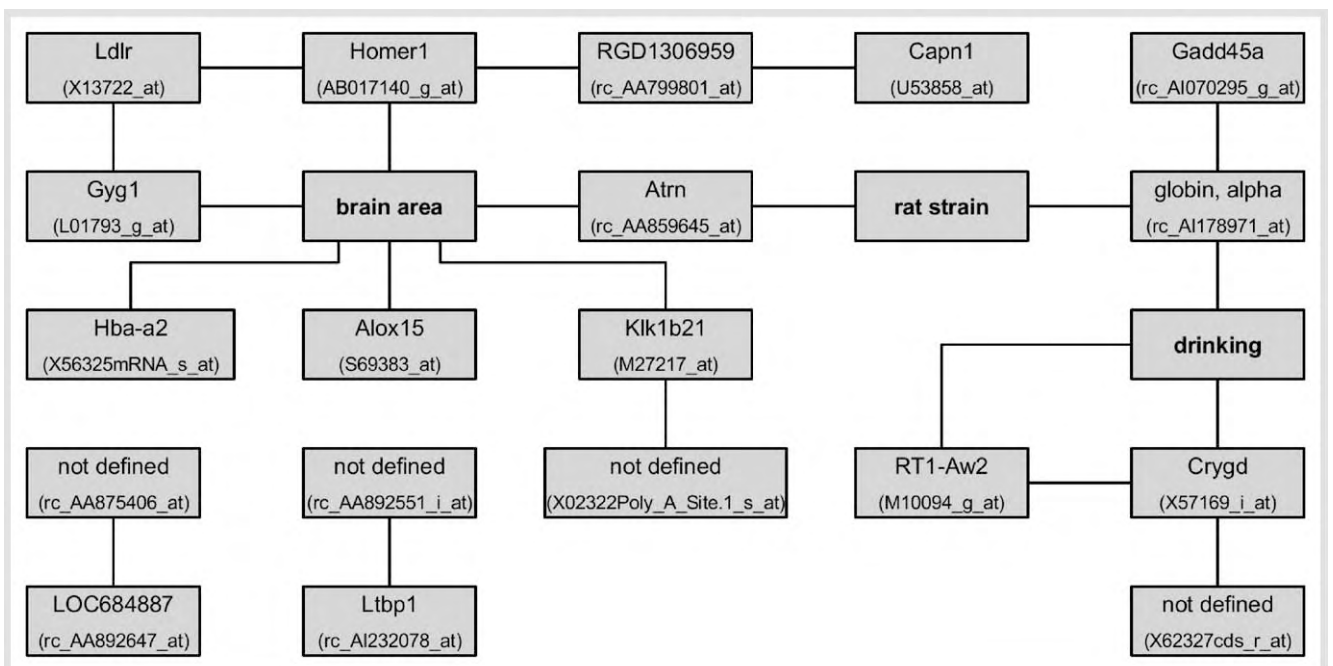
Fig. 6 Links found in all 10 searches of p-value selected genes. Nodes between which all ten searches found a link, in either direction, are shown. Nodes which have no links are omitted from this picture.

Fig. 7 Links found in all 10 searches of naïve Bayes selected genes. Nodes between which all ten searches found a link, in either direction, are shown. The nodes are shaded to represent their naïve Bayes scores with respect to drinking: darker colour means more significant score.

described as an important marker for Alzheimer's disease [31] and has been known for a long time as an acute phase protein, hence, is involved in early phases of the immune response [6]. By thinking also of the apoptosis-related transcript of Gadd45a, one can conclude that there is a strong indication of ongoing immune-related processes in brains of those alcohol-drinking animals.

Cebp beta is a transcription factor that plays an important role during development of the CNS. Multiple gene targets have been identified that may be induced or repressed by this molecule [29]. Amongst others, it is involved in the regulation of immune-related genes, such as IL-1, IL-6, IL-8, TNF-alpha, MIP-1-alpha, chemokine receptor 5 (CCR5) and cyclooxygenease-2 (Cox-2) [42, 52]. In terms of alcoholism, it may be important to mention that it influences genes of the dopamine signalling pathways as shown in striatal neurons [32].

