Method

Choosing the right path: enhancement of biologically relevant sets of genes or proteins using pathway structure

Reuben Thomas\textsuperscript{\pi}*, Julia M Gohlke\textsuperscript{\pi}*, Geffrey F Stopper\textsuperscript{†}, Frederick M Parham\textsuperscript{*} and Christopher J Portier\textsuperscript{*}

Addresses: *Environmental Systems Biology Group, Laboratory of Molecular Toxicology, National Institute of Environmental Health Sciences, RTP, NC 27709, USA. †Department of Biology, Sacred Heart University, Fairfield, CT 06825, USA.

* These authors contributed equally to this work.

Correspondence: Christopher J Portier. Email: portier@niehs.nih.gov

© 2009 Thomas et al.; licensee BioMed Central Ltd. This is an open access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

A method is proposed that finds enriched pathways relevant to a studied condition using the measured molecular data and also the structural information of the pathway viewed as a network of nodes and edges. Tests are performed using simulated data and genomic data sets and the method is compared to two existing approaches. The analysis provided demonstrates the method proposed is very competitive with the current approaches and also provides biologically relevant results.

Background

Data on the molecular scale obtained under different sampling conditions are becoming increasingly available from platforms like DNA microarrays. Generally, the reason for obtaining molecular data is to use these data to understand the behavior of a system under insult or during perturbations such as occurs following exposure to certain toxicants or when studying the cause and progression of certain diseases. Toxins or diseases will hereafter be commonly referred to as perturbations to the biological system. Genomics is capable of providing information on the gene expression levels for an entire cellular system. When faced with such large amounts of molecular data, there are two options available that can enable one to focus on a small number of interesting sets of genes or proteins. One can cluster the data [1] and use the clusters to identify sets of genes that were significantly affected by the perturbations. This represents an unsupervised approach.

Other similar approaches include principal component analysis [2] and self-organizing maps [3].

Alternatively, biologically relevant sets of genes/proteins are deduced to exist \textit{a priori} in the form of biochemical pathways and cytogenetic sets. A supervised approach can be linked with the data to identify these \textit{a priori}-defined sets that are significantly affected by the perturbations seen in the data. The method proposed in this paper is an example of this approach applied to the scenario of distinguishing between two conditions (such as normal patient versus disease patient, or unexposed versus exposed). The data we wish to link to a given set of pathways are assumed to be genomic data such as gene expression levels or the presence of gene polymorphisms known to be associated with diseases.

Supervised approaches for the identification of biologically relevant gene expression sets have typically been identified as
'gene set' or 'pathway enrichment' methods in the literature. Recent years have seen significant work done on proposals for new approaches guided by criticisms and limitations of the existing ones; references [4-8] provide a critical review of the existing methods in terms of their different features, such as the null hypotheses of the underlying statistical tests used and the independence assumption between genes. These reviews essentially inform us that the pathway enrichment methods can be viewed as falling on two sides of a number of different coins. A few of these classifications are given below.

Firstly, methods could be interested in testing either whether the genes in a specific pathway of interest are affected as a result of a treatment (the implied null hypothesis has been referred to as 'self-contained' [4] or denoted as 'Q2' [9]) or whether the genes in the pathway of interest are more affected than the other genes in the system (this implied null hypothesis has been referred to as 'competitive' [4] or as 'class 1, 2, 3' [6] or denoted as 'Q1' [9]). There are of course good reasons for preferring either of these null hypotheses. One would prefer the 'competitive' hypothesis if the treatment had a wide ranging impact on the genes in the system. This could have an undesirable consequence of having randomly chosen (and hence not biologically relevant) sets of genes attaining significance for the 'self-contained' tests; a nice illustration of a case like this is provided in [10]. One could use a 'self-contained' test if the belief is that the treatment had quite a restricted impact on the genes in the system and/or if their only focus is on one or a small number of pathways.

Some of the pathway enrichment methods treat the genes in the system as being independent of each other [7,9,11-22]. Ignoring the gene-gene correlations has been shown to have the effect of elevated false-positive discoveries [4,6]. However, the need to prioritize the different biological pathways with respect to their relevance to the treatment and the lack of a sufficient number of biological replicates (one in some cases) may force the need for this independence assumption. Examples of methods that try to take into account the gene-gene correlations include [6,9,10,23-37].

Pathway enrichment methods can be distinguished by the use or the absence of an explicit gene-wise statistic to measure the gene's association with the treatment in determining a pathway's relevance to the treatment. Examples of gene-wise statistics used include the two-sample t-statistic, log of fold change [35], the significance analysis of microarrays (SAM) statistic [25] and the maxmean statistic [10]. Methods like those in [24,30,31,34,37,38] treat the problem as a multivariate statistical one and avoid the need for an explicit definition of a gene-wise statistic.

The method proposed in this paper defines versions for both the 'self-contained' and the 'competitive' null hypotheses and utilizes the idea of the maxmean statistic [10]. It improves upon the previous methods by its use of structural information present in biochemical pathways. A pathway is said to have structural information if its components can be placed on a network of nodes and edges. For example, a gene set corresponding to a pathway can be viewed to be associated with a network where the nodes represent the gene products (that is, proteins, protein complexes, mRNAs) while the edges represent either signal transfer between the gene products in signaling pathways or the activity of a catalyst between two metabolites in metabolic pathways.

Classic signal transduction pathways, such as the mitogen-activated protein kinase (MAPK) pathways, transduce a large variety of external signals, leading to a wide range of cellular responses, including growth, differentiation, inflammation and apoptosis. In part, the specificity of these pathways is thought to be regulated at the ligand/receptor level (for example, different cells express different receptors and/or ligands). Furthermore, the ultimate response is dictated by the downstream activation of transcription factors. Alternatively, intermediate kinase components are shared by numerous pathways and, in general, do not convey specificity nor do they directly dictate the ultimate response (see [39] for a review). Therefore, we test the value of implementing a Heavy Ends Rule (HER) in which the initial and final components of a signaling pathway are given a higher weight than intermediate components.

Signal transduction relies on the sequential activation of components in order to implement an ultimate response. Therefore, we hypothesize that activation of components that are directly connected to each other in a pathway conveys greater significance than activation of components that are not closely connected to each other. Therefore, we also test the implementation of a Distance Rule (DR) scoring rule in which genes that are closely connected to each other are given a higher score.

