BIOINFORMATICS

Vol. 00 no. 00 2011
Pages 1–9

# From Sets to Graphs: Towards a Realistic Enrichment Analysis of Transcriptomic Systems

Ludwig Geistlinger[1]*, Gergely Csaba[1], Robert Küffner[1], Nicola Mulder[2] and Ralf Zimmer[1]*

[1]Institute for Informatics, Ludwig-Maximilians-Universität Münchchen, Amalienstr. 17, 80333 München, Germany
[2]Department of Clinical Laboratory Sciences, Institute of Infectious Disease and Molecular Medicine, University of Cape Town, Anzio Road, Observatory, Cape Town 7925, South Africa

**ABSTRACT**

**Motivation:** Current gene set enrichment approaches do not take interactions and associations between set members into account. Mutual activation and inhibition causing positive and negative correlation among set members are thus neglected. As a consequence, inconsistent regulations and contextless expression changes are reported and, thus, the biological interpretation of the result is impeded.

**Results:** We analyzed established gene set enrichment methods and their result sets in a large-scale investigation of 1000 expression datasets. The reported statistically significant gene sets exhibit only average consistency between the observed patterns of differential expression and known regulatory interactions. We present *Gene Graph Enrichment Analysis* (GGEA) to detect consistently and coherently enriched gene sets, based on prior knowledge derived from directed gene regulatory networks (GRNs). Firstly, GGEA improves the concordance of pairwise regulation with individual expression changes in respective pairs of regulating and regulated genes, compared to set enrichment methods. Secondly, GGEA yields result sets where a large fraction of relevant expression changes can be explained by nearby regulators, such as transcription factors, again improving on set based methods. Thirdly, we demonstrate in additional case studies that GGEA can be applied to human regulatory pathways, where it sensitively detects very specific regulation processes, which are altered in tumors of the central nervous system. GGEA significantly increases the detection of gene sets where measured positively or negatively correlated expression patterns coincide with directed inducing or repressing relationships thus facilitating further interpretation of gene expression data.

**Availability:** The method and accompanying visualization capabilities have been bundled into an `R` package and tied to a grahical user interface, the `Galaxy` workflow environment, that is running as a web server.

**Contact:** {Ludwig.Geistlinger, Ralf.Zimmer}@bio.ifi.lmu.de

## 1 INTRODUCTION

Transcriptomic studies measure gene expression in different conditions. Striking genes, which are differentially regulated between the conditions, are of primary interest and investigated for common features and membership in group of genes, which have the same function or belong to the same biochemical pathway.

A first impression of similar behavior of genes can be achieved via clustering of genes (Eisen *et al*., 1998). The usually more effective overrepresentation analysis (ORA) tests the overlap of a predefined group of genes and the set of differentially expressed genes assuming the hypergeometrical distribution under the null hypothesis (Breitling *et al*., 2004). The method is widely accepted and has been subject to modifications of diverse visual and model related features (see Khatri and Draghici, 2005, for an overview), though the basic statistical principle remained unchanged. However, Goeman and Bühlmann (2007) criticize that the sampling procedure of ORA is statistically invalid and leads to a hazardous interpretation of the resulting $p$-value. Furthermore, the concentration on the usually small group of significantly differentially expressed genes, compared to the set of all the other, usually thousands of genes analysed in the study that are ignored, is not suitable for an investigation on a global scale.

Both points of criticism are resolved in *Gene Set Enrichment Analysis* (GSEA) as it uses a valid sampling procedure and computes over the whole scope of genes (Subramanian *et al*., 2005). A Kolmogorov-Smirnov test statistic is applied to test whether the ranks of the $p$-values of the genes in the gene set can be a sample from a uniform distribution. Several modifications of GSEA have been published (see Dinu *et al*., 2009, for an overview).

Though ORA and GSEA are convenient in the analysis of genes that are independently expressed, a serious problem arises when these methods are applied to gene set definitions extracted from regulatory networks and metabolic pathways. The assumption of independence among set members does not hold anymore; genes are found to be correlated due to mechanisms of co-regulation and co-expression. Initial steps to deal with that problem include implicit accounting for the correlation structure (e.g. Barry *et al*., 2005) and integration of network topology of undirected interaction networks (e.g. Ulitsky and Shamir, 2007). Based on these first efforts, Liu *et al*. (2007) have proposed *Gene Network Enrichment Analysis* (GNEA) that uses ORA to test for overrepresentation of gene sets in transcriptionally affected subnetworks of a global interaction network.

As the sign of gene expression changes and the direction of regulatory interactions are so far not taken into account, substantial features of the data are still ignored and the dynamics of the transcriptomic system are not realistically reflected. Activation and

*to whom correspondence should be addressed

inhibition are essential regulatory mechanisms in the transcriptional machinery of the cell and are causes for up- and down-regulation of particular genes. Although processes like post-translational modification and combinatorial effects between regulatory proteins impair a straightforward causal relationship between regulation and gene expression, it was shown that coexpression is correlated with functional relationships between genes (Lee *et al*, 2004). Additionally, integrative analysis of transcriptome, proteome and interactome data revealed significant correlations between expression profiles and regulatory interaction on the protein level (Jansen *et al*., 2002; Ge *et al*., 2001). Hence, we explain positive correlation in gene expression with activating edges of the transcriptional network. Vice versa, we assume inhibition to cause observed anti-correlation in gene expression patterns. In our following definition of *Gene Graph Enrichment Analysis* (GGEA), we exploit both fundamental regulation types in a novel enrichment framework for signed and directed gene regulatory networks, to judge whether the topology of the network is well fitted by the expression data.

## 2 METHODS

### Gene Graph Enrichment Analysis (GGEA)

Given gene regulatory information, for example extracted from biochemical pathways or a global transcriptional network, a gene set under investigation and gene expression data sampling different conditions, GGEA performs three essential steps (Fig. 1): First, the gene set is mapped onto the underlying regulatory network, yielding an induced subnetwork. That is the affected part of the network, which consists of edges that involve members of the gene set. Second, each edge of the induced network is scored for consistency with the expression data, i.e. the signs of the expression changes of two interaction partners are evaluated for agreement with the regulation type (activation/inhibition) of the link that connects both genes. Third, the edge consistencies are summed up over the induced network, normalized and estimated for significance using a permuation procedure.

*Experimental Setup* In the following, we consider the classical setup of a transcriptomic study. This incorporates a set $G$ of usually several thousand genes $g_i$ $(i = 1, \ldots, n)$ measured for differential expression between two conditions, each represented by a group of samples $S_1 = \{s_1, \ldots, s_k\}$ and $S_2 = \{s_{k+1}, \ldots, s_m\}$, respectively. The function

$$\text{expr} : G \times (S_1 \cup S_2) \to \mathbb{R} \tag{1}$$

returns the expression value for a gene and a sample at a time.

