A systems biology approach for pathway level analysis

Sorin Draghici,1,2,4 Purvesh Khatri,2 Adi Laurentiu Tarca,1,2,3 Kashyap Amin,2 Arina Done,2 Calin Voichita,2 Constantin Georgescu,2 and Roberto Romero3

1,2,3Karmanos Cancer Institute, Wayne State University, Detroit, Michigan 48202, USA; 2Department of Computer Science, Wayne State University, Detroit, Michigan 48202, USA; 3Perinatology Research Branch, NIH/NICHD, Detroit, Michigan 48201, USA

A common challenge in the analysis of genomics data is trying to understand the underlying phenomenon in the context of all complex interactions taking place on various signaling pathways. A statistical approach using various models is universally used to identify the most relevant pathways in a given experiment. Here, we show that the existing pathway analysis methods fail to take into consideration important biological aspects and may provide incorrect results in certain situations. By using a systems biology approach, we developed an impact analysis that includes the classical statistics but also considers other crucial factors such as the magnitude of each gene's expression change, their type and position in the given pathways, their interactions, etc. The impact analysis is an attempt to a deeper level of statistical analysis, informed by more pathway-specific biology than the existing techniques. On several illustrative data sets, the classical analysis produces both false positives and false negatives, while the impact analysis provides biologically meaningful results. This analysis method has been implemented as a Web-based tool, Pathway-Express, freely available as part of the Onto-Tools (http://vortex.cs.wayne.edu).

*Corresponding author.
E-mail sod@cs.wayne.edu; fax (313) 577-0868.

Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.6202607.

Together with the ability of generating a large amount of data per experiment, high-throughput technologies also brought the challenge of translating such data into a better understanding of the underlying biological phenomena. Independent of the platform and the analysis methods used, the result of a high-throughput experiment is, in many cases, a list of differentially expressed genes. The common challenge faced by all researchers is to translate such lists of differentially expressed genes into a better understanding of the underlying biological phenomena and, in particular, to put this in the context of the whole organism as a complex system. In 2002, a computerized analysis approach using the Gene Ontology (GO) was proposed to deal with this issue (Khatri et al. 2002; Draghici et al. 2003). This approach takes a list of differentially expressed genes and uses a statistical analysis to identify the GO categories (e.g. biological processes, etc.) that are over- or under-represented in the condition under study. Given a set of differentially expressed genes, this approach compares the number of differentially expressed genes found in each category of interest with the number of genes expected to be found in the given category just by chance. If the observed number is substantially different from the one expected just by chance, the category is reported as significant. A statistical model (e.g. hypergeometric) can be used to calculate the probability of observing the actual number of genes just by chance, i.e., a P-value. Currently, there are over 20 tools using this over-representation approach (ORA) (Khatri and Draghici 2005). In spite of its wide adoption, this approach has a number of limitations related to the type, quality, and structure of the annotations available. An alternative approach considers the distribution of the pathway genes in the entire list of genes and performs a functional class scoring (FCS), which also allows adjustments for gene correlations (Goeman et al. 2004; Pavlidis et al. 2004). Arguably the state of the art in the FCS category, the Gene Set Enrichment Analysis (GSEA) (Mootha et al. 2003; Subramanian et al. 2005; Tian et al. 2005), ranks all genes based on the correlation between their expression and the given phenotypes, and calculates a score that reflects the degree to which a given pathway P is represented at the extremes of the entire ranked list. The score is calculated by walking down the list of genes ordered by expression change. The score is increased for every gene that belongs to P and decreased for every gene that does not. Statistical significance is established with respect to a null distribution constructed by permutations.