The use of structural information based on an underlying network in an analysis of gene expression data is not new. Similar ideas have been used to identify activated pathways from time profile data (here the attempt was to distinguish between two phenotypes) [40], while structural information of the pathways has been used to enhance the clusters deduced from the gene expression data [41] and to find differentially expressed genes [42]. The study by Draghici et al. [43] appears to be the only existing work that incorporates pathway network information to the problem of pathway enrichment. However, this appears to be limited by the need to define an arbitrary cut-off for differential expression, the assumption of independence between genes and the parametric assumption of an exponential distribution for computing the significance.
Results and discussion

The method proposed in this paper is named 'structurally enhanced pathway enrichment analysis' (SEPEA). It is a pathway enrichment method that incorporates the associated network information of the biochemical pathway using two rules, the HER and DR. SEPEA provides three options for null hypothesis testing (SEPEA\_NT1, SEPEA\_NT2 and SEPEA\_NT3) that depend on the goal of the pathway enrichment analysis and the properties of genomic data available. SEPEA\_NT1 and SEPEA\_NT2 require multiple array samples per gene and are tests that take into account inherent gene-gene correlations. SEPEA\_NT3 just requires a summary statistic per gene (that indicates association with the treatment) but assumes that genes are independent of each other. The need for the test SEPEA\_NT3 is motivated by the fact that there are situations where the data are just not sufficient to estimate gene-gene correlations, such as the case where the only information available is whether a gene is or is not affected by the treatment; analyzing the situation of having a set of gene polymorphisms known to be associated with breast cancer is one such example. SEPEA\_NT1 and SEPEA\_NT3 are proposed to be used in situations where the goal is to compare the genes in the pathway of interest to the other genes in the system in terms of their associations with the treatment. SEPEA\_NT2 is used for analyses involving only the genes in the pathway in relation to the treatment. The main objective of this paper is to demonstrate the utility of incorporating pathway network information in a pathway enrichment analysis. Therefore, comparisons are made with results from corresponding versions of SEPEA that do not use the network information - SEPEA\_NT1*, SEPEA\_NT2* and SEPEA\_NT3*. In addition, two literature methods are used for comparison with the results from SEPEA\_NT1 - gene set enrichment analysis (GSEA) [35] and the maxmean method [10] - the null hypotheses of GSEA and maxmean being very similar to SEPEA\_NT1.

Motivation for the Heavy Ends Rule score

By giving greater weight to genes whose products are nearest to the terminal gene products of a pathway, the HER score gives more weight to genes specific to a particular pathway. This is illustrated in Figure 1, which uses the concept of terminal gene products. They are gene products like either receptors that initiate the pathway activity or transcription factors that are made to initiate transcription as a result of the pathway activity (see Materials and methods for a more mathematical definition). The genes involved in each of the signaling pathways in the Kyoto Encyclopedia of Genes and

![Figure 1](image-url)

**Figure 1**
Empirical distribution function of number of pathways associated with genes at given distances from terminal nodes. Empirical cumulative distribution function of the number of pathways that are associated with genes that have gene products located at a given distance, \(d = 0, 1, 2, 3, 4\), from a terminal node of the pathway network. Gene products that are at a distance \(d = 0\) are the terminal gene products. The data used were those of all the genes associated with human signaling pathways in the KEGG pathway database [44].
Genomes (KEGG) pathway database [44] were evaluated for
the position of their gene products with respect to the termi-
nal gene products and the total number of signaling pathways
that these genes are involved in. It is clear from Figure 1 that
genes associated with products that are closer to the terminal
gene products are more pathway-specific.

Justification for the Distance Rule score
To illustrate the utility of the DR as a scoring method, we con-
sider the linkage between the full set of pathways in KEGG
[44]: that is, the pathways themselves can be viewed to be
part of a higher level network, the nodes of which are path-
ways while the edges indicate the transfer of signal or mate-
rial between pathways (Figure S1 in Additional data file 2).
For example, the MAPK signaling pathway and the p53 sig-
aling pathway can be considered to be linked. It seems rea-
sonable to expect that after perturbation of the system, the
affected pathways that are linked are more likely to respond
similarly. We test this intuition using different microarray
data (from the Gene Expression Omnibus (GEO) database
[45] in a statistical test on the above network of pathways. The
details are provided in the Materials and methods section.
The P-values for the eight comparisons (estimated using
1,000 random networks) are given in Table 1. Significant P-
values across the comparisons support our use of the DR as a
reasonable score for differentiating between pathways.

Analysis using simulated data
Simulated data were generated from two pathway networks
having different patterns of correlation between the various
genes in the pathway, with each network having genes in a
pool of genes representing a biological system. The pair of
networks and the correlation patterns of genes in the path-
way, denoted by pattern numbers, are listed in Table 2. Pat-
terns 1, 2, 3 and 4 have non-zero correlation between a subset
of genes in the system. All genes in pattern 5 are assumed to
be independent of each other. Patterns 1 and 3 are biased to
the scoring rules proposed here whereas patterns 2 and 4 are
not. The treatments had the effect of increasing (as given in
the variable, pert) the expressions of certain genes in the sys-
tem.

Table 3 gives estimates of the type 1 errors of the five meth-
ods, at the 0.01 and 0.05 significance levels, for patterns 1 and
5. Table 4 gives estimates of the power of the SEPEA_NT1,
GSEA and SEPEA_NT2 methods at 0.01 and 0.05 signifi-
cance levels, for a pert value of 1.2 and for patterns 1-4. The
empirical sizes of the methods maxmean and SEPEA_NT3 do
not match their nominal sizes. So the results are provided at
empirical sizes of 0.07 and 0.05 (corresponding to a nominal
size of 0.001 for both cases).

Only patterns 1 and 5 were used to analyze the type 1 error
behavior because they represented the two scenarios (pres-
ence or absence of gene-gene correlations) where pathway
enrichment methods have been shown to have different
behaviors [4,10]. Because of the presence of correlations in
the data, SEPEA_NT3 gives an incorrect type 1 error value for
pattern 1 (Table 3). As has been stated previously, in spite of
this incorrect behavior, there are situations (like those in
which the only information available for each gene is a sum-
mary statistic representing the effect of the treatment) where
methods like SEPEA_NT3 need to be used in order to create
relevant hypotheses regarding affected processes due to the
treatment. SEPEA_NT1, SEPEA_NT2 and GSEA do maintain
the right type 1 error behavior in both the presence and
absence of gene-gene correlations. In the presence of gene-
gene correlations, the maxmean method [10] also does not
maintain the appropriate type 1 error behavior. As expected,
the power estimates of all three SEPEA methods for patterns
1 and 3 were significantly higher (P < 0.05, two-sample test of
proportions) than those for patterns 2 and 4, respectively.
The power estimates for patterns 1 and 3 using SEPEA_NT1
were higher than those for GSEA, demonstrating improve-

Table 1

| GEO accession number | Description                                               | P-value |
|----------------------|------------------------------------------------------------|---------|
| [GEO:GDS2744]        | MCF-7 breast cancer cells - dioxin treatment versus control| 0.005   |
| [GEO:GDS2649](1)     | Early HIV infection CD8+T cells versus uninfected          | <0.001  |
| [GEO:GDS2649](2)     | Chronic HIV infection CD8+T cells versus uninfected        | 0.001   |
| [GEO:GDS2649](3)     | Non-progressive HIV infection CD8+T cells versus uninfected| 0.004   |
| [GEO:GDS2852](1)     | Bronchial A549 cells - cytokine treatment at 0 h versus control| 0.001   |
| [GEO:GDS2852](2)     | Bronchial A549 cells - cytokine treatment at 4 h versus control| <0.001  |
| [GEO:GDS2852](3)     | Bronchial A549 cells - cytokine treatment at 12 h versus control| <0.001  |
| [GEO:GDS2852](4)     | Bronchial A549 cells - cytokine treatment at 24 h versus control| 0.016   |

Different control versus treated conditions in three microarray datasets indicated by the GDS accession numbers [GEO:GDS2744],
[GEO:GDS2649] and [GEO:GDS2852] from the GEO database were used [45] to compare the DR scores across all the pathways on the pathway
network (Figure S1 in Additional data file 2) using the meta_DR term in Equation 9. The P-value for the significance of meta_DR is computed using
1,000 random networks whose generation is described in the Materials and methods section.
Different correlation patterns (1-5) considered for the generation of simulated data along with the underlying networks, the set of correlated genes, \( \Sigma \), and the set of genes that are the targets of the treatment, \( \mathcal{V} \). \( \mathcal{U} \) denotes a uniformly randomly drawn set of seven genes drawn from the set of genes associated with the pathway displayed in Figure 1a. \( \mathcal{V}^E \) denotes a set of 41 randomly drawn genes from the set of 470 genes not associated with the pathway displayed in Figure 1a. \( \mathcal{U}^E \) denotes a uniformly randomly drawn set of seven genes drawn from the set of genes associated with the pathway displayed in Figure 1b. \( \mathcal{V}_i^E \) denotes a set of three randomly drawn genes from the set of 413 genes not associated with the pathway displayed in Figure 1b. \( \emptyset \) denotes the empty set. The symbol \( \cup \) denotes the set union operation.