*Measures of Differential Expression* The most intuitive measure for expression changes of a single gene between two conditions is the fold change

$$\text{fc} : G \to \mathbb{R}, \tag{2}$$

defined as the ratio of the estimated expression values of a particular gene in both sample groups

$$\text{fc}(g_i) = \frac{\text{expr}(g_i, S_1)}{\text{expr}(g_i, S_2)}, \tag{3}$$

where $\text{expr}(g, S)$ computes the mean expression level of gene $g$ in condition $S$. We compute $t$-test derived $p$-values to assess the statistical significance of the expression changes (Pan, 2002) and correct them for multiple testing. Both measures are log-transformed

$$\tilde{\text{fc}} := \log_2(\text{fc}), \quad \tilde{p} := -\log_{10}(p), \tag{4}$$

and the significance thresholds $\alpha = -\log(0.05)$ and $\beta = 1$ (two-fold) are used as defaults for $\tilde{p}$ and $\tilde{\text{fc}}$, respectively. Such sharp thresholds are of course quite artificial and discriminate drastically between genes just over and just below $\alpha$ or $\beta$. In addition, noise in the data, such as imprecise and erroneous measurements of gene expression values, has to be expected and to be dealt with. Hence, we divide the range of both measures into two main categories and smooth the borders via introduction of a degree of uncertainty, according to the mathematical concept of *fuzzyfication* (Zadeh, 1963; Windhager and Zimmer, 2008; Windhager *et al*., 2010). For the fold change, we map

$$(\tilde{\text{fc}} < 0,\ \tilde{\text{fc}} > 0) \mapsto (\text{down},\ \text{up}), \tag{5}$$

and compute membership values for both categories via the weighting functions $w : \tilde{\text{fc}} \mapsto [0, 1]$ (displayed in Fig. 2b), resulting in a pair

$$\langle \text{fc} \rangle \ := \ fuzzy(\tilde{\text{fc}}) = \langle\ w_{\text{down}}(\tilde{\text{fc}}),\ w_{\text{up}}(\tilde{\text{fc}})\ \rangle. \tag{6}$$

Analogously, we map $\tilde{p}$, using Fig. 2a, to areas of *low* and *high* significance in the fuzzy concept

$$\langle \text{sig} \rangle \ := \ fuzzy(\tilde{p}) = \langle\ w_{\text{low}}(\tilde{p}),\ w_{\text{high}}(\tilde{p})\ \rangle. \tag{7}$$

For both measures, a third category can optionally be introduced to account for unspecific signals in case of very noisy data. The fold change and $p$-value categories are combined to a single measure of differential expression

$$\text{de} := \langle \text{fc}, \text{sig} \rangle, \tag{8}$$

in order to simultaneously summarize and express whether the transcriptional activity of a particular gene is reduced or enhanced in one sample group, compared to the other.

*Induced Gene Regulatory Networks* Enrichment analysis is the determination of significant gene sets out of a predefined universe of gene sets $U$, s.t. result sets accumulate differentially expressed features of the gene expression data. GGEA uses an *a priori* defined gene regulatory network (GRN), typically extracted from respective databases or compiled from the relevant literature, to introduce and exploit the interdependencies between gene set members. We model a regulatory interaction of the GRN as a transition $t$ (see Fig. 3) with an input place for the regulator and an output place for its target, as well as an associated effect (activation, inhibition) and the direction of the interaction. For a gene set $u \in U$, we construct the *induced* subnetwork

$$\text{GRN}(u) := \{t \in \text{GRN} \mid u \cap (\text{in}(t) \cup \text{out}(t)) \neq \emptyset\}, \tag{9}$$

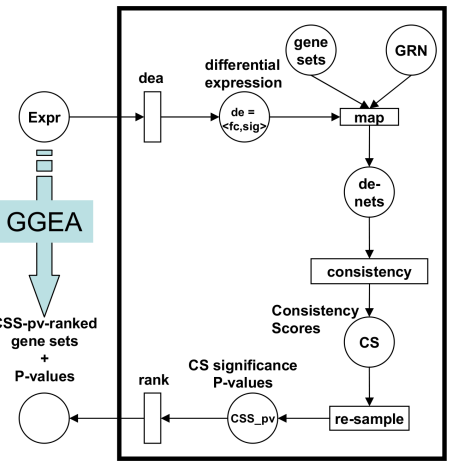
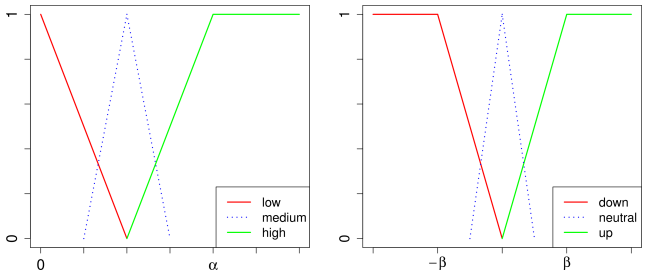



**Fig. 1. Key Steps of GGEA.** Subsequent to differential expression analysis *dea* of expression data *Expr*, yielding fuzzified measures *de* of differential expression, target gene sets are first mapped onto the gene regulatory network (GRN). The *de*-values are assigned to corresponding places in resulting induced nets (*de-nets*). Second, consistency scores are computed for each *de*-net and third, significance of the scores is estimated via re-sampling, and exploited to rank the gene sets.




(a) $p$-value (b) fold change

**Fig. 2. Fuzzyfication of $p$-value and fold change.** Both measures are mapped onto two main categories, each having a membership function to express the uncertainty of the mapping. Additional categories, e.g. a third category *medium* and *neutral*, respectively, can be introduced for a more detailed representation.

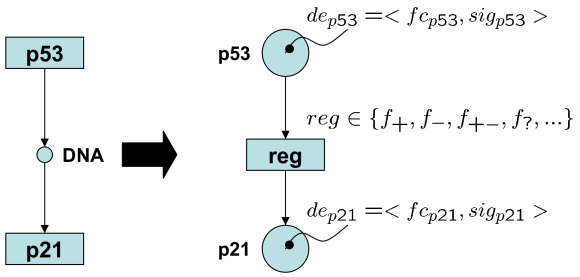



**Fig. 3. Modelling regulatory interactions using PNFL.** Shown is a KEGG style representation of an activation and its transformation into a PNFL transition $f_+$. Tokens of combined fuzzy measures $de$ of differential expression assigned to Petri net (PN) places, represent the regulator and its target. The regulatory effect is defined via a specific fuzzy rule for every effect type of the GRN.

s.t. for each gene $g$ of the gene set $u$ all transitions are extracted, where $g$ is either the regulating or the regulated gene.