Cytoskeletal Subnetwork

Cadherin22 (Cdh22) is a typical cell-cell adhesion molecule expressed at neuronal pre- and postsynaptic sites [53]. Its interactions with beta-catenin and F-actin are well documented. Cadherin22 expression appears to strengthen synapse formation and maintenance [22]. In this way, the cadherins are critically involved in dynamic molecular networks communicating between structural components of pre- and postsynaptic elements. Apart from that, Cdh13 has been found to be affected by alcohol in a genome-wide association study, that has just recently been published [48].

Reactive, Compensatory Network

In this regard, the expression of trehalase (Treh) appears to be of special interest. The enzyme degrades the disaccharide trehalose to two glucose molecules. Trehalose is supplied by nutrition. Very often, it has been used in vitro as a cryopreservative to prevent denaturation of proteins and damage to cells or even to whole organs upon deep-freezing [25]. Along with alpha-globin, trehalose has been found associated with beta-amyloid plaques in animal models of Alzheimer's disease and has been shown to exert inhibitory effects on huntingtin and A-beta peptide aggregates [1], and reduced aggregate formation in oculopharyngeal muscular dystrophy [43]. Downregulation of trehalase in AA rats

could be interpreted as a beneficial, compensatory response to the adverse effects of alcohol on cellular and molecular structures. However, this is clearly not more than one little piece in a mosaic of molecular networks involved in compensatory mechanisms.

Tentative Networks

Unfortunately, there is presently no biological explanation for a tentative connection between RT1-Aw2 and crystallin-gamma (Crygd). Crystallin-gamma, like crystallin-alpha, is a major protein of the eye lens [2]. The eye, like the CNS has been considered as an immune-privileged organ and, therefore, may maintain comparable, and very special interactions with the immune system [4]. Crystallin-gamma reportedly plays a role in differentiation of the epidermis [2]. Regulation of crystallin-gamma D transcription during development has been thoroughly studied by Klok et al. [30]. Moreover, the molecule appears to influence Na,K-ATPase and, hence the resting potential of neurons [38]. This gene pair may be viewed as a good example that mathematical modelling studies on laboratory data can provide hints in which direction it may be rewarding to further extend laboratory investigations. It, however, also highlights a general problem inherent to gene lists: the cut-off at certain thresholds for sake of statistics. Genes above that level may, hence, display no evident relationship to each other but, nevertheless, may be connected through genes hidden in statistical noise. We assume the existence of an underlying network, but are unable to recognize it.

Mathematical Modelling

Principal component analysis (PCA) based on ANOVA criteria revealed three results: a highly significant correlation (differential expression) between the two brain regions under investigation, a moderately high correlation between rat strains and a lower correlation between drinking and non-drinking conditions. These findings highlight the well known requirement to study gene expression in well-circumscribed functional brain units (such as N. accumbens shell vs. core) – ideally in single cells – rather than including regions encompassing multiple units, like striatum or hippocampus.

The differential gene expression observed in the different rat strains is not unexpected but raises questions about a generalized concept of "alcohol-dependent" gene regulation. Those strain differences may indicate distinct susceptibilities and, hence, distinct responses of the genetic "background" to ethanol exposure. This makes the search for "alcohol-responsive" genes more difficult, since those genes may show strain-dependent differences. Moreover, in this light, translation of those data to therapeutic strategies in the clinical setting requires even more caution.

The comparison of p-values of genes with the naïve Bayesian approach leads to interesting results. The genes with a high log-ratio identified by the Bayesian approach frequently have very high (non-significant) p-values resulting in gene lists with little overlapping genes. Admittedly, there is loss of information about the variance of data in the Bayesian approach. This, however, may even be advantageous compared to ANOVA-based computations, since it reduces noise and includes non-monotonic changes. This has been nicely illustrated in ○ **Fig. 4** on the right compared to the left panel in that figure. Bayesian networks tend to be highly conservative, providing a few, strong interactions at the expense of potentially missing weaker ones and thus leading to many "false negatives"; in contrast, ANOVAs use of parametric statistics enables them to pull out even weak relationships from highly factorized data, as we have here. Thus, ANOVAs may be more beneficial for providing an overview of the effects of a treatment, while BNs may be more useful for pulling out a handful of promising targets to further investigate.