Both ORA and FCS techniques currently used are limited by the fact that each functional category is analyzed independently without a unifying analysis at a pathway or system level (Tian et al. 2005). This approach is not well suited for a systems biology approach that aims to account for system level dependencies and interactions as well as to identify perturbations and modifications at the pathway or organism level (Stelling 2004). Several pathway databases such as KEGG (Ogata et al. 1999), BioCarta (http://www.biocarta.com), and Reactome (Joshi-Tope et al. 2005) currently describe metabolic pathway and gene signaling networks offering the potential for a more complex and useful analysis. A recent technique, ScorePage, has been developed in an attempt to take advantage of these types of data for the analysis of metabolic pathways (Rahnenfuhrer et al. 2004). Unfortunately, no such technique currently exists for the analysis of gene signaling networks. All pathway analysis tools currently available use one of the ORA approaches above and fail to take advantage of the much richer data contained in these resources. GenMAPP/MAPPfinder (Doniger et al. 2003; Dahlquist et al. 2002) and Gene-Sifter use a standardized Z-score. PathwayProcessor (Grosu et al. 2002), PathMAPA (Pan et al. 2003), Cytoscape (Shannon et al. 2003), and PathwayMiner (Pandey et al. 2004) use Fisher's
exact test. MetaCore uses a hypergeometric model, while ArrayX-Path (Chung et al. 2004) offers both Fisher’s exact test and a false discovery rate (FDR). Finally, VitaPad (Holford et al. 2004) and Pathway Studio (Nikitin et al. 2003) focus on visualization alone and do not offer any analysis.

The approaches currently available for the analysis of gene signaling networks share a number of important limitations. First, these approaches consider only the set of genes on any given pathway and ignore their position in those pathways. This may be unsatisfactory from a biological point of view. If a pathway is triggered by a single gene product or activated through a single receptor and if that particular protein is not produced, the pathway will be greatly impacted, probably completely shut off. A good example is the insulin pathway (http://www.genome.ac.jp/KEGG/pathway/hsa04910.html). If the insulin receptor (INSR) is not present, the entire pathway is shut off. Conversely, if several genes are involved in a pathway but they only appear somewhere downstream, changes in their expression levels may not affect the given pathway as much.

Second, some genes have multiple functions and are involved in several pathways but with different roles. For instance, the above INSR is also involved in the adherens junction pathway as one of the many receptor protein tyrosine kinases. However, if the expression of INSR changes, this pathway is not likely to be heavily perturbed because INSR is just one of many receptors on this pathway. Once again, all these aspects are not considered by any of the existing approaches.

Probably the most important challenge today is that the knowledge embedded in these pathways about how various genes interact with each other is not currently exploited. The very purpose of these pathway diagrams is to capture some of our knowledge about how genes interact and regulate each other. However, the existing analysis approaches consider only the sets of genes involved on these pathways, without taking into consideration their topology. In fact, our understanding of various pathways is expected to improve as more data are gathered. Pathways will be modified by adding, removing or redirecting links on the pathway diagrams. Most existing techniques are completely unable to even sense such changes. Thus, these techniques will provide identical results as long as the pathway diagram involves the same genes, even if the interactions between them are completely redefined over time.

Finally, up to now the expression changes measured in these high-throughput experiments have been used only to identify differentially expressed genes (ORA approaches) or to rank the genes (FCS methods), but not to estimate the impact of such changes on specific pathways. Thus, ORA techniques will see no difference between a situation in which a subset of genes is differentially expressed just above the detection threshold (e.g., twofold) and the situation in which the same genes are changing by many orders of magnitude (e.g., 100-fold). Similarly, FCS techniques can provide the same rankings for entire ranges of expression values, if the correlations between the genes and the phenotypes remain similar. Even though analyzing this type of information in a pathway and system context would be extremely meaningful from a biological perspective, currently there is no technique or tool able to do this.

We propose a radically different approach for pathway analysis that attempts to capture all aspects above. An impact factor (IF) is calculated for each pathway incorporating parameters such as the normalized fold change of the differentially expressed genes, the statistical significance of the set of pathway genes, and the topology of the signaling pathway. We show on a number of real data sets that the intrinsic limitations of the classical analysis produce both false positives and false negatives while the impact analysis provides biologically meaningful results.