*Gene Regulatory Networks as Petri Nets* Petri net models are well established in information theory (see Murata, 1989, for a review) and have been extensively applied to biochemical processes, like metabolic pathways (e.g. Küffner *et al*, 2000) and gene regulatory networks (reviewed in Chaouiya, 2007). Given a GRN under investigation, we construct a corresponding Petri net (PN) having features of fuzzy logic (FL), as it is introduced as PNFL in Küffner *et al*. (2010), and illustrated in Fig. 3. The regulations of the GRN are required to be specified with direction and effect. In our model, regulator (R) and regulated target (RT) are represented via PN places holding tokens of fuzzy values for both fold change (fc) and significance of fc (sig). The variety of regulatory effects occuring in the GRN are defined by specific fuzzy rules $reg \in \{f_+, f_-, f_{+-}, f_?, ...\}$ (Table 1), meaning activation $f_+$, inhibition $f_-$ and dual effects $f_{+-}$. The concept is extendable, e.g. to other effects like interactions of unknown type $f_?$. The fuzzy rules compute output tokens from given input tokens. Thus, consistency between expected (i.e. modeled) behavior and the measured values can be evaluated. Consistency takes the direction of the effect, the amount (fc) and its significance into account and is a straightforward extension of the discrete notion of consistency (e.g. R *up* and $f_+ \implies$ RT *up*). Moreover, it appropriately models noise in the actual experimental measurements.

*Consistency of Regulatory Interactions* The major problem of set enrichment strategies, when applied to GRN-based gene sets, is that they accumulate evidence for differential expression of single genes to estimate the enrichment of the whole set. Interfering and potentially contrary constraints of the underlying GRN are ignored. For example, two significantly up-regulated genes increase the enrichment of the set, even if one gene inhibits the other. For that reason, we introduce the concept of consistency.

**Table 1.** Fuzzy rule set for activation and inhibition.

| | ⟨fc⟩ | | ⟨sig⟩ | |
|---|---|---|---|---|
| | down | up | low | high |
| $f_+$ | down | up | low | high |
| $f_-$ | up | down | low | high |

> **Definition** (*consistency*): A transition of a PNFL is consistent with given expression data, if the measured and the modeled expression of the regulated gene is in agreement. The modeled expression is estimated from the regulatory effect and the expression of the regulator.

Intuitively, consistency for the special case of a simple activating or inhibiting edge requires fold changes for regulator and target of the same or opposite directions, respectively. It is implied for the above example that an up-regulated inhibitor should result in reduced expression of the affected gene.
For the PN constructed above, a *consistent* transition $t$ with fuzzy regulation function $f_t$ between an input place $i$ and an output place $o$ satisfies

$$\mathrm{de}_o \approx f_t(\mathrm{de}_i), \tag{10}$$

i.e. the modeled predicted expression behavior agrees with the actual observed behavior.

*Scoring* To determine if and to which extent $t$ is consistent with the given expression data, we calculate the consistency

$$C(t) := \mathrm{cons}\left(\mathrm{de}_o, f_t(\mathrm{de}_i)\right), \tag{11}$$

where the function $cons$ estimates the (fuzzy) similarity between the predicted and measured token on the output place of transition $t$. Consistency computation is generic, an example implementation of $cons$ incorporates defuzzyfication of the fuzzy values back into real numerical values (Küffner *et al*., 2010) and taking their reciprocal absolute difference. We compute the raw GGEA consistency score $S$ for the subnetwork GRN$(u)$, induced by the gene set $u \in U$, via summation over the consistencies of all transitions $T_u$ of GRN$(u)$

$$S := \sum_{t \in T_u} C(t), \tag{12}$$

and normalize it by the number of transitions $|T_u|$

$$\bar{S} := \frac{S}{|T_u|}, \tag{13}$$

to adjust for the size of GRN$(u)$.

*Significance and Ranking* According to the recommendations of Goeman and Bühlmann (2007) and Gatti *et al*. (2010), statistical significance of the consistency score is estimated via a permutation approach based on subject sampling, which is defined in a self-contained way:

1. Permute group assignment of samples $N$ times.
2. Recalculate differential expression measures for each permutation.
3. Recalculate consistency score for each permutation.
4. Find the consistency $p$-value as the proportion of permutation scores that are larger than the observed score.

We compute the consistency $p$-value for each gene set $u \in U$ and rank the gene sets by the adjusted $p$-values, i.e. $p$-values corrected for multiple testing (see again Fig. 1). Gene sets below the chosen significance niveau are classified as *significantly and consistently* enriched.

*Extensions* To apply to regulation processes involving multiple regulators and transcription complexes composed of several genes, we allow a transition $t$ to have an arbitray number of inputs $I_t = \{i_t^1, \ldots, i_t^k\}$ and outputs $O_t = \{o_t^1, \ldots, o_t^l\}$. This is accomplished via generalization of equation (10) to

$$\left(\mathrm{de}(o_t^1), \ldots, \mathrm{de}(o_t^l)\right) \approx f_t\left[\left(\mathrm{de}(i_t^1), \ldots, \mathrm{de}(i_t^k)\right)\right]. \qquad (14)$$

We model the combined effect via computation of the average behavior of all effects, or optionally, by the effect of highest statistical significance (the effect could, of course, also be modeled as a full-blown $k$-dimensional (fuzzy) function).
Missing data, i.e. genes of the GRN, which are not measured in the study, is resolved using transitivity. By going up and down, respectively, the regulation path until a non-empty place is reached, an empty origin is filled with the found token, which is adjusted to path length of the transitive relation. The adjustment is due to the fact that the evidence for regulation weakens, as the path length increases.

*Implementation and Availability* GGEA is implemented in the statistical language `R` (Ihaka and Gentleman, 1996) and makes use of the `Bioconductor` software suite (Gentleman *et al*., 2004). The GGEA method and accompanying visualization capabilities have been bundled into an `R` package and tied to a grahical user interface, the `Galaxy` workflow environment (Goecks *et al*., 2010), that is running as a web server.

## Consistency and Explainability Study Setup

*Data Sampling and Network Construction* Gene expression data of *E. coli* was collected and sampled from the `M3D` database (Many Microbe Microarrays Database, Faith *et al*., 2008). 1000 datasets were designed in a two-class fashion, s.t. each class contained 15 samples. It was assured that real-world distributions of fold changes and differential expression $p$-values were matched. A global gene regulatory network for *E. coli* was constructed using the regulatory interactions provided in the `RegulonDB` database (Gama-Castro *et al*., 2008). From the union of all stored TF/gene, TF/operon, TF/TF, $\sigma$/gene and $\sigma$/TU regulatory interactions (TF stands for *transcription factor*, TU for *transcriptional unit* and $\sigma$ for the RNA polymerase $\sigma$-factor), we removed duplicated and ambiguous edges. The final network connected 2097 unique nodes by 5784 edges, which were clearly annotated as either activating or inhibiting.

