A systems biology approach for pathway level analysis (Supplementary materials)

Sorin Drăghici[1,2]*, Purvesh Khatri[2], Adi Laurentiu Tarca[3], Kashyap Amin[2],
Arina Done[2], Calin Voichita[2], Constantin Georgescu[2] and Roberto Romero[3]
[1]Karmanos Cancer Institute, Wayne State University
[2]Department of Computer Science, Wayne State University
[3]Perinatology Research Branch, NIH/NICHD, Detroit, MI 48201

June 27, 2007

Contents

* *To whom the correspondence should be addressed. E-mail: sorin@wayne.edu, Fax: +1-313-577-6868*

1 A statistical perspective on the impact analysis

The approach proposed here evaluates the strength of the null hypothesis H_0 (that the pathway is not significant), by combining two types of evidence. In a first analysis, a classical over-representation analysis (ORA) approach provides a p-value defined as the probability that the number of differentially expressed genes, X, is larger than or equal to the observed number of differentially expressed genes, N_{de}, just by chance (when the null hypothesis H_0 is true):

$$p_i = P(X \geq N_{de}|H_0) \tag{1}$$

Next, in a separate perturbation analysis, the impact of topology, gene interactions, and gene fold changes come into play and are captured thought the pathway perturbation factor:

$$PF = \frac{\sum_{g \in P_i} |PF(g)|}{\overline{|\Delta E|} \cdot N_{de}(P_i)} \tag{2}$$

where $N_{de}(P_i)$ is the number of differentially expressed genes on the given pathway P_i, $PF(g)$ is the perturbation of the gene g:

$$PF(g) = \Delta E(g) + \sum_{u \in US_g} \beta_{ug} \cdot \frac{PF(u)}{N_{ds}(u)} \tag{3}$$

and $\overline{|\Delta E|}$ is the mean fold change over the entire set of N differentially expressed genes:

$$\overline{|\Delta E|} = \frac{\sum_{k=1}^{N} |\Delta E|}{N} \tag{4}$$

Let PF denote the perturbation factor as a random variable and pf be the observed value for a particular pathway. The score pf is always positive, and the higher its value, the less likely the null hypothesis (that the pathway is not significant). Moreover this likelihood decays very fast as pf gets away from zero. These features point to the exponential distribution as an appropriate model for the random variable PF: Under the null hypothesis, differentially expressed genes would fall on the pathway randomly, and would not interact with each other in any concerted way. In other words, in the second term in Eq. 3 (which captures the influence of the genes upstream) roughly half of those influences will be positives, and half negative, canceling each other out. In such circumstances, the perturbation of each gene would be limited to its own measured fold change (due to random unrelated causes):

$$PF(g) = \Delta E(g) + \sum_{u \in US_g} \beta_{ug} \cdot \frac{PF(u)}{N_{ds}(u)} = \Delta E(g) + 0 = \Delta E(g) \tag{5}$$

Consequently, under the same null hypothesis, the expected value for the perturbation of a pathway (from Eq. 2) will be:

$$E(PF) = E\left(\frac{\sum_{g \in P_i} |PF(g)|}{\overline{|\Delta E|} \cdot N_{de}(P_i)}\right) = E\left(\frac{1}{\overline{|\Delta E|}} \frac{\sum_{k=1}^{N_{de}(P_i)} |\Delta E(g)|}{N_{de}(P_i)}\right) = E\left(\frac{\overline{|\Delta E_{P_i}|}}{\overline{|\Delta E|}}\right) = 1 \tag{6}$$

The last fraction above is the ratio between the mean fold change on the given pathway, P_i, and the mean fold change in the entire data set. Under the null hypothesis, the genes are distributed randomly across pathways and the two means should be equal. Since this expected value is 1, the distribution of the random variable PF can be modeled by the exponential of mean 1, $exp(1)$.

If we use the PF score as a test statistics and assume its null distribution is exponential with mean 1, then the p-value p_{pf} resulting from the perturbation analysis will have the form:

$$p_{pf} = P(PF \geq pf|H_0) = e^{-pf} \tag{7}$$

This is the probability of observing a perturbation factor, PF, greater or equal to the one observed, pf, when the null hypothesis is true.

Let us now consider that for a given pathway we observe a perturbation factor equal to pf and a number of differentially expressed genes equal to N_{de}. A 'global' probability p_{global}, of having just by chance both a higher than expected number of differentially expressed genes AND a significant biological perturbation (large PF in the second term), can be defined as the joint probability:

$$p_{global} = P(X \geq N_{de}, PF \geq pf | H_0) \tag{8}$$

Since the pathway perturbation factor in Eq. (2) is calculated by dividing the total pathway perturbation by the number of differentially expressed genes on the given pathway, the PF will be independent of the number of differentially expressed genes X, and the joint probability above becomes a product of two single probabilities:

$$p_{global} = P(X \geq N_{de} | H_0) \cdot P(PF \geq pf | H_0) \tag{9}$$

This p_{global} provides a global significance measure that requires both a statistically significant number of differentially expressed genes on the pathway, N_{de}, and at the same time, large perturbations on the same pathway as described by pf. Using equations (1) and (7), the formula (9) becomes:

$$p_{global} = p_i \cdot e^{-pf} \tag{10}$$

We take a natural log of both sides and obtain:

$$\log(p_{global}) = \log(p_i) - pf \tag{11}$$

which can be re-written as:

$$-\log(p_{global}) = -\log(p_i) + pf \tag{12}$$

in which we can substitute the definition of pf from (2) above to yield:

$$-\log(p_{global}) = -\log(p_i) + PF \tag{13}$$

The right hand side of this expression is exactly our definition of the impact factor:

$$IF = -log(p_i) + \frac{\sum_{g \in P_i} |PF(g)|}{|\overline{\Delta E}| \cdot N_{de}(P_i)} \tag{14}$$

This shows that the proposed impact factor, IF, is in fact the negative log of the global probability of having both a statistically significant number of differentially expressed genes and a large perturbation in the given pathway.