**Type 1 error of different pathway enrichment methods**

| Pattern number | A | SEPEA_NT1 | GSEA | Maxmean | SEPEA_NT2 | SEPEA_NT3 |
|----------------|---|-----------|------|---------|-----------|-----------|
| 1              | 0.01 | 8 | 5 | 85 | 13 | 135 |
| 2              | 0.05 | 36 | 51 | 187 | 44 | 266 |
| 5              | 0.01 | 7 | 12 | 12 | 7 | 15 |
|                | 0.05 | 51 | 45 | 48 | 52 | 53 |

Type 1 errors (in terms of the number of experiments out of 1,000 that gave P-values for the randomization tests below \( \alpha = 0.01 \) and 0.05 levels) for each of the five methods and for correlation patterns 1 and 5.

**Analysis using lung cancer data**

The study by Raponi et al. [46] analyzes gene expression data taken from 130 lung cancer patients in different stages of the disease. They also provide survival times for each patient. The data are divided into two groups of 85 patients (training set) and 45 patients (test set). This was done such that the proportion of patients in each stage was approximately the same for the two groups. Using these data, the Cox proportional hazards statistic is computed for each gene on the microarray (indicating how predictive it is of the survival time of a patient). The next logical step is then an attempt to find what biochemical pathways are predictive of survival. Of all the human KEGG [44] pathways are used in this analysis. The methods used were SEPEA_NT1, GSEA and maxmean. Also, to estimate the value of including information on the network structure, SEPEA_NT3 was applied to the data assuming that all the genes in the pathway are given equal weight and the DR score is zero. This analysis is denoted by SEPEA_NT1*. The goal of our analysis is to evaluate consistency in choosing ‘significant’ pathways found using the training set versus the test set. Curves for sensitivity versus ‘1 - specificity’ and positive predictive value versus negative predictive value are obtained by using different cut-offs for the log of the P-values obtained using each method; the results are shown in Figure 2. The sensitivity, specificity, positive predictive and negative predictive values for SEPEA analyses have better ranges than those for GSEA and maxmean. For a significant portion of the ranges of sensitivity and specificity for GSEA and maxmean, the SEPEA analyses provide higher sensitivity for a given level of false positives (a point on the ‘1 - specificity’ axis). The same can be said about the portion of the ranges of the positive and negative predictive values of maxmean dominated by the SEPEA analyses. From the curves for SEPEA_NT1 and SEPEA_NT1*, we also observe the benefit of incorporating pathway network information. An updated Figure 2 that also includes results from SEPEA_NT2 and SEPEA_NT3 is provided as Figure S2 in Additional data file 3.

**Analysis using exposure of Xenopus laevis to cycloamine data**

Enriched KEGG pathways using SEPEA_NT2 and SEPEA_NT2* (which is essentially the SEPEA_NT2 analysis but does not make use of the network information of the pathways and is identical to the analysis of the Q2 test in [9]) methods were determined for a microarray dataset (see Materials and methods section) examining the consequences of inhibition of Sonic hedgehog (SHH) signaling by cycloamine treatment of developing Xenopus laevis (Tables 5 and 6).
Based on the specificity of cycloamine to inhibit the SHH pathway, we expected to see the SHH signaling pathway significantly enriched; however, the $P$-value for this pathway was not significant using either method (SEPEA_NT2 and SEPEA_NT2*). This may be due to the time point at which gene expression was evaluated, which was optimized to evaluate downstream effectors of SHH pathway inhibition. Alternatively, this result may also reflect the limitation of the method when using only gene expression datasets, as several components of the SHH pathway, including Hedgehog (Hh) and Patched (PTCH), are known to be regulated at the protein level. Finally, when results obtained using SEPEA_NT2 ver-

### Table 4

| Pattern number | A  | SEPEA_NT1 | GSEA  | Maxmean | SEPEA_NT2 | SEPEA_NT3 |
|----------------|----|-----------|-------|---------|-----------|-----------|
| 1              | 0.01 | 328       | 188   | 52      | 357       | 321       |
|                | 0.05 | 610       | 510   |         | 686       |           |
| 2              | 0.01 | 271       | 189   | 37      | 295       | 39        |
|                | 0.05 | 505       | 508   |         | 580       |           |
| 3              | 0.01 | 344       | 222   | 32      | 347       | 480       |
|                | 0.05 | 692       | 496   |         | 712       |           |
| 4              | 0.01 | 166       | 212   | 32      | 157       | 11        |
|                | 0.05 | 361       | 468   |         | 379       |           |

Power estimates for the SEPEA_NT1, GSEA and SEPEA_NT2 methods (in terms of the number of experiments out of 1,000 that gave $P$-values for the randomization tests below nominal sizes of $\alpha = 0.01$ and 0.05). The estimates for maxmean are given at an empirical size of 0.07 (nominal size of 0.001) and those for SEPEA_NT3 at an empirical size of 0.05 (nominal size of 0.001). These are results from simulations in which the treatment resulted in an over-expression of the mean expression of the target genes by the factor $\text{pert} = 1.2$. The methods were evaluated on correlation patterns 1-4.

![Figure 2](http://genomebiology.com/2009/10/4/R44)

**Figure 2**

Receiver-operator characteristic and positive predictive power versus negative predictive power plots for lung cancer data. (a) Sensitivity versus ’1 - specificity’ of enriched pathways that are predictive of survival from lung cancer for four methods: SEPEA_NT1, SEPEA_NT1*, GSEA and maxmean. SEPEA_NT1* is the same analysis as SEPEA_NT1 except that the pathway network information was not used. (b) Positive predictive power (ppp) versus negative predictive power (npp) for the same data and using the same methods of analysis as in (a).
sus SEPEA_NT2* are examined in the context of pathways linked to the SHH pathway (Figure S1 in Additional data file 2), we see that only the MAPK and Proteasome pathways are reachable from the SHH pathway by two and three edges, respectively, suggesting that results from SEPEA_NT2 may be more consistent with targets downstream of the SHH pathway. None of the other pathways listed in Tables 5 and 6 were reachable along the network of pathways (Figure S1 in Additional data file 2) from the SHH pathway. In fact, recent evidence suggests that SHH promotion of proliferation and differentiation in muscle [47] and gastric mucosal cells [48] is through transcription-independent activation of the MAPK/ERK pathway. This analysis suggests benefits of using pathway network information. Additional results from analysis of these data with SEPEA_NT1, SEPEA_NT3, GSEA and maxmean are provided in Additional data file 4.