Furthermore, when using those two distinct approaches to search for genes directly related to drinking in functional relationships, eventually three genes emerged from the p-value selected genes and five genes surfaced from the Bayesian approach. Since most of those genes are not the same, the combined approaches increased the number of candidate genes and, consequently, the likelihood of identifying pivotal ones.

Admittedly, the statistical power could have been better if less conditions or more animals per condition had been chosen. The differences in gene expression between the strains, however, were somehow surprising, because a set of genes specifically regulated by ethanol and irrespective of the strain was expected.

Finally, despite the low *n*-numbers, no experiments have been done on a **time-scale**. The Bayesian and other mathematical approaches, however, often rely on those kinds of data. Although more experiments are needed, time-course studies would increase the statistical power and more precise biological conclusions, especially about the development of a disease, could be drawn. Examples using Bayesian approaches in microarray studies can already been found in the literature [23, 36, 40].

Altogether, it can be concluded, that mathematical approaches using biological datasets are becoming more and more indispensable to extend existing views of biological interactions and to reveal new aspects that result in new hypotheses upon mechanisms of disease-related molecular networks.

Acknowledgments/Disclosure
▼

Affiliations
[1] Center for Modeling and Simulation in the Biosciences, University of Heidelberg, Heidelberg, Germany
[2] University of St. Andrews, School of Biology, St Andrews, Fife, UK
[3] Interdisciplinary Centre for Mathematical and Computational Modelling, University of Warsaw, Warsaw, Poland
[4] Laboratory for Plant Molecular Biology, University of Warsaw, Warsaw, Poland
[5] Central Institute of Mental Health, Department of Psychopharmacology, Mannheim, Germany
[6] Dipartimento di Medicina Sperimentale e Sanità Pubblica, Università di Camerino, Camerino, Italy
[7] Department of Statistics, Virginia Commonwealth University, Richmond, VA, USA

References
1 *Béranger F, Crozet C, Goldsborough A et al.* Trehalose impairs aggregation of PrPSc molecules and protects prion-infected cells against oxidative damage. Biochem Biophys Res Commun 2008; 374 (1): 44–48
2 *Bhat SP.* Crystallins, genes and cataract. Prog Drug Res 2003; 60: 205–262
3 *Butte A, Kohane I.* Mutual information revelance networks: functional genomic clustering using pairwise entropy measurements. Pac Symp Biocomput 2000; 5: 418–429
4 *Charukamnoetkanok P, Fukushima A, Whitcup SM et al.* Expression of ocular autoantigens in the mouse thymus. Curr Eye Res 1998; 17 (8): 788–792
5 *Chen T, He HL, Church GM.* Modeling gene expression with differential equations. Pacific Symposium Biocomputing 1999; 4: 29–40
6 *Chu CT, Pizzo SV.* Alpha 2-Macroglobulin, complement, and biologic defense: antigens, growth factors, microbial proteases, and receptor ligation. Lab Invest 1994; 71 (6): 792–812
7 *DeRisi J, Penland L, Brown PO et al.* Use of a cDNA microarray to analyse gene expression patterns in human cancer. Nat Genet 1996; 14 (4): 457–460
8 *DeRisi JL, Iyer VR, Brown PO.* Exploring the metabolic and genetic control of gene expression on a genomic scale. Science 1997; 278 (5338): 680–686
9 *Eriksson K.* The estimation of heritability for the self-selection of alcohol in the albino rat. Ann Med Exp Biol Fenn 1969; 47: 172–174
10 *Friedman N, Geiger D, Goldschmidt M.* Bayesian network classifiers. Machine Learning 1997; 29: 131–163
11 *Friedman N, Murphy K, Russell S.* Learning the structure of dynamic probabilistic networks. Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence. San Francisco, CA; 1998, Morgan Kaufmann 139–147
12 *Friedman N, Goldszmidt M, Wyner A.* Data analysis with Bayesian networks: a bootstrap approach. Proc Fifteenth Conf on Uncertainty in Artificial Intelligence (UAI) 1999; 206–215
13 *Friedman N, Linial M, Nachman I et al.* Using Bayesian networks to analyze expression data. J Comp Bio 2000; 7: 601–620
14 *Friedman N.* Inferring cellular networks using probabilistic graphical models. Science 2004; 303: 799–805
15 *Gebicke-Haerter PJ, Tretter F.* The Systems view in addiction research. Addiction Biol 2008; 13: 449–454
16 *Gentleman RC, Carey VJ, Bates DM et al.* Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 2004; 5 (10): R80
17 *Hammock EA.* Gene regulation as a modulator of social preference in voles. Adv Genet 2007; 59: 107–127
18 *Hartemink AJ, Gifford D, Jaakkola T et al.* Combining location and expression data for principled discovery of genetic regulatory network models. Pac Symp Biocomput 2002; 7: 437–449
19 *Heckerman D, Geiger D, Chickering DM.* Learning Bayesian networks: The combination of knowledge and statistical data. Mach Learn 1995; 20: 197–243