Ignoring the discrete character of the hypergeometric distribution, under the null hypothesis $p_i = P(X \geq N_{RP} | H_0)$ has a uniform distribution. By taking negative log, the distribution changes into exponential with parameter 1, similar to the distribution we assumed for PF, the second term in IF formula.

$$-log(p_i) \sim exp(1); \;\; PF \sim exp(1); \;\; exp(1) = \Gamma(1, 1) \tag{15}$$

Then, as the sum of two independent exponential random terms, the IF will follow a Gamma distribution $\Gamma(2, 1)$ (Hogg, 1978). The pdf of this distribution is:

$$f(x) = xe^{-x}, \;\; x \geq 0 \tag{16}$$

Finally, the p-value corresponding to the observed value if of the statistic IF can be easily computed by integrating the density (16):

$$p = P(IF \geq if | H_0) = \int_{if}^{\infty} f(x)dx = \int_{if}^{\infty} xe^{-x}dx = (if + 1) * e^{-if} \tag{17}$$

2 Special cases for the impact analysis

The impact analysis proposed includes and extends the classical approach both with respect to individual genes and with respect to pathways. We discuss briefly a few interesting particular cases. These cases illustrate how, when the limitations of the classical approach are forcefully imposed (e.g., ignoring the magnitude of the measured expression changes or ignoring the regulatory interactions between genes), the impact analysis reduces to the classical approach and yields the same results.

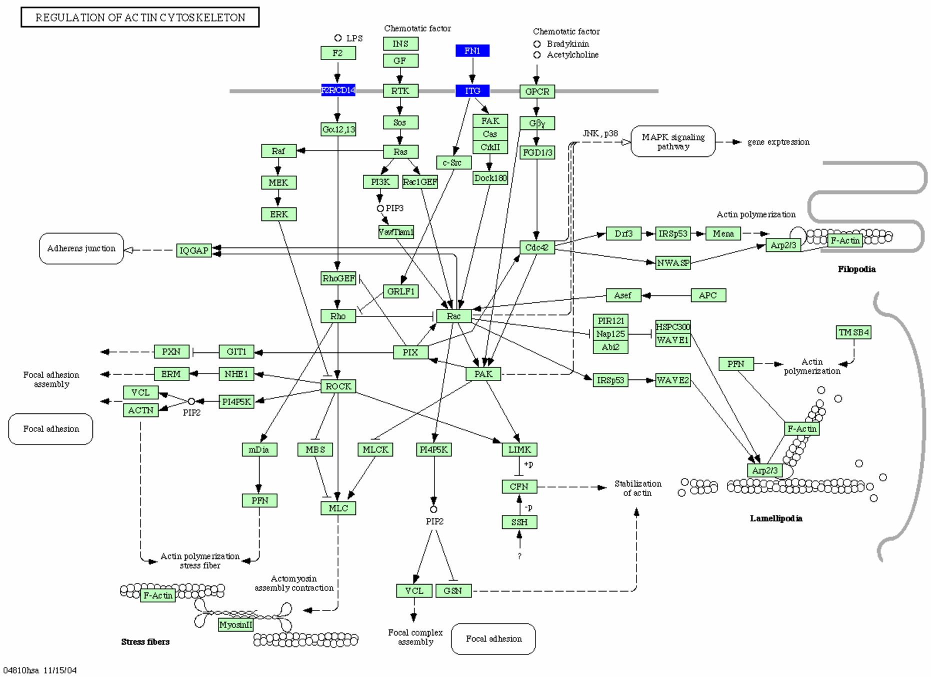


Figure 1: The hepatic cell line treated with palmitate - differentially expressed genes on the actin cytoskeleton pathway. For genes with no measured changes upstream, such as FN1 and CD14, the gene perturbation will be equal to the measured expression change. The perturbation of genes such as ITG will be higher in absolute value, reflecting both its own measured change as well as the fact that the FN1 gene immediately upstream is also differentially expressed.

2.1 Gene perturbations for genes with no upstream activity

In our analysis, the gene perturbation factor for a gene g is defined as:

$$PF(g) = \Delta E(g) + \sum_{u \in US_g} \beta_{ug} \cdot \frac{PF(u)}{N_{ds}(u)} \tag{18}$$

If there are no measured differences in the expression values of any of the genes upstream of g, $PF(u) = 0$ for all genes in US_g, and the second term becomes zero. In this case the perturbation factor reduces to:

$$PF(g) = \Delta E \tag{19}$$

This is exactly the classical approach, in which the amount of perturbation of an individual gene in a given condition is measured through its expression change ΔE. Examples could include the genes FN1 and CD14 in Fig. 1.