### Impact analysis

Our goal is to develop an analysis model that would require both a statistically significant number of differentially expressed genes and biologically meaningful changes on a given pathway. In this model, the IF of a pathway $P_i$ is calculated as the sum of two terms:

$$IF(P_i) = \log \left( \frac{1}{P_i} \right) + \sum_{g \in P_i} \frac{|PF(g)|}{|\Delta g| \cdot N_{g1}(P_i)}.$$  

The first term is a probabilistic term that captures the significance of the given pathway $P_i$ from the perspective of the set of genes contained in it. This term captures the information provided by the currently used classical statistical approaches and can be calculated using either an ORA (e.g., z-test [Doniger et al. 2003]), contingency tables [Pan et al. 2003; Pandey et al. 2004], etc.), a FCS approach (e.g., GSEA; Mootha et al. 2003; Subramanian et al. 2005) or other more recent approaches (Robinson et al. 2004; Breslin et al. 2005; Tian et al. 2005). The $P_i$ value corresponds to the probability of obtaining a value of the statistic used at least as extreme as the one observed, when the null hypothesis is true. The results presented here were obtained using the hypergeometric model (Tavazoie et al. 1999; Draghici et al. 2003) in which $P_i$ is the probability of obtaining at least the observed number of differentially expressed gene, $N_{g1}$, just by chance.

The second term in Equation 1 is a functional term that depends on the identity of the specific genes that are differentially expressed as well as on the interactions described by the pathway (i.e., its topology). In essence, this term sums up the absolute values of the perturbation factors (PFs) for all genes $g$ on the given pathway $P_i$. The PF of a gene $g$ is calculated as follows:

$$PF(g) = \Delta E(g) + \sum_{u \in g} \beta_{ug} \frac{PF(u)}{N_{u1}(u)}.$$  

In this equation, the first term captures the quantitative information measured in the gene expression experiment. The factor $\Delta E(g)$ represents the signed normalized measured expression change of the gene $g$ determined using one of the available methods (Quackenbush 2001; Churchill 2002; Draghici 2002; Yang and Speed 2002). The second term is a sum of all PFs of the genes $u$ directly upstream of the target gene $g$, normalized by the number of downstream genes of each such gene $N_{g1}(u)$, and weighted by a factor $\beta_{ug}$, which reflects the type of interaction: $\beta_{ug} = 1$ for induction, $\beta_{ug} = -1$ for repression. (In KEGG, which is the source of the pathways used here, this information about the type of interaction is available for every link between two genes in the description of the pathway topology.) $U_{g1}$ is the set of all such genes upstream of $g$. The second term here is similar to the Page-Rank index used by Google (Page et al. 1998), only we weight the downstream instead of the upstream connections (a Web page is important if other pages point to it, whereas a gene is important if it influences other genes).

Under the null hypothesis, which assumes that the list of differentially expressed genes only contains random genes, the likelihood that a pathway has a large IF is proportional to the
number of such “differentially expressed” genes that fall on the pathway, which in turn is proportional to the size of the pathway. Thus, we need to normalize with respect to the size of the pathway by dividing the total perturbation by the number of differentially expressed genes on the given pathway, \( N_{de}(P_i) \). Furthermore, various technologies can yield systematically different estimates of the fold changes. For instance, the fold changes reported by microarrays tend to be compressed with respect to those reported by RT-PCR (Canales et al. 2006; Draghici et al. 2006). In order to make the \( IFs \) as independent as possible from the technology, and also comparable between problems, we also divide the second term in Equation 1 by the mean absolute fold change \( \Delta E \), calculated across all differentially expressed genes. Assuming that there are at least some differentially expressed genes anywhere in the data set, both \( \Delta E \) and \( N_{de}(P_i) \) are different from zero so the second term is properly defined. (If there are no differentially expressed genes anywhere, the problem of finding the impact on various pathways is meaningless.)

It can be shown that the \( IFs \) correspond to 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. \( IF \) values, \( if \), will follow a \( \Gamma(2,1) \) distrib-
tion from which $P$-values can be calculated as $P = \left( \frac{f}{f + 1} \right) \cdot e^{-\frac{1}{f}}$ (for details, see Supplemental materials).