*Methods Collection and Gene Set Definitions* For each dataset, we applied the standard hypergeometrical overrepresentation test ORA1, and a collection of array resampling methods that correctly control false positive rates and gene correlation patterns (Gatti *et al*., 2010). These are the modified resampling overrepresentation test ORA2 (Goeman and Bühlmann, 2007), SAFE (Barry *et al*., 2005), GSEA (Subramanian *et al*., 2005) and SAM-GS (Dinu *et al*., 2009). The gene set catalog for analysis was defined on the one hand according to the `KEGG` pathway annotation (Ogata *et al*., 1999) for *E. coli*, and, on the other hand, according to the `GO` classifications (Ashburner *et al*., 2000) of *E. coli*. We restricted both catalogs to gene sets having at minimum five and at maximum 500 set members. This yielded 83 and 446 gene sets for `KEGG` and `GO`, respectively.

*Consistency Benchmark* For each method, we collected for all datasets with statistical significant outcome ($p < 0.05$) the top ranked gene sets. As not all datasets produced significant outcome for all methods, we uniformly chose 700 sets at random from these top ranked gene sets and computed the percentage of consistent relations in the corresponding induced regulatory networks. We took regulation direction, type and strength into account and distinguished respective categories. Activating relations required both interaction partners to be expressed in the same direction to be consistent, while inhibiting relations required them to be expressed in the opposite direction. Regulation strength was categorized as *weak* and *strong*, depending on the differential expression $p$-value of the regulator. We chose 0.5 and 0.05 as the thresholds for the weak and the strong category, respectively. To estimate the null distribution in each category, we computed the consistency of all gene sets in all datasets.

*Explainability Benchmark* The selected 700 top ranked sets were restricted to differentially expressed genes of high statistical significance. The significance niveau was set to 0.1. Minimum spanning trees (MST) were computed for each of the reduced gene sets according to the underlying global GRN, s.t. each significant gene of a top ranked set could be reached by all other significant members of that set. Moreover, the corresponding MST for such a set minimized the number of genes not contained in the set. The direction of the regulatory link between two genes in the network (activation/inhibition) as well as the direction of the expression change of individual genes (down-/up-regulation) was ignored. We classified a restricted result set as *fully explainable* if all members were directly connected to another member in the corresponding MST. Otherwise, we counted the number $x$ of genes in the MST, which were not a member of the set, and classified the set as *explainable with* $x$ additional genes. As a measure of explainability achieved by a method in all its 700 top ranked sets, we calculated, for a chosen number $x$, the percentage of sets that were explainable with at most $x$ additional genes.

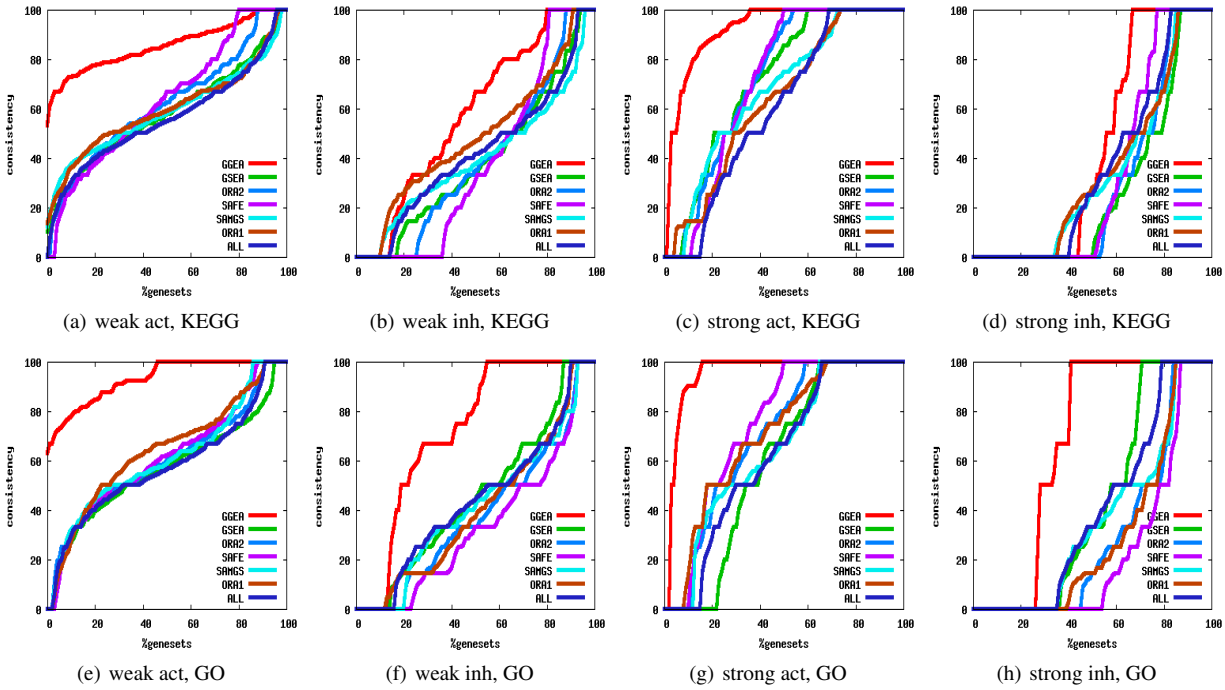
## Case Study Setup

*FiDePa and Local GGEA* We applied GGEA to the glioma dataset that has been investigated before with the method FiDePa (Keller *et al*., 2009). The method exploits GSEA first to determine differentially regulated paths of a particular length and uses the resulting paths for the construction of a consensus network, which is subsequently tested for overrepresentation of gene sets. In a similiar approach, we computed consistency scores of regulatory links in all human non-metabolic `KEGG` pathways (gene regulatory and signaling pathways) and the ten edges with the highest consistency score were extracted from each of them. Duplicated edges were removed and the consensus graph was further reduced via application of a high pass consistency filter using the mean consistency score as threshold. That yielded a total of 378 edges connecting 342 unique nodes, which were tested, as in FiDePa, for overrepresentation.