20 *Heckerman D*, *Mamdani A*, *Wellman M*. Real-world applications of bayesian networks. Communications of the ACM 1995b; 38 (3): 24–30

21 *Hipólito L*, *Sánchez MJ*, *Polache A et al*. Brain metabolism of ethanol and alcoholism: an update. Curr Drug Metab 2007; 8 (7): 716–727

22 *Huntley GW*, *Gil O*, *Bozdagi O*. The cadherin family of cell adhesion molecules: multiple roles in synaptic plasticity. Neuroscientist 2002; 8 (3): 221–233

23 *Husmeier D*, *Werhli AV*. Bayesian integration of biological prior knowledge into the reconstruction of gene regulatory networks with Bayesian networks. Comput Syst Bioinformatics Conf 2007; 6: 85–95

24 *Hussain MM*, *Strickland DK*, *Bakillah A*. mammalian low-density lipoprotein receptor family. Annu Rev Nutr 1999; 19: 141–172

25 *Jain NK*, *Roy I*. Effect of trehalose on protein structure. Protein Sci 2009; 18 (1): 24–36

26 *Jolliffe IT*. Principal Component Analysis, Springer Series in Statistics 2nd ed. Springer, NY; 2002

27 *Kaderali L*, *Radde N*. Inferring Gene Regulatory Networks from Gene Expression Data In: Kelemen A, Abraham A, Chen Y, Eds. "Computational Intelligence in Bioinformatics. Studies in Computational Intelligence". Springer-Verlag; 2008

28 *Kauffman S*. Metabolic stability and epigenesis in randomly constructed genetic nets. J Theor Biol 1969; 22: 437–467

29 *Kfoury N*, *Kapatos G*. Identification of neuronal target genes for CCAAT/ Enhancer binding proteins. Mol Cell Neurosci 2008, [Epub ahead of print]

30 *Klok EJ*, *Genesen ST van*, *Civil A et al*. Regulation of expression within a gene family. The case of the rat gammaB- and gammaD-crystallin promoters. J Biol Chem 1998; 273 (27): 17206–17215

31 *Kovacs DM*. alpha2-macroglobulin in late-onset Alzheimer's disease. Exp Gerontol 2000; 35 (4): 473–479

32 *Kovács KA*, *Steinmann M*, *Magistretti PJ et al*. C/EBPbeta couples dopamine signalling to substance P precursor gene expression in striatal neurones. J Neurochem 2006; 98 (5): 1390–1399

33 *Kuehn BM*. Scientists probe immune system's role in brain function and neurological disease. JAMA 2008; 299 (6): 619–620

34 *Langley P*, *Iba W*, *Thompson K*. An analysis of Bayesian classiers. In Proceedings, Tenth National Conference on Articial Intelligence 1992. Menlo Park, CA: AAAI Press, 223–228

35 *Li T-K*, *Lumeng L*, *Doolittle DP*. Selective breeding for alcohol preference and associated responses. Behav Genet 1993; 23: 163–170