2.2 Pathway impact analysis when the expression change is ignored

The pathway analysis framework can also be used in the framework in which the ORA approach is usually used. If the expression changes measured for the pathway genes are to be ignored (as they are in the ORA approach), the pathway impact analysis can still be used to assess the impact of a condition upon specific pathways. This is achieved by setting all measured expression changes $\Delta E(g) = 0$ for all genes on the given pathway $g \in P_i$. This will make all gene perturbation factors zero:

$$PF(g) = \Delta E(g) + \sum_{u \in US_g} \beta_{ug} \cdot \frac{PF(u)}{N_{ds}(u)} = 0 \tag{20}$$

Assuming that there are at least some differentially expressed genes somewhere in this data set[a], (i.e. $\overline{|\Delta E|} \neq 0$) the pathway impact factor in Eq. 14 becomes:

$$IF(P_i) = log\left(\frac{1}{p_i}\right) + \frac{\sum_{g \in P_i} |PF(g)|}{\overline{|\Delta E|} \cdot N_{de}(P_i)} = log\left(\frac{1}{p_i}\right) + 0 = -log(p_i) \tag{21}$$

Since the expression now involves a single random variable, the IF values will follow a $\Gamma(1,1) = exp(1)$, rather than a $\Gamma(2,1)$ distribution, and our p value can be calculated as:

$$p = P(IF \geq -log(p_i)|H_0) = \int_{-log(p_i)}^{\infty} e^{-x} dx = -e^{-x}|_{-log(p_i)}^{\infty} = e^{log(p_i)} = p_i \tag{22}$$

This expression shows that in this particular case, the impact analysis reduces to exactly the classical approach which measures the impact of a pathway by looking exclusively at the probability of the given number of differentially expressed genes occurring just by chance, i.e., the p-value yielded by an analysis in which only the set of genes is considered.

2.3 Impact analysis involving genes with no measured expression change

It is entirely possible that certain genes are in fact changing their expression level but the change is below the sensitivity threshold of the technology, or below the threshold used to select differentially expressed genes. It is also possible that the regulation between genes happens at levels other than that of the mRNA (e.g., phosphorylation, complex formation, etc.). Hence, signals should be allowed to be propagated around the pathway even through those genes for which no expression change has been detected at the mRNA level. The perturbation factor model accounts for these situations. If the measured expression change is zero, the perturbation of the gene becomes:

$$PF(g) = \sum_{u \in US_g} \beta_{ug} \cdot \frac{PF(u)}{N_{ds}(u)} \tag{23}$$

In this case, the perturbation of a given gene is due to the perturbations of the genes upstream, propagated through the pathway topology.

2.4 Impact analysis in the absence of perturbation propagation

In certain situations, one might not wish that the analysis propagate the gene perturbations through specific graph edges or types of graph edges (e.g., for edges corresponding to indirect effects or state changes). This can be easily achieved by setting $\beta = 0$ for the desired edges or edge types.

If no perturbation propagations are to be allowed at all, the expression of the gene perturbation in Eq.18 reduces to:

$$PF(g) = \Delta E \tag{24}$$

and the impact factor for the pathway becomes:

$$IF(P_i) = \log\left(\frac{1}{p_i}\right) + \frac{\sum_{g \in P_i} |PF(g)|}{\overline{|\Delta E|} \cdot N_{de}(P_i)} = \log\left(\frac{1}{p_i}\right) + \frac{1}{\overline{|\Delta E|}} \frac{\sum_{k=1}^{N_{de}(P_i)} |\Delta E(g)|}{N_{de}(P_i)} = \log\left(\frac{1}{p_i}\right) + \frac{\overline{|\Delta E_{P_i}|}}{\overline{|\Delta E|}} \tag{25}$$

In this case, the impact analysis would assess the pathways based not only on the number of differentially expressed genes that fall on each pathway but also based on the ratio between the average expression change on the pathway and the average expression change in the entire set of differentially expressed genes.

[a] If there are no differentially expressed genes anywhere, the problem of finding the impact on various pathways is meaningless.

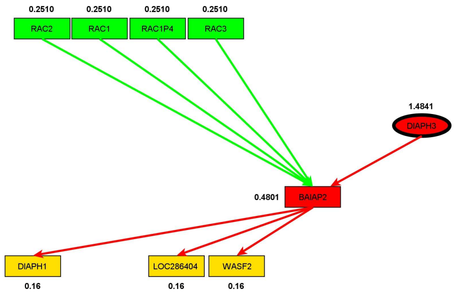


Figure 2: The computation of the PF for a gene and the subsequent propagation of the perturbation according to Eq. 3. The genes are part of regulation of actin cytoskeleton pathway shown in Fig. 6. Some of the interactions between the genes have been removed in order to simplify the figure. The labels next to each gene indicate the PF.