# 3 RESULTS

## Consistency Study

We conducted a meta-analysis of 1000 *E. coli* datasets and evaluated the consistency within results of gene set enrichment methods, based on the regulatory interactions found in the transcriptional network of *E. coli*. Details of the study setup, the consistency benchmark and the classification of interaction strength as *weak* and *strong* are described in METHODS. The results are shown in Fig. 4. We observe that the set enrichment methods systematically neglect mutual regulation among set members. For `KEGG` gene sets, weak regulations (Fig. 4a and Fig. 4b) are only slighly more consistent than average (the null consistency) and the gene set with maximal consistency is frequently not reported by the set enrichment methods, regardless of activatory or inhibitory links. Strong activators, with an expression change of high statistical significance, and the effects on their targets are more consistently aligned (Fig. 4c). However, the consistency gained in strong activations is lost for strong inhibitions (Fig. 4d). The results for `KEGG` sets are nearly replicated in `GO` gene sets (Fig. 4e-h). In contrast, GGEA, which takes consistency into account for selecting relevant gene sets in the first place, yields the most consistent gene sets in all categories for both, `KEGG` and `GO` gene set definitions.




(a) weak act, KEGG (b) weak inh, KEGG (c) strong act, KEGG (d) strong inh, KEGG

(e) weak act, GO (f) weak inh, GO (g) strong act, GO (h) strong inh, GO

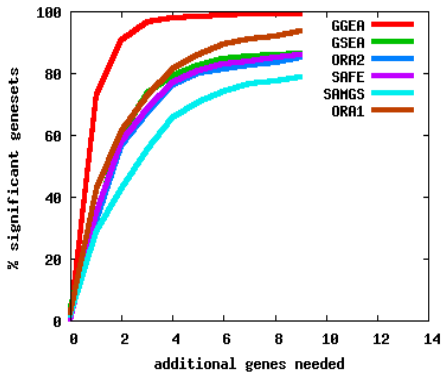
**Fig. 4. Consistency of Regulatory Interactions in Top Ranked Sets.** Each of the set enrichment methods was applied to 1000 *E. coli* datasets using KEGG and GO gene set definitions, respectively. From datasets with statistical significant outcome, the top ranked gene sets were collected and investigated for consistency of *weak* and *strong* activation and inhibition (as described in METHODS). GGEA results are displayed in red. The plots show which fraction ($x$-axis) of the identified gene sets had at most a consistency of $y\%$. The $y$-axis shows the consistency of sets as the fraction of consistent regulatory interactions in the respective gene set. The null consistencies were estimated via the overall consistency of all gene sets in all datasets and are displayed in dark blue.

Activations and inhibitions are similarily consistent, if adjusted to background distributions of both regulation types, and stronger signals are properly weighted in order to preserve the regulation kinetics. Although stronger signals have an higher impact on the GGEA score, weak regulations are also highly consistent in the sets found by GGEA. In general, these findings are more pronounced for `GO` sets, compared to `KEGG` gene set definitions. This is due to the fact that the `GO` catalog (446 gene sets) is nearly six times larger and contains more diverse composed gene sets than the `KEGG` catalog (83 gene sets), which emphasizes differences between the set and graph enrichment methods.

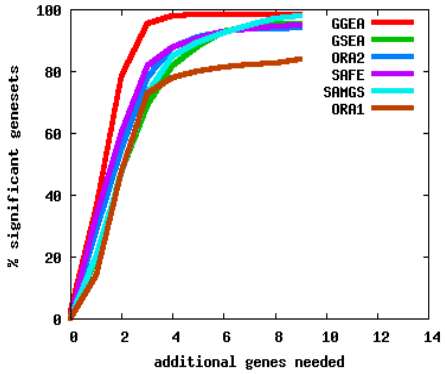
## Explainability Study

As the consistency is substantially incorporated in the GGEA score, we performed a second evaluation using the more independent benchmark of explainability, as described in METHODS. The main target of this investigation was to determine to which extent statistical significant expression changes of single genes can be explained by other set members. Considering that a statistical significant finding for a gene set indicates differential regulation of the corresponding biological process, it is in turn implied that a part of the global regulatory network (here a subgraph of `RegulonDB`) exists, which connects the differentially expressed genes in this set. However, it is frequently observed that important regulators or mediators are missing in a particular gene set, leaving its differentially expressed genes not connected with each other. As a result, the biological interpretation of the observed effect is impeded. Based on these considerations, we have introduced above the terms *fully explainable* and *explainable with $x$* additional genes, to assess how easily a result set can be interpreted. Intuitively, the less additional genes needed, the easier the interpretation: a single additional gene could possibly be a regulator or mediator not contained in the set, while the need of several additional genes requires more complex assumptions to make the outcome interpretable. For the explainabilty study, we explicitly made the input regulatory network undirected, generalizing the edges, s.t. possibly unknown inverse regulations are allowed. We enhanced this feature by additionally removing the sign of the fold change and only judged whether a gene was differentially expressed or not. The results are shown in Fig. 5.

GGEA systematically reports more easily explainable sets than all other methods for both, `KEGG` and `GO` gene set definitons. Similar to the results of the consistency study, the gap is much bigger between the performance of GGEA and the other methods when using `GO` sets definitions, as also observed in the consistency study. For example, GGEA needs in 73% of its top ranked gene sets a single additional gene to make the differentially expressed genes in a particular set connected, whereas the best set enrichment method, ORA1, can explain only 42% of the sets with a single addition gene (SAFE: 35%, GSEA: 34%, ORA2: 32%, SAMGS: 29%). Allowing two additional genes, GGEA can explain more than 90% of all reported genesets, while all other methods produce results around 60% or below.




(a) GO




(b) KEGG

**Fig. 5. Explainability of Expression Changes in Top Ranked Sets.** The 700 top ranked gene sets (introduced in the consistency study above) of each method were restricted to genes with expression changes of statistical significance. For each restricted set, we computed the minimal number of genes not in the set, but needed to connect the significantly regulated genes of that set, to a regulation network. Displayed is the percentage of gene sets, for which $x$ or less additional genes are needed. E.g. for GO sets, a single additional gene makes 73% of GGEA's top ranked sets explainable, while in case of ORA1 or SAMGS a single gene makes only 42% or 29% of the top ranked sets, corresponding to each method, explainable.