Analysis using OMIM breast cancer data

Genes associated with breast cancer were downloaded from the Online Inheritance in Man (OMIM) database [49]. This group of genes was pruned to include only those genes that participate in a pathway in the KEGG pathway database [44]. The list of genes used is provided in Table S5 in Additional data file 1. The SEPEA analysis was used to test whether there is an overabundance of important (as defined by the scoring rules) breast cancer genes in pathways relative to the remaining set of genes that participate in some pathway in the KEGG pathway database [44]. Using these data, SEPEA_NT3 and SEPEA_NT3* (which is essentially the SEPEA_NT3 analysis but does not make use of the network information of the pathways and is very similar to those used in [7,9,11-22]) was used to find the enriched human pathways associated; the results are given in Table 7. Several of the pathways known to be important for breast cancer initiation and progression are significant using either method, such as the ErbB, p53, and apoptosis pathways. In contrast, the adherens junction, regulation of actin cytoskeleton, cell adhesion molecules, and focal adhesion pathways are significant using SEPEA_NT3, but are not considered significant using the SEPEA_NT3* method ($P \leq 0.05$). These pathways, in particular the focal and cell adhesion pathways, all deal with cell to cell communication and are thought to be key modulators of progression and invasion of malignant phenotypic characteristics [50]. In fact, several novel cancer chemotherapy drugs are being designed to specifically act on the focal adhesion pathway and

**Table 5**

| KEGG pathway ID | Pathway description | P-value |
|-----------------|---------------------|---------|
| [path:xla03022] | Basal transcription factors | 0.01 |
| [path:xla04010] | MAPK signaling | 0.02 |
| [path:xla00460] | Cyanoamino acid metabolism | 0.024 |
| [path:xla00550] | Peptidoglycan biosynthesis | 0.031 |
| [path:xla02010] | ABC transporters | 0.045 |
| [path:xla03050] | Proteasome | 0.05 |
| [path:xla00982] | Drug metabolism - cytochrome P450 | 0.053 |
| [path:xla00830] | Retinol metabolism | 0.059 |
| [path:xla04630] | Jak-STAT signaling | 0.07 |
| [path:xla04012] | ErbB signaling | 0.1 |

Enriched KEGG [44] pathways (with $P$-value $\leq 0.1$) due to cycloamine treatment of developing X. laevis, designed to inhibit SHH signaling, using microarray data from GEO [45] [GEO:GSE8293]. P-values were obtained using the SEPEA_NT2 analysis with 1,000 randomizations to compute significance.

**Table 6**

| KEGG pathway ID | Pathway description | P-value |
|-----------------|---------------------|---------|
| [path:xla00930] | Caprolactam degradation | 0.006 |
| [path:xla03030] | DNA replication | 0.011 |
| [path:xla00480] | Glutathione metabolism | 0.016 |
| [path:xla00561] | Glycerolipid metabolism | 0.023 |
| [path:xla03010] | Ribosome | 0.045 |
| [path:xla00982] | Drug metabolism - cytochrome P450 | 0.057 |
| [path:xla00983] | Drug metabolism - other enzymes | 0.057 |
| [path:xla04012] | ErbB signaling | 0.067 |
| [path:xla03060] | Protein export | 0.072 |
| [path:xla00562] | Inositol phosphate metabolism | 0.086 |
| [path:xla04914] | Progesterone-mediated oocyte maturation | 0.087 |
| [path:xla04020] | Calcium signaling pathway | 0.089 |

Enriched KEGG [44] pathways (with $P$-value $\leq 0.1$) due to cycloamine treatment of developing X. laevis, designed to inhibit SHH signaling, using microarray data from GEO [45] [GEO:GSE8293]. P-values were obtained using the SEPEA_NT2* analysis with 1,000 randomizations to compute significance.
many standard chemotherapy drugs modulate this pathway in conjunction with their primary mode of action [51]. So this analysis again suggests gains in the pathway enrichment analysis when network details of pathways are incorporated in the analysis.

Conclusions

This paper presents a new method that uses biological data in order to find biochemical pathways that are relevant to the different responses of an organism to two different conditions. Biochemical pathways, instead of being treated as just sets of genes, are viewed as a network of interactions between proteins or metabolites. The extensive analysis using simulated and real data clearly demonstrates the utility of incorporating information on the interactions between the genes present in a pathway network.

Materials and methods

Notation

Assume there are $m$ genes (identified by indices in the set $G = \{1, 2, \ldots, m\}$) in the system and $n$ array measurements ($n_c$ control and $n_t$ treated, $n_c + n_t = n$) per gene. We will analyze one particular pathway made up of a subset $m_p$ of the $m$ genes in the system. Without loss of generality, assume that these genes correspond to the first $m_p$ gene indices in $G$. The genes in this pathway are part of an underlying network of their gene products. On the basis of this network, gene $i$ of the pathway is assigned a weight $w_i$ and a gene pair ($i$ and $j$) is assigned two weights $d_{ij}$ (denoting a measure of the distance between these two genes on the network) and $e_{ij}$ (which is equal to 1 for a non-zero value of $d_{ij}$). Each of the $m$ genes is also assigned a value, $t_{stat, k}$ for gene $k$ capturing the treatment effect on it as found in the observed data. This value obtained under the different null distributions (as defined in the next section) is denoted by $T_{stat, k}$. The two scores, from the Heavy Ends Rule and the Distance Rule are denoted by $H_R$ and $D_R$, respectively. They are a function of $t_{stat, k}$. $HER_{obs}$ and $DR_{obs}$ denote those obtained from the observed experimental data while $HER_{rand}$ and $DR_{rand}$ those obtained from the different null distributions.

Null hypotheses

Null hypotheses for the three statistical tests performed are given below and share similarities with those stated in [6].

Table 7

| KEGG pathway ID | Pathway description            | SEPEA_NT3 | SEPEA_NT3* |
|-----------------|-------------------------------|-----------|------------|
| [path:hsa04370] | VEGF signaling pathway        | 1.69E-04  | 5.14E-04   |
| [path:hsa04662] | B-cell receptor signaling     | 3.32E-04  | 3.51E-04   |
| [path:hsa04630] | Jak-STAT signaling           | 8.91E-04  | 0.0417     |
| [path:hsa04520] | Adherens junction             | 0.0014    | 0.1438     |
| [path:hsa04810] | Regulation of actin cytoskeleton | 0.0027  | 0.0717     |
| [path:hsa04150] | mTOR signaling                | 0.0047    | 0.0052     |
| [path:hsa04664] | Fc epsilon RI signaling       | 0.0081    | 5.99E-04   |
| [path:hsa04510] | Focal adhesion                | 0.0103    | 0.0648     |
| [path:hsa04012] | ErbB signaling                | 0.0103    | 8.51E-04   |
| [path:hsa04210] | Apoptosis                     | 0.0108    | 7.97E-04   |
| [path:hsa03440] | Homologous recombination      | 0.0147    | 0.0016     |
| [path:hsa04660] | T cell receptor signaling     | 0.0182    | 0.001      |
| [path:hsa04010] | MAPK signaling                | 0.0183    | 0.0183     |
| [path:hsa04910] | Insulin signaling             | 0.0191    | 0.0032     |
| [path:hsa04514] | Cell adhesion molecules       | 0.0274    | 0.2407     |
| [path:hsa04115] | P53 signaling                 | 0.0306    | 0.0093     |
| [path:hsa04620] | Toll-like receptor signaling pathway | 0.0391 | 0.0193 |

Enriched KEGG [44] pathways (with $P$-value $\leq 0.05$) obtained using genes from the OMIM database [49] that confer susceptibility to breast cancer. $P$-values were obtained using the SEPEA_NT3 and SEPEA_NT3* analysis.
Network test 3 (NT3): $T^{\text{stat}}, i = 1, 2, ..., m$ are independent and identically distributed with a common distribution, $F$ (which can take any form).