36 *Li P*, *Zhang C*, *Perkins EJ et al*. Comparison of probabilistic Boolean network and dynamic Bayesian network approaches for inferring gene regulatory networks. BMC Bioinformatics 2007; 8 Suppl 7: S13

37 *Lipshutz RJ*, *Morris D*, *Chee M et al*. Using oligonucleotide probe arrays to access genetic diversity. Biotechniques 1995; 19 (3): 442–447

38 *Moseley A*, *Graw J*, *Delamere NA*. Altered Na,K-ATPase pattern in gamma-crystallin mutant mice. Invest Ophthalmol Vis Sci 2002; 43 (5): 1517–1519

39 *Ott S*, *Imoto S*, *Miyano S*. Finding optimal models for small gene networks. Pac Symp Biocomput 2004; 9: 557–567

40 *Polanski A*, *Polanska J*, *Jarzab M et al*. Application of Bayesian networks for inferring cause-effect relations from gene expression profiles of cancer versus normal cells. Math Biosci 2007; 209 (2): 528–546 Epub 2007 Mar 27

41 *Reimers M*, *Heilig M*, *Sommer WH*. Gene discovery in neuropharmacological and behavioral studies using Affymetrix microarray data. Methods 2005; 37 (3): 219–228

42 *Rosati M*, *Valentin A*, *Patenaude DJ et al*. CCAAT-Enhancer-Binding Protein-beta (C/EBP-beta) activates CCR5 Promoter: Increased C/EBP-beta and CCR5 in T Lymphocytes from HIV-1-Infected Individuals. J Immunol 2001; 167: 1654–1662

43 *Sarkar S*, *Davies JE*, *Huang Z et al*. Trehalose, a novel mTOR-independent autophagy enhancer, accelerates the clearance of mutant huntingtin and alpha-synuclein. J Biol Chem. 2007; 282 (8): 5641–5652

44 *Schelshorn DW*, *Schneider A*, *Kuschinsky W et al*. Expression of hemoglobin in rodent neurons. J Cereb Blood Flow Metab 2008

45 *Schena M*, *Shalon D*, *Davis RW et al*. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 1995; 270: 467–470

46 *Simon I*, *Siegfried Z*, *Ernst J et al*. Combined static and dynamic analysis for determining the quality of time-series expression profiles. Nature Biotechnology 2005; 23 (12): 1503–1508

47 *Sommer WH*, *Arlinde C*, *Heilig M*. The search for candidate genes of alcoholism: Evidence from expression profiling studies. Addiction Biology 2005; 10 (1): 71–80

48 *Treutlein J*, *Cichon S*, *Ridinger M et al*. Genome-wide association study of alcohol dependence. Arch Gen Psych [in press] 2009

49 *Vengeliene V*, *Siegmund S*, *Singer MV et al*. A comparative study on alcohol-preferring rat lines: effects of deprivation and stress phases on voluntary alcohol intake. Alcohol Clin Exp Res 2003; 27 (7): 1048–1054

50 *Vengeliene V*, *Leonardi-Essmann F*, *Perreau-Lenz S et al*. The dopamine D3 receptor plays an essential role in alcohol-seeking and relapse. FASEB J 2006; 20 (13): 2223–2233

51 *Wada K*, *Sakaguchi H*, *Jarvis ED et al*. Differential expression of glutamate receptors in avian neural pathways for learned vocalization. J Comp Neurol 2004; 476: 44–64

52 *Wu KK*. Aspirin and other cyclooxygenase inhibitors: new therapeutic insights. Semin Vasc Med 2003; 3 (2): 107–112

53 *Yamagata M*, *Sanes JR*, *Weiner JA*. adhesion molecules. Curr Opin Cell Biol 2003; 15 (5): 621–632

54 *Yu J*, *Smith VA*, *Wang PP et al*. Advances to Bayesian network inference for generating causal networks from observational biological data. Bioinformatics 2004; 20: 3594–3603

55 *Yu J*. Developing Bayesian network inference algorithms to predict causal functional pathways in biological systems [PhD Thesis]05 Duke University, Durham, NC