## Case Study

In a final case study, we investigated two expression datasets of human neuronal tumours and compared results of GGEA and set enrichment strategies. Though a comparative benchmark is hard to find, due to a missing gold standard that classifies detected pathways as right or wrong in the context of the investigated expression data, we approached this matter via collection of biological evidence in the scientific literature and focussed on the specificity of the findings and the sensitivity of the method used. For consistency evaluation, we used the regulatory interactions occuring in human non-metabolic KEGG pathways (gene regulatory and signaling pathways). In the first analysis, we applied GGEA to the glioma dataset that was investigated before by Keller *et al.* (2009) with the method FiDePa (see METHODS for details). We observe large agreement in the result lists of both methods (Table 2); 17 pathways listed in the FiDePa result also occur in the top 25 of the GGEA ranking. The positive control Glioma is better ranked

**Table 2. Result comparison of GGEA and FiDePa application to the glioma dataset.** Arrows in the first column denote whether a pathway is ranked higher or lower by GGEA, compared to FiDePa.

| Pathway | ORA $p$ (GGEA) | ORA $p$ (FiDePa) | Rank (FiDePa) |
|---|---|---|---|
| ↑ Pathways in cancer | 1.8e-24 | – | – |
| ↑ Focal adhesion | 1.4e-18 | 2.5e-06 | 5 |
| ↑ T cell receptor signaling | 1.2e-17 | 1.5e-05 | 7 |
| ↑ Neurotrophin signaling | 5.5e-15 | – | – |
| ↑ Colorectal cancer | 1.1e-14 | 9.4e-05 | 11 |
| ↑ Pancreatic cancer | 3.8e-14 | 0.0001 | 12 |
| ↑ Renal cell carcinoma | 1.3e-13 | – | – |
| ↑ VEGF signaling | 1.5e-13 | 0.006 | 22 |
| ↔ Fc epsilon RI signaling | 4.1e-13 | 1.9e-05 | 9 |
| ↓ Chronic myeloid leukemia | 6.3e-13 | 1.65e-05 | 8 |
| ↑ ErbB signaling | 8.9e-13 | – | – |
| ↑ B cell receptor signaling | 4.2e-12 | 0.001 | 17 |
| ↑ Glioma | 5.1e-12 | 0.003 | 20 |
| ↑ Insulin signaling | 3.2e-11 | 0.001 | 18 |
| ↑ Leukocyte trans. migration | 3.9e-11 | 0.01 | 24 |
| ↓ Adherens junction | 4.9e-11 | 1.4e-05 | 6 |
| ↓ GnRH signaling | 6.5e-11 | 0.0003 | 16 |
| ↓ Nat. killer cell med. cytotox. | 6.5e-11 | 1.4e-11 | 2 |
| ↑ Wnt signaling | 1.2e-10 | – | – |
| ↓ Toll-like receptor signal. | 1.2e-09 | 5.5e-05 | 10 |
| ↑ Endometrial Cancer | 1.6e-07 | – | – |
| ↑ Non-small cell lung cancer | 3.4e-07 | – | – |
| ↑ Acute myeloid leukemia | 3.9e-07 | – | – |
| ↓ mTOR signaling | 1.2e-06 | 0.0002 | 15 |
| ↓ MAPK signaling | 4.4e-06 | 1.6e-25 | 1 |
| ... | ... | ... | ... |
| ↓ Apoptosis | 0.04 | 9.3e-11 | 3 |

(and has higher significance) by GGEA. Further, several unspecific and disease unrelated pathways detected by FiDePa (e.g. Type I/II diabetes mellitus, Cell cycle) are discarded by GGEA and replaced by specific, cancer related pathways (e.g. Renal cell carcinoma, Endometrial cancer). For the top rank, GGEA (Pathways in Cancer; not detected by FiDePa) gives a clear disease related hint, while FiDePa (MAPK signaling pathway) reports a general signaling process. The Neurotrophin signaling pathway, which promotes neuronal tumors via modulation of neuronal apoptosis (Miller and Kaplan, 2001), is not identified by FiDePa, but listed by GGEA on rank 4.

In the second evaluation study, we used neuroblastoma expression data that was investigated for enrichment of metabolic pathways before (Schramm *et al.*, 2010). The application of GGEA to the neuroblastoma dataset identified 17 significantly and consistently enriched pathways (Table 3). Best ranked is the Neurotrophin signaling pathway, which was already detected in the glioma study to play an essential role in the development of neuronal tumors. As this pathway seemed to be particularly striking for both tumors, we determined regulations with highest consistency in that pathway, in order to get a deeper insight into the disease causing dynamics: We found that the high affinity nerve growth factor receptor, which in humans is encoded by the NTRK1 gene, is up-regulated in

**Table 3.** Result of GGEA application to the neuroblastoma dataset.

| Pathway | $p$-value |
|---|---|
| Neurotrophin signaling | 7.5e-06 |
| Chemokine signaling | 0.0004 |
| Cell adhesion molecules (CAMs) | 0.0021 |
| Regulation of actin cytoskeleton | 0.0068 |
| Focal adhesion | 0.0091 |
| Nat. killer cell med. cytotox. | 0.0092 |
| Leukocyte trans. migration | 0.0099 |
| Pathways in cancer | 0.01 |
| T cell receptor signaling | 0.016 |
| Fc epsilon RI signaling | 0.019 |
| Long-term depression | 0.023 |
| Axon guidance | 0.033 |
| Vasc. smooth muscle contraction | 0.035 |
| p53 signaling pathway | 0.035 |
| Melanogenesis | 0.039 |
| MAPK signaling | 0.043 |
| Thyroid cancer | 0.05 |

neuroblastoma cells and activates the adaptor protein SH2B3, the growth factor receptor-bound protein 2 (GRB2), the Abelson murine leukemia viral oncogene homolog 1 (ABL1), the phospholipase gamma 2 (PLCG2) and the SHC-transforming protein 1 (SHC1). A literature search revealed that all of the activated and associated proteins are proliferating, oncogenic and/or apoptosis influencing and thus, of cancer promoting importance (e.g. Ohmichi *et al*., 1991; Borrello *et al*., 1994). In addition, the up-regulation of the whole NTRK1 proliferation module in neuroblastoma was experimentally validated (Evangelopoulos *et al*., 2004) some years ago. This sensitive finding motivated a similar investigation for the other pathways in Table 3, which we identified to be throughout substantially involved in neuroblastoma formation. As an example: GGEA detects the Chemokine signaling pathway. We found that neuroblastoma impairs chemokine-mediated dendritic cell migration (Walker *et al*., 2006) and chemokines strongly promote neuroblastoma primary tumor and metastatic growth (Meier *et al*., 2007).