In all three hypotheses, $\text{HER}_{\text{obs}}$ and $\text{DR}_{\text{obs}}$ are each drawn from the distribution of $\text{HER}_{\text{rand}}$ and $\text{DR}_{\text{rand}}$, respectively.

**Association value computations**

For each gene we define by a pair of values $(t^+_i, t^-_i)$ corresponding to the association with the treatment in the context of the observed data. The association of any given gene with treatment is given in terms of the square of the two-sample t-statistic (similar to what has been done in [6,25,35]) and also shares similarities with the $\text{maxmean}$ statistic defined in [10]. Mathematically:

$$t_{\text{stat}, i} = \frac{(x^C_i - x_i)}{\sqrt{\left(\frac{s^2_i}{n_t} + \frac{s^C_i}{n_c}\right)}}$$  \hspace{1cm} (1)

$$t^+_i = \left(1 - \frac{r^+_i}{m}\right) \cdot \max\left(t_{\text{stat}, i}, 0\right)^2$$  \hspace{1cm} (2)

$$t^-_i = \left(1 - \frac{r^-_i}{m}\right) \cdot \min\left(t_{\text{stat}, i}, 0\right)^2$$

$$CF = \max\left\{\frac{\text{mean}(t^2_{\text{stat}, i})^{mP}}{\text{var}(t^2_{\text{stat}, i})^{mP}} - \frac{\text{mean}(t^2_{\text{stat}, i})^{-mP+1}}{\text{var}(t^2_{\text{stat}, i})^{-mP+1}}\right\}, 0$$  \hspace{1cm} (3)

where $x^C_i$, $x_i$ are the sample mean expression for gene $g_i$ of the control and treated data, respectively, $s^2_i$, $s^C_i$ are the associated standard deviations, $I_{NT1}$ is equal to 1 when the NT1 test is being used and is equal to zero otherwise, $r^+_i$ denotes the position of gene $i$ in the sorted (in descending order) list of $\max(t_{\text{stat}, i}, 0)$ over all the $m$ genes, and, similarly, $r^-_i$ denotes the position of gene $i$ in the sorted (in ascending order) list of $\min(t_{\text{stat}, i}, 0)$. $a$ and $b$ are parameters chosen empirically in order to control for the selection of the pathway with the most significant genes (relative to the other genes in the system). The first terms in the products on the right-hand side of Equation 2 will be called importance factors for a gene. These are values between 0 and 1. The functions ‘mean’ and ‘var’ refer to the standard definitions of mean and variance.

The term $CF$ denotes a (competitive) factor that is a measure of difference in the mean of differential expression of the genes in the pathway and that of the other genes in the system. Higher $CF$ values indicate higher individual association values for genes in the pathway relative to the other genes and vice versa. Therefore, for similar values for changes in gene expression ($t_{\text{stat}, i}$) the power to detect treatment effect decreases as the $CF$ factor decreases (or as more genes in the system are affected as a result of the treatment). For high values of the $CF$ factor, parameter $a$ controls the (decreasing) importance of genes along the sorted list. The parameter $b$ provides a much steeper decrease in the importance of genes down the sorted list for small values of the $CF$ factor.

Here, $t_{\text{stat}, i}$ is the standard two sample t-statistic. In some instances, the only information of the association of a gene with a treated condition may be just a summary statistic. For example, there are a set of known gene polymorphisms associated with breast cancer; in trying to identify pathways relevant for breast cancer, these genes would then be arbitrarily assigned a $t_{\text{stat}, i}$ equal to 1 while the other genes would be given values of 0. Note that in these situations, $n$, the number of array measurements per gene, is zero.

**Definition of the scoring rules**

The score for linking the observed expression data to a given pathway has two components. The first component is called the Heavy Ends Rule score $\text{HER}_{\text{obs}}$ and will have a high value when a combination of the more 'important' genes (those associated with gene products close to a terminal of a pathway) is significantly associated with the treated condition. The second component called the Distance Rule score $\text{DR}_{\text{obs}}$ has a high value when the genes that are significantly associated with the treated condition have their gene products located close together. It is in fact the reciprocal of the weighted average distance between the genes in the network. The weights $w_i$, $d_{ij}$ and $e_{ij}$ are defined in a subsequent section. Each score is defined as the maximum of individual expressions dependent either only on the genes whose expression increased due to the treatment or on the genes whose expression decreased as a result of the treatment. This should make it more robust to detect changes in both scale and location as discussed in [10]. The two scores are defined as:

$$\text{HER}_{\text{obs}} = \max\left\{\sum_{i=1}^{mP} w_i t^+_i, \sum_{i=1}^{m} w_i t^-_i\right\}$$

$$\text{DR}_{\text{obs}} = \max\left\{\sum_{i=1}^{mP} \sum_{j=1}^{mP} t^+_i t^+_j e_{ij}, \sum_{i=1}^{m} \sum_{j=1}^{m} t^-_i t^-_j e_{ij}\right\}$$

$$\sum_{i=1}^{mP} \sum_{j=1}^{mP} t^+_i t^+_j d_{ij}, \sum_{i=1}^{m} \sum_{j=1}^{m} t^-_i t^-_j d_{ij}\right\}$$

(4)
For the \( DR \) score computation, \( 0/0 \) is defined to be equal to zero. The scores obtained under the null distributions are denoted by \( \text{HER}_{\text{rand}} \) and \( \text{DR}_{\text{rand}} \) and are defined as in Equation 4 with \( t_i \) replaced by \( T_i \).

**Test statistic and significance evaluation**

For each of the three hypotheses (NT1, NT2 or NT3) the test statistic is defined as:

\[
S = \frac{\text{HER}-\text{mean}(\text{HER})_{NT}}{\text{std}(\text{HER})_{NT}} + \frac{\text{DR}-\text{mean}(\text{DR})_{NT}}{\text{std}(\text{DR})_{NT}}
\]

(5)

where \( \text{mean}(\text{HER}) \) and \( \text{std}(\text{HER}) \) refer to the mean and standard deviation of the \( \text{HER} \) score for the given test and \( \text{mean}(\text{DR}) \) and \( \text{std}(\text{DR}) \) are those for the \( \text{DR} \) score.

For the NT1 and NT2 tests, multiple random samples of arrays are taken from the common set of treated and control data (without replacement) and randomly assigned to control or treated groups. For each random sample, the \( T_{\text{stat},i} \) s are calculated and then \( \text{HER}_{\text{rand}} \) and \( \text{DR}_{\text{rand}} \) are computed. The NT1 test requires \( T_{\text{stat},i} \) to be computed for all the \( m \) genes while the NT2 test requires computation for just the \( m_p \) genes that are part of the pathway. For the NT3 test, multiple random samples of \( m_p T_{\text{stat},i} \) s are drawn from the global set of \( m \) observed \( t_{\text{stat},i} \).