3 An example of perturbation factor propagation in a pathway

Fig. 2 illustrates the computation and propagation of the perturbations in a small area of a the actin cytoskeleton pathway (shown in its entirety in Fig. 6). As already mentioned, in all data shown here the regulatory efficiency is $\beta = 1$ for all genes. In this case, the gene *DIAPH3* is the input gene with an observed fold change $\Delta E = 1.4841$. Since there are no genes upstream of *DIAPH3*, its second term in Eq. 3 is zero. Using Eq. 3, the PF of gene *DIAPH3* is simply its measured expression change:

$$PF(DIAPH3) = 1.4841 + 0 = 1.4841 \tag{26}$$

The next step involves the computation of the perturbation for *BAIAP2*. This gene receives signals from *DIAPH3* but also from *RAC1*, *RAC1P4*, *RAC2* and *RAC3*. Using Eq. 3, the PF for the gene *BAIAP2* can be calculated as:

$$\begin{aligned} PF(BAIAP2) \quad = \quad & \Delta E(BAIAP2) + \frac{PF(DIAPH3)}{N_{ds}(DIAPH3)} + \frac{PF(RAC1)}{N_{ds}(RAC1)} + \frac{PF(RAC1P4)}{N_{ds}(RAC1P4)} \\ & + \frac{PF(RAC2)}{N_{ds}(RAC2)} + \frac{PF(RAC3)}{N_{ds}(RAC3)} \end{aligned}$$

The previously calculated perturbations for *RAC1*, *RAC1P4*, *RAC2* and *RAC3* are:

$$PF(RAC1) = PF(RAC1P4) = PF(RAC2) = PF(RAC3) = -0.251 \tag{27}$$

Each of these genes signals only to *BAIAP2* so for each of them the number of downstream genes, N_{ds} will be equal to 1. Hence, the PF for the gene *BAIAP2* can be calculated as:

$$PF(BAIAP2) = 1.4841 + \frac{-0.251}{1} + \frac{-0.251}{1} + \frac{-0.251}{1} + \frac{-0.251}{1} = 0.4801 \tag{28}$$

Similarly, using Eq. 3, the PF for the gene *DIAPH1* is

$$PF(DIAPH1) = \Delta E(DIAPH1) + \frac{PF(BAIAP2)}{N_{ds}(BAIAP2)} = 0 + \frac{0.4801}{3} = 0.16 \tag{29}$$

The perturbation of the other two genes, *LOC286404* and *WASF2* is analogous and yields the same numerical value.

3.1 Pathways involving loops

If the pathway includes loops, Eq. 18 becomes recursive, and the computation of the gene PFs will involve an iterative process. The best way of treating such loops would probably involve modeling the pathways as dynamical systems and using differential (or difference) equations to study them from the point of view of stability and convergence. However, at the moment, the expression data generated by the currently available techniques do not appear to be sufficiently accurate to allow this type of analysis. The very same biological samples analyzed on various platforms yield numbers that often correlate only around 0.7 (Draghici et al., 2006, Jenssen et al., 2002, Kuo et al., 2002, Tan et al., 2003). Treating the pathways as dynamical systems with such data runs quickly into stability problems. In order to address this in a feasible way, we perform the computation of the perturbation factors by going around each loop once. This approach appears to be a good compromise for the nature of the data: loops are not completely ignored and, at the same time, stability problems created by noisy data are avoided. The drawback is that the impact factor can only be interpreted in a probabilistic framework, and cannot be put into any type of quantitative correspondence with any biochemical product anywhere on the pathway.

4 Some simulations

We also performed a number of simulations in order to study the behavior of the proposed method. The research hypothesis here is that somewhere in the biological system studied, there is a subgroup of genes that are behaving in a way that is significantly different from the way they behave in the reference condition (fold changes different from zero with respect to this reference condition). The null hypothesis is that there is no such subgroup of genes, but rather that all gene expression changes are only due to random causes. In order to run simulations that would preserve as much as possible of the characteristics of the real data, we used one of the data sets already included in the paper, the lung adenocarcinoma data set. In this data set, we swapped randomly both class labels and the genes (both rows and columns in the gene expression matrix). The goal here is to see if any of the methods will report any significant pathways which in this case would be false positives. After swapping both genes and experiments randomly, we used a t-test to select the top N "differentially expressed" (DE) genes. We then used this set of DE genes with the hypergeometric and the pathway analysis to identify the "significantly impacted" pathways. Since GSEA does not require a set of differentially expressed genes, we used the entire (permuted) data set for its analysis. We repeated this for N=100, 200, 500, 1000, 2000, 3000 and 5000 DE genes out of the pool of 12,000 genes, for each N repeating the process 20 times on each of the 25 pathways.

Note that this type of simulations does not model very well the null hypothesis because it artificially divides the genes into DE genes, whose measured fold changes will be provided as input to the pathway analysis, and non-DE genes, whose measured fold changes will not be provided to the algorithm and hence will be ignored during the analysis. Because of this, certain pathways might find themselves with an arbitrary subgroup of genes with substantial fold changes while the remaining genes on the pathway will artificially be set to zero measured fold change. Even though this experimental setup does not model well the null hypothesis, this is the way the technique may be used in practice, and hence it is important to assess the effect of submitted a limited list of DE genes to the algorithm.

Varying the number of DE genes during these simulations, did not affect GSEA because GSEA does not use a subset of differentially expressed genes. Hence, GSEA yielded an approximately uniform distribution of raw p-values with a rate of raw false positives close to the expected 5%. The hypergeometric analysis also provided a similar distribution, approximately uniform and with a false positive rate close to 0.5%.