Moreover, we wanted to know whether the findings of GGEA are in concordance with the results for metabolic pathways. As a showcase, we demonstrate this via the detected Fc epsilon RI signaling pathway. In Schramm *et al*. (2010), only moderate attention (discussed in their supplement) is paid to the extremely significant findings for Phosphatidylinositol metabolism ($p$ = 9e-12) and for several pathways concerning the metabolism of lipids and fatty acids, e.g. Fatty acid metabolism ($p$ = 1.7e-9) and Glycerophospholipid metabolism ($p$ = 3.9e-7), which are listed in Table 1 of that publication. As it can be verified in the corresponding `KEGG` pathway maps, Fc epsilon RI signaling has a regulatory impact on both - the Phosphatidylinositol metabolism via modulation of the phospholipase (affected by the Neurotrophin pathway); and the metabolism of lipids in general via stimulation of arachidonic acid synthesis. Arachidonic acid is a polyunsaturated fatty acid that is required for membrane phospholipid synthesis. It is also involved in cellular signaling and known to activate syntaxin-3, which causes cell membrane expansion of neuronal cells (Darios and Davletov, 2006). Schramm *et al*. explain the several revealed signals in lipid related metabolisms with TCA based energy production; the GGEA results, explaining stimulation of arachidonic acid synthesis, imply that the observed activated production of fatty acids and lipids (which is based on the latter) is rather due to the increased requirement of neuronal membrane material (i.e. specific lipids) in the fast growing and dividing neuroblastoma cells.

## 4 DISCUSSION

In this work, we presented *Gene Graph Enrichment Analysis* (GGEA), a novel algorithmic framework to detect increased agreement between positively and negatively correlated expression patterns of genes, connected by activating and inhibiting edges in signed and directed transcriptional networks. The method exploits directed regulatory relations represented as fuzzy logic rules to assess and identify graphs, which maximize the consistency between the regulatory network and the expression data. GGEA is a major improvement to current gene set enrichment strategies, as we found experimentally validated regulatory interactions not to be consistent *per se* with the expression data in top ranked and statistically significant result sets of these methods. That was validated in a large-scale consistency study of 1000 *E. coli* chips using the *E. coli* `RegulonDB`, currently the best curated regulatory network, for the investigation of consistency. As set enrichment strategies ignore mutual regulation among set members, we observed that activations and inhibitions are only average consistent with the gene expression in these result sets. Even strong causal signals, i.e. a regulator with differential expression of high statistical significance, in pairwise directed regulations were frequently not properly reflected. Inhibitions were more seriously neglected than activations. This is partly due to a data bias, as there are more activations than inhibitions in the database. Hence, more genes, and thus also more significant genes, are involved in activations just by chance. As gene set enrichment analysis mainly computes upon the leading edge of the ranked $p$-value vector of genewise differential expression (see Subramanian *et al*., 2005), gene sets with a majority of activating genes are more likely to be reported. On the other hand, we found activitations clearly better conserved than inhibitions across all experiments stored in the `M3D` database. For GGEA, we observed, under consideration of this bias, that activations were nearly optimally consistent and inhibitions were preserved in a large fraction of regulations. GGEA achieved the highest concordance between the regulation direction and the expression behavior of the incorporated regulator and regulated target gene. It should be emphasized that GGEA consistently aligned weak (only moderately differentially expressed) signals, which are usually not taken into account by set enrichment methods. That improved sensitivity enables preference of weak, but coherent regulations over strong, but contextless signals. This is expected to better reflect the nature of key cellular regulators.

As GGEA exploits the consistency for the computation of its score, we additionally carried out a more independent benchmark to investigate how well statistically significant expression changes of single genes can be explained by other set members. As a measure of explainability, we used the number of additional genes,

which were needed to connect significant members of a set to a regulatory network. We found this evaluation of particular interest, as it tries to approximate the process of the human interpretation. For all set enrichment methods, only a small amount of genes could be explained by other set members in a significant result and we observed frequently that several additional genes were needed. Implied is that set enrichment indeed indicates that there is *something* striking happening in a certain result set, however, conclusions whether observed expression changes are coherent and in context with the surrounding regulators cannot be drawn. This is resolved by GGEA. It systematically reports more easily explainable sets than all other methods, and the fraction of explainable sets with a single additional gene is increased by over 30% in comparison with the best set enrichment method.

Furthermore, we applied GGEA in two pilot case studies of human neuronal tumors using regulatory interactions of signaling pathways, though incorporated protein-protein regulations cannot be measured at the transcriptional level. Nonetheless, we again hypothesize that genes, annotated to be associated in a pathway, should show higher correlation patterns than arbitrary genes, which are not. On the other hand, we argue that signal cascades normally target altered gene regulation.

On the glioma dataset, GGEA discovered throughout specific and disease related pathways. Induced by increasing specificity, the fraction of false positives decreases. Unspecific and inconsistent pathways are replaced by more appropriate pathways. An example is the detection of the Neurotrophin signaling pathway that modulates neuronal apoptosis (a very specific finding), while general apoptosis is downgraded.

The Neurotrophin signaling pathway also has a major influence on the development of neuroblastoma, another neuronal tumor type. The experimentally verified connection was detected by GGEA with high significance, while GSEA failed to detect it. The discovery of such false negatives of the set enrichment analysis is due to improved sensitivity already observed in the consistency study. However, it is surprising that only GGEA is sensitive enough to detect the Neurotrophin signaling pathway, the Chemokine signaling pathway and the Fc epsilon RI signaling pathway - all of which have been shown to be of crucial importance in neuroblastoma formation - while standard GSEA does not detect them. Best ranked pathways of GSEA are: Cell cycle, Ribosom and Olfactory transduction. The connection to the disease is incomprehensible and explanations are almost arbitrary.

## 5 CONCLUSION

We showed in three independent and differently designed studies that GGEA consistently aligns regulation and expression and yields result sets where statistically significant expression changes can be explained by regulators within the set. Moreover, GGEA eases the biological interpretation of reported gene sets, as they are more coherent than sets reported by set enrichment methods. This means many more of their relevant genes are connected or can be connected by a minimum number of additional factors. In summary, our new method *Gene Graph Enrichment Analysis* (GGEA) is an intuitive enrichment method, which uses gene regulatory information to improve consistency and coherence of detected enriched gene sets and, thus, substantially reduces the fraction of false positive and false negative classifications of relevant gene sets. GGEA significantly improves the detection of gene sets where measured positively or negatively correlated expression patterns coincide with directed inducing or repressing relationships between the respective pairs of genes. Hence, gene set regulators, such as transcription factors, can explain a significant portion of the observed expression changes. As GGEA is as fast and easy to apply to experimental data as state-of-the-art set enrichment analysis methods, it provides an alternative for interpreting gene expression measurements and for deriving first insights into the relevant processes. The advantages of GGEA will increase in the future with the availabilty of better GRNs and better models for regulatory relations in these GRNs.

## ACKNOWLEDGEMENTS

LG has been partly funded by the DFG international research training group 1563/1 RECESS.

## REFERENCES

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25-9.

Barry WT, Nobel AB, Wright FA (2005) Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics*, **21**, 1943-9.

Breitling R, Amtmann A, Herzyk P (2004) Iterative Group Analysis (iGA): a simple method to enhance sensitivity and facilitate interpretation of microarray experiments. *BMC Bioinformatics*, **5**, 34.