The estimate of the \( P \)-value for each of the tests is computed as:

\[
p = \sum_{\text{randomizations}} \frac{I(S_{i} \geq S_{\text{obs}})}{\# \text{randomizations}}
\]

(6)

where \( I(S_{i} \geq S_{\text{obs}}) \) is an indicator function that equals 1 when the \( i \)th randomly estimated test statistic value, \( S_{i} \), equals or exceeds the observed value and 0 otherwise. The estimation procedure used for the special case when the data are in the form of a list of differentially expressed genes or a list of genes associated with a disease is provided in Additional data file 1.

The way the significance computations are performed, tests NT1 and NT3 could be viewed as belonging to the class of 'competitive' hypotheses (as elaborated in the Background section) while NT2 could be viewed as a 'self-contained' hypothesis.

The method when applied to each of the three null hypotheses NT1, NT2 and NT3 is denoted by SEPEA_NT1, SEPEA_NT2 and SEPEA_NT3, respectively.

**Generation of simulated data**

Data were simulated from two genetic systems (\( \text{Linear} (L) \) and \( \text{ErbbSignaling} (E) \)) of 500 genes \( \{g_n^L\}_{n=1,2,...,500} \) and \( \{g_n^E\}_{n=1,2,...,500} \). Each system had two subnetworks of interest and each subnetwork was assumed to have no interactions with the other subnetwork. The \( \text{Linear} \) network had a set of 30 genes \( \{g_n^L\}_{n=1,2,...,30} \) that were connected in a linear fashion (Figure 3a). A set of 87 genes \( \{g_n^E\}_{n=1,2,...,87} \) in the \( \text{ErbbSignaling} \) network interacted in the same manner as described by the Erbb signaling pathway in the KEGG pathway database [44] (Figure 3b). Pathway enrichment analysis was performed on these two subnetworks.

Each set \( \Lambda \) and \( H \) had a subset of genes (with indices \( \{i_j\}_{j=1}^{n_{corr}} \) ), \( \Sigma = \{g_{i_j}^\Lambda\}_{j=1}^{n_{corr}} \), whose expressions were perfectly correlated with each other (\( \Sigma^L \) had \( n_{corr} = 0 \) or 9 genes and \( \Sigma^E \) had \( n_{corr} = 7 \) genes). The gene expressions in the complement of each of the sets \( \Sigma^L \) and \( \Sigma^E \), \( (\Sigma^L)^c \) and \( (\Sigma^E)^c \), were assumed to be independent of each other even though some of them could be assumed to be known to have gene products that interact with gene products of genes in \( \Sigma^L \) and \( \Sigma^E \). This could be justified by the fact that the interaction was not at the gene expression level and involved changes in the phosphorylation/binding states of the protein, for example. Let \( \{\text{erb}_{j}\}_{j=1}^{7} \) denote the set of gene indices associated with the proteins circled in Figure 3b, ordered from left to right. The random variable defining the gene expression of gene \( g_n \) is denoted by \( X_n \).

Let \( N(\mu, \sigma) \) represent the normal probability distribution with mean \( \mu \) and standard deviation \( \sigma \). Then data for all the 500 genes in each of the two systems were generated for one experiment under control conditions in the following manner:

\[
X_{ij} = N(10,1) \\
X_{ij} = X_{i,j-1} + 2, j = 2, \ldots, n_{corr} \\
X_n = N(10,1), n \in \{k\}_{k=1}^{500} \setminus \{i_j\}_{j=1}^{n_{corr}}
\]

(7)

Let \( \Phi (\Phi^L \) and \( \Phi^E) \) denote the set of genes that are direct targets of the treatment. The total number of genes in the system affected by the treatment (that includes the set \( \Phi \)) was chosen to be 50 and 10 for the \( \text{Linear} \) and \( \text{ErbbSignaling} \) networks, respectively. The effect of the treatment was to increase the mean of the expressions of the direct targets by a factor \( \text{pert} \), \( \mu = \text{pert} \cdot \mu \). Results from the assignment \( \text{pert} = 1.2 \) are discussed here while those resulting from other assignments are discussed in Table S3 in Additional data file 1. Let \( L^L \) and \( L^E \) denote a uniformly random selection of \( n_{corr} \) genes from the sets \( \Lambda \) and \( H \), respectively, let \( V_n^L \) and \( V_n^E \) denote sets of \( n \) genes drawn from the complements of the sets \( \Lambda \) and \( H \), respectively, and let \( \emptyset \) denote the empty set. The details of the different correlation patterns considered here are given in Table 1. Patterns 1 and 3 were the correlation patterns that were favored by the scoring rules described in this paper.
All methods in this paper were coded using the Java programming language. For each combination of correlation pattern and pert assignment, 1,000 independent experiments were simulated. Each experiment involved the generation of $n_c = 5$ control samples and $n_t = 5$ treatment samples. For the randomization tests of each method, 1,000 randomizations were chosen. The performance measures chosen were the number of experiments out of the 1,000 performed that resulted in $P$-values for the test (Equation 6) below different chosen significance levels. The methods evaluated were GSEA [35], maxmean [10], SEPEA_N1, SEPEA_N2 and SEPEA_N3. For the SEPEA_N1 method, the parameters $a$ and $b$ in Equation 2 were empirically set to equal 2 and 5, respectively. Parameter $a = 2$ provides a quadratic decrease in the importance of genes along the sorted list for high values of the CF factor (when the mean changes in expression of the genes in the pathway are higher than that of the rest of the genes in the system). In the situation of low values of the CF factor, the
value $b = 5$ was chosen such that the top 20% of genes in the sorted list approximately receive importance in the interval $(0.2, 1)$ while the remaining genes receive weights in the interval $(0, 0.2)$. Results from GSEA [35], maxmean [10] and SEPEA-NT1 are comparable because all test a similar null hypothesis. The main difference between these methods is that while GSEA and maxmean are blind to the structure of the biochemical pathway, SEPEA-NT1 is not.

### Assignment of network weights

The pathway network is represented by a set of nodes/gene products and set of edges between these nodes. The nodes represent gene products such as individual proteins or protein complexes. There is an edge from node/protein $u$ to node/protein $v$ if $u$ transfers the signal it received immediately to $v$ (either in the form of increasing the transcription of genes associated with $v$, changing the phosphorylation state of $v$, causing disassociation of $v$ from a complex that it is part of) in the case of signaling pathways or that $u$ and $v$ catalyze two successive reactions in the case of metabolic pathways.

Let $\{v_k\}_{k=1,2,...,p}$ denote the set of $P$ nodes of the network and $\{g_a\}_{a=1,2,...,m_p}$ denote the set of $N$ genes associated with the nodes. The number of edges entering node $v_i$ is defined as its in-degree and the number of edges leaving $v_i$ is defined as its out-degree. We define a node to be a terminal node if either its in-degree or out-degree is zero.

Assume that each edge represents a unit distance between the two nodes that it connects. So if the shortest route between two nodes is via two edges in the pathway network, assuming that there may be more than one path connecting the two nodes in the pathway network. Let $\delta_i$ denote the shortest distance of node $v_i$ to a terminal node of the pathway. Let $G(v_i, g_a)$ denote the indicator function, which is equal to 1 when gene $g_a$ is associated with node $v_i$ and is equal to 0 otherwise. The number of genes associated with node $v_i$ is denoted by $N_i$. Let $s_{ij}$ denote the distance from node $v_i$ to node $v_j$ in the network. $s_{ij}$ is assigned a value of 0 either when $i = j$ or when node $v_j$ is unreachable from node $v_i$. Define the positive indicator function, $I^+(x)$, which is equal to 1 when $x$ is positive and equal to 0 otherwise.