Fig. 3 shows the number of pathways that appeared to be significant in the pathway analysis, according to their raw p-values, as the number of DE genes increased from 100 to 5000 at a significance level of 5%. The figure shows that the practice of submitting a short list of DE genes to this type of analysis might generate a number of false positives higher than the expected 5%. The more genes are included in the analysis the better. Fig. 4 shows the distribution of the raw p-values for when submitting non-zero fold changes for 500 out of the 12,000 genes. The figure shows that this practice leads to a slightly non-uniform distribution with a tendency to generate more p-values lower than 0.5. This is to be expected because having a short list of genes with fold changes while all the other genes are set to 0 fold change might make certain pathways appear as locally perturbed. In this simulation, for 500 genes and $\alpha = 0.05$ the false positive rate was 6.8% slightly higher than the expected 5%. In all cases, performing a suitable correction for multiple comparisons

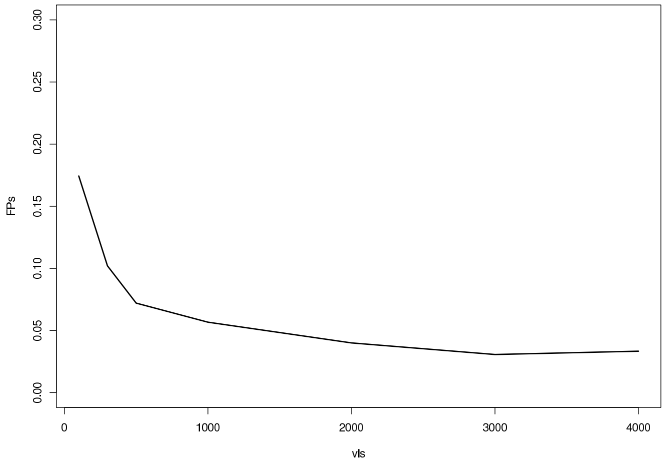


Figure 3: The fraction of false positives (using raw p-values) as a function of the number of genes with non-zero fold changes used as an input in the simulations performed. The more genes are included in the analysis the better. Because the pathway analysis considers both the measured fold changes, as well as the position of all genes on every pathway, the practice of providing the fold changes only for a rather short list of high confidence DE genes (and implicitely ignoring the measured fold changes of the others) can yield more false positives than expected by chance (5% in this case).

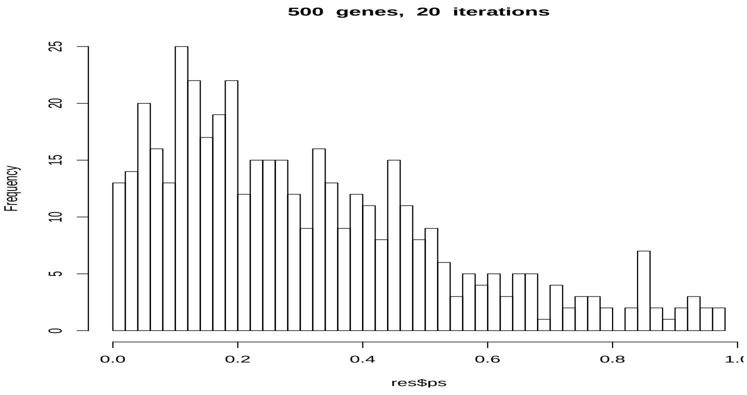
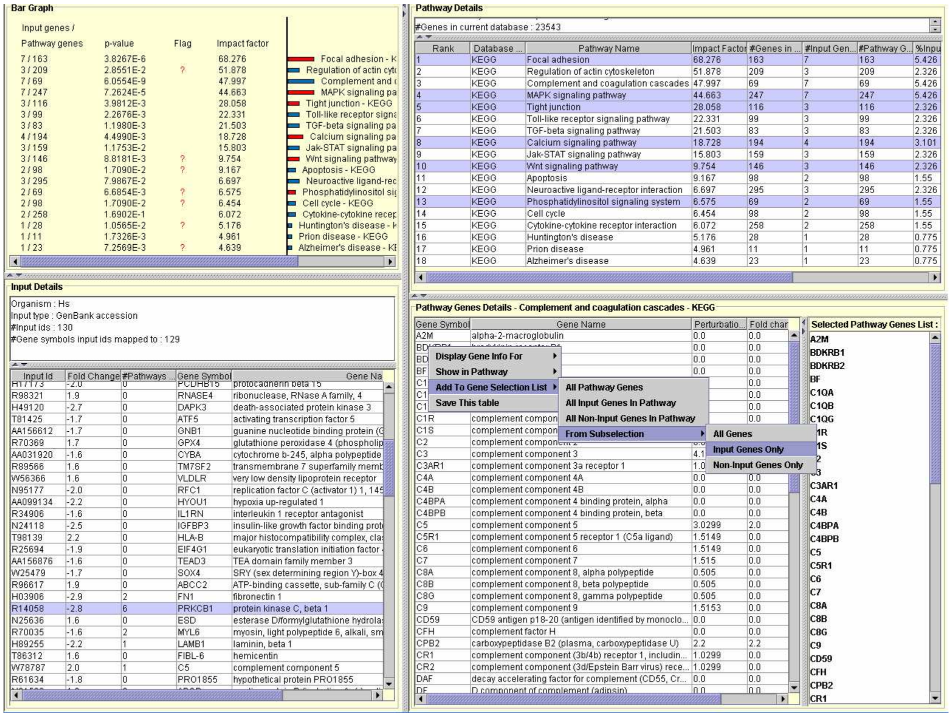


Figure 4: Distribution of the raw p-values provided by the pathway analysis when submitting non-zero fold changes for 500 out of 12,000 genes in 20 random trials. In this simulation, for 500 genes and $\alpha = 0.05$, the false positive rate was 6.8%, slightly higher than the expected 5%. After the correction for multiple comparisons, no false positives are present.