Borrello MG, Pelicci G, Arighi E, De Filippis L, Greco A, Bongarzone I, Rizzetti M, Pelicci P G, Pierotti M A (1994) The oncogenic versions of the Ret and Trk tyrosine kinases bind Shc and Grb2 adaptor proteins. *Oncogene*, **9**, 1661-8.

Chaouiya C (2007) Petri Net Modelling of Biological Networks. *Brief. Bioinform*, **8**, 210-9.

Darios F and Davletov B (2006) Omega-3 and omega-6 fatty acids stimulate cell membrane expansion by acting on syntaxin 3. *Nature*, **440**, 813-7.

Dinu I, Potter JD, Mueller T, Liu Q, Adewale AJ, Jhangri GS, Einecke G, Famulski KS, Halloran P, Yasui Y (2009) Gene-set analysis and reduction. *Brief. Bioinform.*, **10**, 24-34.

Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA*, **95**, 14863-8.

Evangelopoulos ME, Weis J, Kruttgen A (2004) Neurotrophin effects on neuroblastoma cells: correlation with trk and p75NTR expression and influence of Trk receptor bodies. *J. Neurooncol.*, **66**, 101-10.

Faith JJ, Driscoll ME, Fusaro VA, Cosgrove EJ, Hayete B, Juhn FS, Schneider SJ, Gardner TS (2008) Many Microbe Microarrays Database: uniformly normalized Affymetrix compendia with structured experimental metadata. *Nucleic Acids Res.*, **36**, D866-70.

Gama-Castro S, Jimenez-Jacinto V, Peralta-Gil M, Santos-Zavaleta A, Penaloza-Spinola MI, Contreras-Moreira B, Segura-Salazar J, Muniz-Rascado L, Martinez-Flores I, Salgado H, Bonavides-Martinez C, Abreu-Goodger C, Rodriguez-Penagos C, Miranda-Rios J, Morett E, Merino E, Huerta AM, Trevino-Quintanilla L, Collado-Vides J (2008) RegulonDB (version 6.0): gene regulation model of Escherichia coli K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. *Nucleic Acids Res.*, **36**, D120-4.

Gatti DM, Barry WT, Nobel AB, Rusyn I, Wright FA (2010) Heading Down the Wrong Pathway: on the Influence of Correlation within Gene Sets. *BMC Genomics*, **11**, 574.

Ge H, Liu Z, Church GM, Vidal M (2001) Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae. *Nat Genet.*, **29**, 482-6.

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JY, Zhang J (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, **5**, 80.

Goecks J, Nekrutenko A, Taylor J, Galaxy Team (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.*, **11**, 86.

Goeman JJ and Bühlmann P (2007) Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, **23**, 980-987.

Ihaka R and Gentleman R (1996) R: A language for data analysis and graphics. *J. Comp. Graph. Stat.*, **5**, 299-314.

Jansen R, Greenbaum D, Gerstein M (2002) Relating whole-genome expression data with protein-protein interactions. *Genome Res.*, **12**, 37-46.

Keller A, Backes C, Gerasch A, Kaufmann M, Kohlbacher O, Meese E, Lenhof HP (2009) A novel algorithm for detecting differentially regulated paths based on gene set enrichment analysis. *Bioinformatics*, **25**, 2787-94

Khatri P and Draghici S (2005) Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*, **21**, 3587-3595.

Küffner R, Zimmer R, Lengauer T (2000) Pathway analysis in metabolic databases via differential metabolic display (DMD). *Bioinformatics*, **16**, 825-36.

Küffner R, Petri T, Windhager L, Zimmer R (2010) Petri Nets with Fuzzy Logic (PNFL): Reverse Engineering and Parametrization. *PLoS One*, **5**, 12807.

Lee HK, Hsu AK, Sajdak J, Qin J, Pavlidis P (2004) Coexpression analysis of human genes across many microarray data sets. *Genome Res.*, **14**, 1085-94.

Liu M, Liberzon A, Kong SW, Lai WR, Park PJ, Kohane IS, Kasif S (2007) Network-Based Analysis of Affected Biological Processes in Type 2 Diabetes Models. *PLoS Genetics*, **3**, e96.

Meier R, Mühlethaler-Mottet A, Flahaut M, Coulon A, Fusco C, Louache F, Auderset K, Bourloud KB, Daudigeos E, Ruegg C, Vassal G, Gross N, Joseph JM (2007) The chemokine receptor CXCR4 strongly promotes neuroblastoma primary tumour and metastatic growth, but not invasion. *PLoS One*, **2**, e1016.

Miller FD and Kaplan DR (2001) Neurotrophin signalling pathways regulating neuronal apoptosis. *Cell. Mol. Life Sci.*, **58**, 1045-53.

Murata T (1989) Petri Nets: Properties, Analysis and Applications. *Proc. of the IEEE*, **77**, 541-580.

Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.*, **27**, 29-34.

Ohmichi M, Decker S J, Pang L, Saltiel A R (1991) Nerve growth factor binds to the 140 kd trk proto-oncogene product and stimulates its association with the src homology domain of phospholipase C gamma 1. *Biochem. Biophys. Res. Commun.*, **179**, 217-23.

Pan W (2002) A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics*, **18**, 546-54.

Schramm G, Wiesberg S, Diessl N, Kranz AL, Sagulenko V, Oswald M, Reinelt G, Westermann F, Eils R, Koenig R (2010) PathWave: discovering patterns of differentially regulated enzymes in metabolic pathways. *Bioinformatics*, **26**, 1225-1231.

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA*, **102**, 15545-15550.

Ulitsky I and Shamir R (2007) Identification of functional modules using network topology and high-throughput data. *BMC Syst. Biol.*, **1**, 8.

Walker SR, Ogagan PD, DeAlmeida D, Aboka AM, Barksdale EM Jr (2006) Neuroblastoma impairs chemokine-mediated dendritic cell migration in vitro. *J. Pediatr. Surg.*, **41**, 260-5.

Windhager L, Zimmer R (2008) Intuitive Modeling of Dynamic Systems with Petri Nets and Fuzzy Logic. German Conference on Bioinformatics, September 9-12, 2008, Dresden, Germany, Lecture Notes in Informatics, vol P-136, pp. 106-115, Gesellschaft für Informatik, 2008.

Windhager L, Erhard F, Zimmer R (2010) Fuzzy modeling. Koch I, Reisig W, Schreiber F (eds.) Modeling in Systems Biology: The Petri Net Approach, *Springer*, Chapter 9.

Zadeh LA (1963) Fuzzy sets. *Information and Control*, **8**, 338-353.