The weights for gene $g_a$, $w_a$, and gene pair $(g_a, g_b)$, $d_{ab}$ and $e_{ab}$, are given by:

$$w_a = \sum_{i=1}^{p} G(v_i, g_a) \left(1 - \frac{\delta_i}{\text{max}_{k=1,2,..,p}(\delta_k)+1}\right)$$

if the pathway is a signaling on $v_j$, $d_{ab} = \sum_{i=1}^{p} \sum_{j=1}^{p} G(v_i, g_a) G(v_j, g_b) s_{ij}$

$$e_{mn} = \sum_{i=1}^{p} \sum_{j=1}^{p} G(v_i, g_a) G(v_j, g_b) I^+(s_{ij})$$

(8)

The weight $w_a$ is defined such that genes associated with nodes closer to the terminal nodes have higher weights than those that are further away. The choice of a linear function to capture the intuition behind the HER is arbitrary and other functions will be experimented with as part of future work. The non-zero weights $d_{ab}$ for genes $a$ and $b$ are smaller if they are associated with gene products that are closer together in the pathway network than for pairs of genes whose gene products are further away.

### Statistical test for Distance Rule justification

Let the total number of pathways (nodes) in the network in Figure S1 in Additional data file 2 be denoted by $N_p$. Denote the distance between pathways $i$ and $j$ on this pathway network by $d^p_{ij}$. Define $d^p_{ij}$ to be equal to zero if pathway $j$ is not reachable from pathway $i$. Also define variable $e^p_{ij}$, which is equal to 1 for all non-zero values of the corresponding $d^p_{ij}$ and 0 otherwise. Perturbations to one pathway are transferred across the edges of the network to multiple pathways. Using human microarray data randomly chosen from the GEO database [45], we considered eight comparisons between two conditions (Table 1). For each comparison, the $DR$ score was computed (Equation 4) for every human pathway on the network of pathways described above. In order to make the comparison possible across all the pathways, the $DR$ scores obtained above using experimental data were normalized with $DR$ scores obtained by setting the $T_{stat,i}$ values for all the genes equal to 1. Let the normalized $DR$ score for pathway $i$ be denoted by $\overline{DR}_i$. A meta score can now be defined on the pathway network as follows:

$$\text{meta}_DR = \frac{\sum_{i=1}^{N_p} \sum_{j=1}^{N_p} DR_i DR_j e^p_{ij}}{\sum_{i=1}^{N_p} \sum_{j=1}^{N_p} DR_i DR_j d^p_{ij}}$$

(9)

Higher values of $\text{meta}_DR$ would indicate that pathways with higher values of the normalized $DR$ scores $\overline{DR}_i$ are closer to
each other. The significance of the obtained meta\_DR scores are tested using random networks generated by the Markov-chain switching algorithm [52]. The properties of these random networks are that they have the same number of nodes and edges as the original pathway network and the degree sequence among all the nodes is also maintained. These networks differ, however, from the original network due to a number of random edge swaps across the network.

**GeneChip experiments**

Cyclopamine powder (11-deoxyjervine; Toronto Research Chemicals Inc., North York, Ontario, Canada) was dissolved in 100% ethanol to a concentration of 5 mg/ml and this stock solution was stored at -20°C. A similar volume of 100% ethanol was stored at -20°C for use in vehicle control exposures. Approximately 200 tadpoles from each of two clutches (designated ‘clutch A’ and ‘clutch B’) of the species *Xenopus laevis* were obtained from Nasco Biology (Fort Atkinson, WI, USA) for a total of approximately 400 tadpoles. Animals were raised at an air temperature of 25 ± 1°C in tanks of 9 liters of tap water treated with Stress Coat (Aquarium Pharmaceuticals, Chalfont, PA, USA) and aged 1 day. Each day for three consecutive days, as animals reached stage 52 [53], the population of stage 52 individuals from each clutch was removed from the clutch tanks and divided in half indiscriminately, resulting in four exposure groups per day: a control group for clutch A; an experimental group for clutch A; a control group for clutch B; and an experimental group for clutch B. Each exposure tank had between 10 and 20 individuals, with 150 ml treated water per individual. After sorting into exposure tanks, 30 μl per animal of 5 mg/ml cyclopamine solution was added to all experimental tanks, and 30 μl per animal of 100% ethanol was added to each control tank. After 24 hours of exposure, animals were sacrificed by over-anesthesia with MS222, dried on a paper towel, then put into vials of RNAlater (Ambion, Austin, TX, USA). Vials were kept at 4°C overnight, then moved to -20°C for storage. Both hindlimb buds were dissected off each animal at the base of the limb using surgical scissors, placed in fresh vials of RNAlater, and returned to -20°C for continued storage. RNA extractions were performed using the RNeasy Mini Kit and optional RNase-Free DNase Set (QIAGEN, Valencia, CA, USA), with the following notes: limbs were put into a 1.5 ml microcentrifuge tube, residual RNAlater was pipetted off, and limbs were crushed with a homogenizer in 200 μl buffer RLT, then 300 μl more buffer RLT was added; and elution was carried out with two washes of 50 μl RNase-free water. Extracted total RNA was stored at -80°C and transferred to the WM Keck Foundation Biotechnology Resource Center, Affymetrix Resource Center (Yale University, New Haven, CT), where they were again run through DNase treatment. Four control-experimental pairs of samples were chosen, from a total of 12 pairs, based on quantity and quality of RNA as determined by analysis on an Agilent 2100 Bioanalyzer RNA Nano chip (Agilent Technologies Inc., Santa Clara, CA, USA). Samples in each pairwise comparison were extracted from the same number of limbs, were from the same clutch, were exposed to cyclopamine solution or ethanol on the same day, and their total RNA was extracted in the same batch of extractions. The eight chosen samples were each hybridized to an Affymetrix® GeneChip® *Xenopus laevis* Genome Array (Affymetrix, Santa Clara, CA, USA) using 3 μg total RNA. Data have been deposited in the National Center for Biotechnology Information, NCBI GEO with series record ID [GEO:GSE8293].

**Abbreviations**

DR: Distance Rule; GEO: Gene Expression Omnibus; GSEA: gene set enrichment analysis; HER: Heavy Ends Rule; KEGG: Kyoto Encyclopedia of Genes and Genomes; MAPK: mitogen-activated protein kinase; NT: network test; OMIM: Online Mendelian Inheritance in Man; SEPEA: structurally enhanced pathway enrichment analysis; SHH: Sonic hedgehog.

**Authors’ contributions**

RT and JMG designed and evaluated the research with important suggestions from CJP and FMP. RT implemented the research and drafted the manuscript. GFS performed the *Xenopus laevis* experiments. All the authors read and approved the final manuscript.

**Additional data files**

The following additional data are available with the online version of this paper: a Word document that provides a section on a particular estimation of P-values and additional tables of results (Additional data file 1); a figure of the network of pathways in the KEGG pathway database [44] (Additional data file 2); a figure that demonstrates the receiver-operator characteristics of the different methodologies used (Additional data file 3); a table with the P-values for KEGG pathways after cyclopamine treatment of developing *X. laevis*, designed to inhibit SHH signaling, using microarray data from GEO [45] [GEO:GSE8293] (Additional data file 4).