Input genes / Pathway genes	p-value	Flag	Impact factor	
7 / 163	3.8267E-6		68.276	Focal adhesion - K
3 / 209	2.8551E-2	?	51.878	Regulation of actin cyt
7 / 69	6.0554E-9		47.997	Complement and
7 / 247	7.2624E-5		44.663	MAPK signaling pa
3 / 116	3.9812E-3		28.058	Tight junction - KEGG
3 / 99	2.2676E-3		22.331	Toll-like receptor signa
3 / 83	1.1980E-3		21.503	TGF-beta signaling pa
4 / 194	4.4990E-3		18.728	Calcium signaling pa
3 / 159	1.1753E-2		15.803	Jak-STAT signaling pa
3 / 146	8.8181E-3	?	9.754	Wnt signaling pathway
2 / 98	1.7090E-2	?	9.167	Apoptosis - KEGG
3 / 295	7.9867E-2		6.697	Neuroactive ligand-rec
2 / 69	6.6854E-3	?	6.575	Phosphatidylinositol si
2 / 98	1.7090E-2	?	6.454	Cell cycle - KEGG
2 / 258	1.6902E-1		6.072	Cytokine-cytokine recep
1 / 28	1.0565E-2	?	5.176	Huntington's disease - K
1 / 11	1.7326E-3		4.961	Prion disease - KEGG
1 / 23	7.2569E-3	?	4.639	Alzheimer's disease - KE

Pathway Details

#Genes in current database : 23543

Rank	Database ...	Pathway Name	Impact Factor	#Genes in ...	#Input Gen...	#Pathway G...	%Inpu
1	KEGG	Focal adhesion	68.276	163	7	163	5.426
2	KEGG	Regulation of actin cytoskeleton	51.878	209	3	209	2.326
3	KEGG	Complement and coagulation cascades	47.997	69	7	69	5.426
4	KEGG	MAPK signaling pathway	44.663	247	7	247	5.426
5	KEGG	Tight junction	28.058	116	3	116	2.326
6	KEGG	Toll-like receptor signaling pathway	22.331	99	3	99	2.326
7	KEGG	TGF-beta signaling pathway	21.503	83	3	83	2.326
8	KEGG	Calcium signaling pathway	18.728	194	4	194	3.101
9	KEGG	Jak-STAT signaling pathway	15.803	159	3	159	2.326
10	KEGG	Wnt signaling pathway	9.754	146	3	146	2.326
11	KEGG	Apoptosis	9.167	98	2	98	1.55
12	KEGG	Neuroactive ligand-receptor interaction	6.697	295	3	295	2.326
13	KEGG	Phosphatidylinositol signaling system	6.575	69	2	69	1.55
14	KEGG	Cell cycle	6.454	98	2	98	1.55
15	KEGG	Cytokine-cytokine receptor interaction	6.072	258	2	258	1.55
16	KEGG	Huntington's disease	5.176	28	1	28	0.775
17	KEGG	Prion disease	4.961	11	1	11	0.775
18	KEGG	Alzheimer's disease	4.639	23	1	23	0.775

Input Details

Organism : Hs
Input type : GenBank accession
#Input ids : 130
#Gene symbols input ids mapped to : 129

Input Id	Fold Change	#Pathways ...	Gene Symbol	Gene Na
H17173	-2.0	0	PCDHB15	protocadherin beta 15
R98321	1.9	0	RNASE4	ribonuclease, RNase A family, 4
H49120	-2.7	0	DAPK3	death-associated protein kinase 3
T81425	-1.7	0	ATF5	activating transcription factor 5
AA156612	-1.7	0	GNB1	guanine nucleotide binding protein (G
R70369	1.7	0	GPX4	glutathione peroxidase 4 (phospholip
AA031920	-1.6	0	CYBA	cytochrome b-245, alpha polypeptide
R89566	1.6	0	TM7SF2	transmembrane 7 superfamily memb
W56366	1.6	0	VLDLR	very low density lipoprotein receptor
N95177	-2.0	0	RFC1	replication factor C (activator 1) 1, 145
AA099134	-2.2	0	HYOU1	hypoxia up-regulated 1
R34906	-1.6	0	IL1RN	interleukin 1 receptor antagonist
N24118	-2.5	0	IGFBP3	insulin-like growth factor binding prot
T98139	2.2	0	HLA-B	major histocompatibility complex, clas
R25694	-1.9	0	EIF4G1	eukaryotic translation initiation factor
AA156876	-1.6	0	TEAD3	TEA domain family member 3
W25479	-1.7	0	SOX4	SRY (sex determining region Y)-box 4
R96617	1.9	0	ABCC2	ATP-binding cassette, sub-family C (C
H03906	-2.9	2	FN1	fibronectin 1
R14058	-2.8	6	PRKCB1	protein kinase C, beta 1
N25636	1.6	0	ESD	esterase D/formylglutathione hydrolas
R70035	-1.6	2	MYL6	myosin, light polypeptide 6, alkali, sm
H89255	-2.2	1	LAMB1	laminin, beta 1
T86312	1.6	0	FIBL-6	hemicentin
W78787	2.0	1	C5	complement component 5
R61634	-1.8	0	PRO1855	hypothetical protein PRO1855