**Acknowledgements**

This research was supported by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences (NIEHS). RT especially thanks Dr Shyamal Peddada of the Biostatistics Branch at NIEHS for valuable suggestions regarding the methodology. The *Xenopus* cyclopamine exposure experiments were performed by GFS in Yale University’s Department of Ecology and Evolutionary Biology, and were supported in part by a grant to Gunter P Wagner by the Yale Core Center for Musculoskeletal Disorders, which is funded by NIH grant P30 AR-46032 from the National Institute of Arthritis and Musculoskeletal and Skin Diseases (NIAMS).

**References**

1. Eisen MB, Spellman PT, Brown PO, Botstein D: Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA* 1998, 95:14863-14868.

2. Raychaudhuri S, Stuart JM, Altman RB: Principal components analysis for high-dimensional data from genomics.
analysis to summarize microarray experiments: application to spatiotemporal series. Proc Symp Biocomput 2000, S:455-466.
2. Toronen P, Kolehmainen M, Wong G, Castren E: Analysis of gene expression data using self-organizing maps. FEBS Lett 1999, 451:142-146.
3. Goeman JJ, Buhmann P: Analyzing gene expression data in terms of gene sets: methodological issues. Bioinformatics 2007, 23:980-987.
4. Liu Q, Dinu I, Adedewale Aj, Potter JD, Yashu Y: Comparative evaluation of gene-set analysis methods. BMC Bioinformatics 2007, 8:431.
5. Newton MA, Quintana FA, den Boon JA, Sengupta S, Ahlquist P: Random-set methods identify distinct aspects of the enrichment signal in gene-set analysis. Ann Appl Stat 2008, 2:286-315.
6. Newton MA, Quintana FA, den Boon JA, Sengupta S, Ahlquist P: Random-set methods identify distinct aspects of the enrichment signal in gene-set analysis. Ann Appl Stat 2007, 1:85-106.
7. Newton MA, Quintana FA, den Boon JA, Sengupta S, Ahlquist P: Random-set methods identify distinct aspects of the enrichment signal in gene-set analysis. Ann Appl Stat 2007, 1:85-106.
8. Newton MA, Quintana FA, den Boon JA, Sengupta S, Ahlquist P: Random-set methods identify distinct aspects of the enrichment signal in gene-set analysis. Ann Appl Stat 2007, 1:85-106.
9. Beissbarth T, Speed TP: Gene enrichment analysis of groups of genes in expression profiling studies. Proc Nat Acad Sci USA 2005, 102:13544-13549.
10. Efron B, Tibshirani R: Estimating the significance of the set of genes. Ann Appl Stat 2007, 1:107-129.
11. Al-Shahrour F, Arbia A, Dopoza H, Huerta-Cepas J, Minguiz P, Monter D, Dopazo J: From genes to functional classes in the study of biological systems. BMC Bioinformatics 2007, 8:1-14.
12. Al-Shahrour F, Dopoza J, Dazio-Ariarte R, Mataos A, Vasquez J, Mon D, Dopazo J, GEFA: A web tool for finding significant associations of Gene Ontology terms with groups of genes. Bioinformatics 2004, 20:578-580.
13. Beissbarth T, Speed TP: GStat: faithfully overrepresented Gene Ontologies within a group of genes. Bioinformatics 2004, 20:2156-2158.
14. Gross P, Tsoukridis J, Hard D, Cavali E: Pathway processor: A tool for integrating whole-genome expression results into metabolic networks. Genome Res 2002, 12:1121-1126.
15. Herrero J, Al-Shahrour F, Dazio-Ariarte R, Mataos A, Vasquez J, Mando D, Dopazo J, GEFA: A web tool for finding significant associations of Gene Ontology terms with groups of genes. Bioinformatics 2004, 20:578-580.
16. Qiu P, Townsend J, Hard D, Cavali E: Pathway processor: A tool for integrating whole-genome expression results into metabolic networks. Genome Res 2002, 12:1121-1126.
17. Pan DY, Sun N, Cheung KH, Guan Z, Ma LG, Holford M, Deng XW, Zhao HY: PathMAPA: a tool for displaying gene expression data and performing statistical tests on metabolic pathways at multiple levels for Arabidopsis. BMC Bioinformatics 2003, 4:56.
18. Pandey R, Guru RK, Mount D: Pathway Miner: extracting gene association networks from molecular pathways for predicting the biological significance of gene expression microarray data. Nucleic Acids Res 2004, 32:346-347.
19. Segal E, Friedman N, Koller D, Regev A: A module map showing functional and experimental relevance of high-throughput gene expression experiments. Nucleic Acids Res 2004, 32:W449-W456.
20. Pan DY, Sun N, Cheung KH, Guan Z, Ma LG, Holford M, Deng XW, Zhao HY: PathMAPA: a tool for displaying gene expression data and performing statistical tests on metabolic pathways at multiple levels for Arabidopsis. BMC Bioinformatics 2003, 4:56.
21. Pan DY, Sun N, Cheung KH, Guan Z, Ma LG, Holford M, Deng XW, Zhao HY: PathMAPA: a tool for displaying gene expression data and performing statistical tests on metabolic pathways at multiple levels for Arabidopsis. BMC Bioinformatics 2003, 4:56.
22. Zeeberg BR, Feng WM, Wang G, Wang MD, Fojo AT, Sunshine M, Kawashima S, Katayama T, Araki M, Hiramakawa M: From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res 2006, 34:D354-357.
23. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Sobolev A, Tomashevsky M, Edgar R: NCBI GEO: mining tens of millions of expression profiles - database and tools update. Nucleic Acids Res 2007, 35:D760-765.
24. Raponi M, Zhang Y, Yu J, Chen G, Lee G, Taylor JF, McAdams J, Thomas D, Moskaluk C, Wang Y, Beer DG: Gene expression signatures for predicting prognosis of squamous cell and adenocarcinomas of the lung. Cancer Res 2006, 66:7466-7472.
25. Elia D, Madhala D, Ardon E, Reshef R, Haley O: Sonic hedgehog promotes proliferation and differentiation of adult muscle cells: Involvement of MAPK/ERK and PI3K/Akt pathways. Biochim Biophys Acta 2007, 1773:1438-1446.
26. Osawa H, Ohnishi H, Takano K, Noguti T, Masuda H, Hoshino H, Kita H, Sato K, Matsui H, Sugano K: Sonic hedgehog stimulates the proliferation of rat gastric mucosal cells through ERK activation by elevating extracellular calcium concentration. Biochem Biophys Res Commun 2006, 344:680-687.
27. Online Mendelian Inheritance in Man (OMIM) [http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?DB=OMIM&cmd=Retrieve&dopt=Summary&uid=15018351]
50. Behmoaram E, Bijian K, Bismar TA, Alaoui-Jamali MA: Early stage cancer cell invasion: signaling, biomarkers and therapeutic targeting. Front Biosci 2008, 13:6314-6325.
51. Chatzizacharias NA, Kouraklis GP, Theocharis SE: Focal adhesion kinase: a promising target for anticancer therapy. Expert Opin Ther Targets 2007, 11:1315-1328.
52. Maslov S, Sneppen K: Specificity and stability in topology of protein networks. Science 2002, 296:910-913.
53. Nieuwkoop PD, Faber J: Normal Table of Xenopus laevis (Daudin): A Systematical and Chronological Survey of the Development from the Fertilized Egg Till the End of Metamorphosis. New York: Routledge; 1994.