Pathway Genes Details - Complement and coagulation cascades - KEGG

Gene Symbol	Gene Name	Perturbatio...	Fold char
A2M	alpha-2-macroglobulin	0.0	0.0
BD		0.0	0.0
BD		0.0	0.0
BF		0.0	0.0
C1			0.0
C1			0.0
C1			0.0
C1R	complement compon		0.0
C1S	complement compon		
C2	complement compon		
C3	complement component 3	4.1	
C3AR1	complement component 3a receptor 1	1.0	
C4A	complement component 4A	0.0	0.0
C4B	complement component 4B	0.0	0.0
C4BPA	complement component 4 binding protein, alpha	0.0	0.0
C4BPB	complement component 4 binding protein, beta	0.0	0.0
C5	complement component 5	3.0299	2.0
C5R1	complement component 5 receptor 1 (C5a ligand)	1.5149	0.0
C6	complement component 6	1.5149	0.0
C7	complement component 7	1.515	0.0
C8A	complement component 8, alpha polypeptide	0.505	0.0
C8B	complement component 8, beta polypeptide	0.505	0.0
C8G	complement component 8, gamma polypeptide	0.505	0.0
C9	complement component 9	1.5153	0.0
CD59	CD59 antigen p18-20 (antigen identified by monoclo...	0.0	0.0
CFH	complement factor H	0.0	0.0
CPB2	carboxypeptidase B2 (plasma, carboxypeptidase U)	2.2	2.2
CR1	complement component (3b/4b) receptor 1, includin...	1.0299	0.0
CR2	complement component (3d/Epstein Barr virus) rece...	1.0299	0.0
DAF	decay accelerating factor for complement (CD55, Cr...	0.0	0.0
DF	D component of complement (adipsin)	0.0	0.0

Selected Pathway Genes List :

A2M
BDKRB1
BDKRB2
BF
C1QA
C1QB
C1QG
1R
1S
2
3
C3AR1
C4A
C4B
C4BPA
C4BPB
C5
C5R1
C6
C7
C8A
C8B
C8G
C9
CD59
CFH
CPB2
CR1

Figure 5: The main output window of Pathway-Express.

eliminated all false positives.

5 Implementation details

This analysis method has been implemented as a web-based tool, Pathway-Express, freely accessible as part of Onto-Tools. All data sets discussed here are available at the click of a button, as demo runs. The typical output window is shown in Fig. 5. The tool uses pathway data from KEGG and implements both the classical statistical approach (ORA), as well as the impact analysis described above, allowing a side by side comparison. The tool also allows rapid queries for genes or pathways, visualization of entire pathways (see Fig. 1), etc.

Currently, all signaling pathways for human, mouse and rat are downloaded from KEGG and stored in a relational database. In order to calculate the impact factor for a given pathway, the pathway database is queried to retrieve all genes and gene interactions in the pathway, and a graph data structure for this pathway is created. The genes are represented as nodes, and the gene interactions as edges of the graph (see Fig. 6). The user-provided normalized fold changes are mapped on the pathway graph and used to calculate the gene perturbation factors as described in Eq. 18. Once the perturbation factors of all genes in a given pathway are calculated, Eq. 14 is used to calculate the impact factor of each pathway. The impact factor of each pathway is then used as a score to assess the impact of a given gene expression data set on all pathways (the higher the impact factor the more significant the pathway).

6 Data sets availability

The data sets used in this paper are available for download from the following web addresses:

1. Breat cancer data set - van't Veer *et al.* Nature, 415(6871):530-536, January 2002. `http://www.rii.com/publications/2002/vantveer.html`

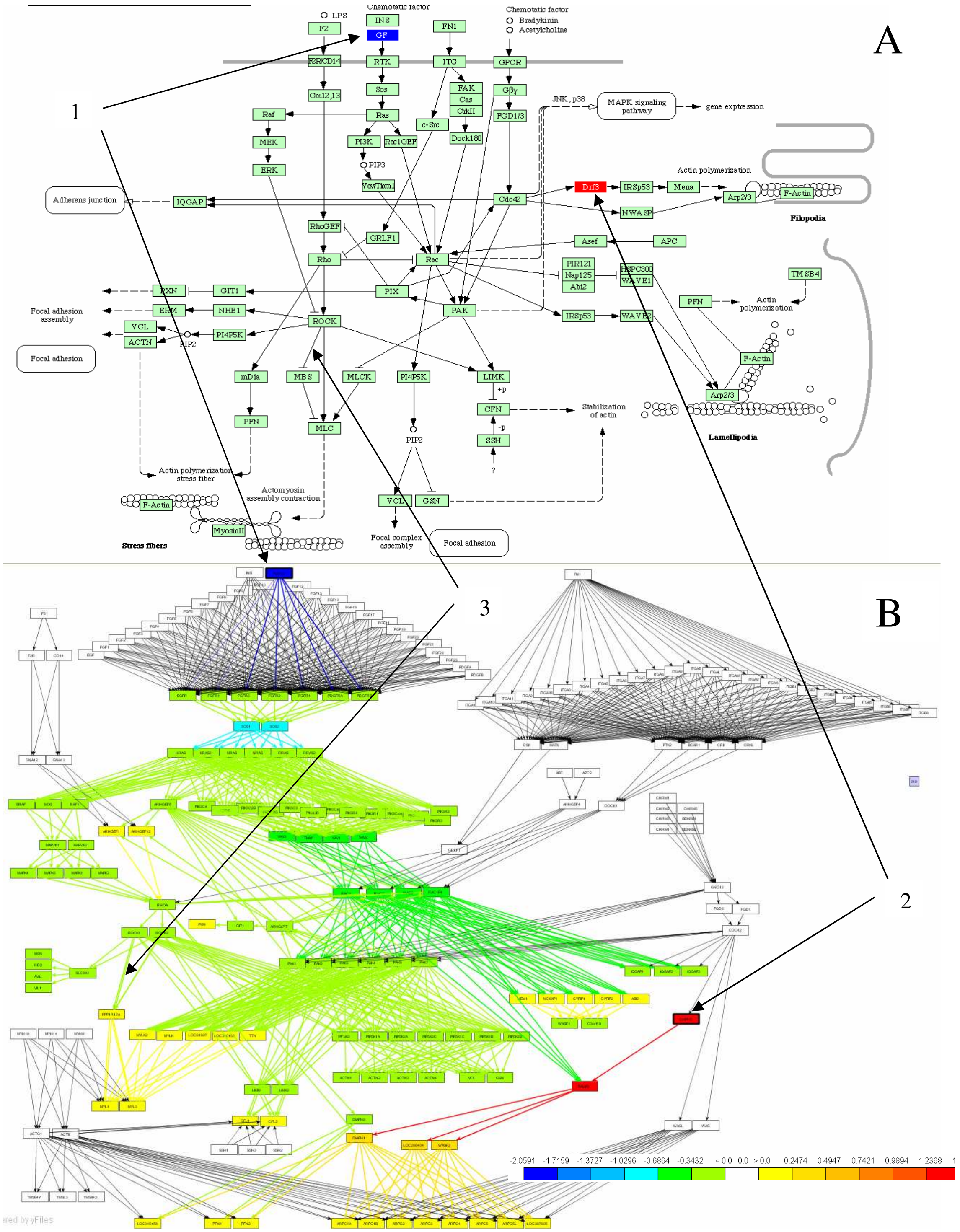


Figure 6: Regulation of actin cytoskeleton as impacted in breast cancer (van't Veer et al., 2002): the KEGG pathway diagram (A) and its internal graph representation (B). Note that the unique symbol GF (blue) in the KEGG diagram A, actually stands for 25 FGF genes in the internal graph B, only one of which is differentially expressed (1). The colors show the propagation of the gene perturbations throughout the pathway. The differentially expressed genes are FGF18 (1) and DIAPH3 (2). Changes from blue/green to yellow/red and viceversa correspond to inhibitory interactions. For instance, since ROCK inhibits MBS, the negative perturbation of ROCK propagates as a positive perturbation of MBS (3).

2. Lung cancer data set - Beer *et al.* Nature Medicine, 8(8):816-824, Jul 2002. `http://dot.ped.med.umich.edu:2000/ourimage/pub/Lung/gene_expression_data_for_all_samples.txt`

3. Hepatic cell line data set - Swagell *et al.* Biochemical and Biophysical Research Communications, 328: 432-441, 2005. The list of differentially expressed genes and their corresponding fold changes were obtained from table 2 in the manuscript.

The input files used in Pathway-Express for each of these data sets are also available at `http://vortex.cs.wayne.edu/ontoexpress/servlets/pedemoruninfo.html`.

References

Draghici, S., Khatri, P., Eklund, A. C., and Szallasi, Z., 2006. Reliability and reproducibility issues in DNA microarray measurements. *Trends Genet*, **22**(2):101–9.

Hogg, R. V., 1978. *Introduction to Mathematical Statistics.* New York: Macmillan.

Jenssen, T.-K., Langaas, M., Kuo, W. P., Smith-Sorensen, B., Myklebost, O., and Hovig, E., 2002. Analysis of repeatability in spotted cDNA microarrays. *Nucleic Acids Research*, **30**(14):3235–3244.

Kuo, W. P., Jenssen, T.-K., Butte, A. J., Ohno-Machado, L., and Kohane, I. S., 2002. Analysis of matched mRNA measurements from two different microarray technologies. *Bioinformatics*, **18**(3):405–412.

Tan, P. K., Downey, T. J., Jr., E. L. S., Xu, P., Fu, D., Dimitrov, D. S., Lempicki, R. A., Raaka, B. M., and Cam, M. C., 2003. Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Research*, **31**(19):5676–5684.

van't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveenothers, A. T., *et al.*, 2002. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**(6871):530–536.