The impact analysis proposed here extends and enhances the existing statistical approaches by incorporating the novel aspects discussed above. For instance, the second term of the gene perturbation (in Equation 2) increases the PF scores of those genes that are connected through a direct signaling link to other differentially expressed genes (e.g., the PFs of F5 and F11 in Fig. 1 are both increased because of the differentially expressed SERPINC1 and SERPINA1). This will yield a higher overall score for those pathways in which the differentially expressed genes are localized in a connected subgraph, as in this example. Interestingly, when the limitations of the existing approaches 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 statistics and yields the same results. For instance, if there are no perturbations directly upstream of a given gene, the second term in Equation 2 is zero and the PF reduces to the measured expression change $E$, which is the classical way of assessing the impact of a condition upon a given gene. A more detailed discussion of various particular cases is included in the Supplemental materials.

### Results

We have used this pathway analysis approach to analyze several data sets. A first such set includes genes associated with better survival in lung adenocarcinoma (Beer et al. 2002). These genes have the potential to represent an important tool for the therapeutic decision, and if the correct regulatory mechanisms are identified, they could also be potential drug targets. The expression values of the 97 genes associated with better survival identified by Beer and colleagues were compared between the cancer and healthy groups. These data were then analyzed using a classical ORA approach (hypergeometric model), a classical FCS approach (GSEA), and our impact analysis. Figure 2 shows a comparison between the results obtained with the three approaches.

From a statistical perspective, the power of both classical techniques appears to be very limited. The corrected $P$-values do not yield any pathways at the usual 0.01 or 0.05 significance levels, independently of the type of correction. If the significance levels were to be ignored and the techniques used only to rank the pathways, the results would continue to be unsatisfactory. According to the classical ORA analysis, the most significantly affected pathways in this data set are *prion disease*, *focal adhesion*, and *Parkinson’s disease*. In reality, both prion and Parkinson’s diseases are pathways specifically associated to diseases of the central nervous system and are unlikely to be related to lung adenocarcinomas. In this particular case, prion disease ranks at the top only due to the differential expression of LAMB1. Since this pathway is rather small (14 genes), every time any one gene is differentially expressed, the hypergeometric analysis will rank it highly. A similar phenomenon happens with Parkinson’s disease, indicating that this is a problem associated with the method rather than with a specific pathway. At the same time, pathways highly relevant to cancer such as *cell cycle* and *Wnt signaling* are ranked in the lower half of the pathway list. The most significant pathways reported as enriched in cancer by GSEA (Subramanian et al. 2005) are *cell cycle*, *Huntington’s disease*, *DRPLA*, *Alzheimer’s disease*, and Parkinson’s disease (see Fig. 2). Among these, only cell cycle is relevant, while Huntington’s, Alzheimer’s and Par-

![Figure 2](image-url)
von Essen's diseases are clearly incorrect. However, although ranked first, cell cycle is not significant in GSEA, even at the most lenient 10% significance and with the least conservative correction.

In contrast, the impact analysis reports cell cycle as the most perturbed pathway in this condition and also as highly significant from a statistical perspective ($P = 1.6 \times 10^{-6}$). Since early articles on the molecular mechanisms perturbed in lung cancers (Slebos and Rodenhuis 1989; Nau et al. 1985) until the most recent articles on this topic (Panani and Roussos 2006; Coe et al. 2006), there is a consensus that the cell cycle is highly deranged in lung cancers. Moreover, cell cycle genes have started to be considered both as potential prognostic factors and therapeutic targets (Vincenzi et al. 2006). The second most significant pathway as reported by the impact analysis is focal adhesion. An inspection of this pathway (shown in Fig. 3) shows that in these data, both ITG and RTK receptors are perturbed, as well as the VEGF ligand. Because these three genes appear at the very beginning and affect both entry points controlling this pathway, their perturbations are widely propagated throughout the pathway. Furthermore, the CRK oncogene was also found to be up-regulated. Increased levels of CRK proteins have been observed in several human cancers, and over-expression of CRK in epithelial cell cultures promotes enhanced cell dispersal and invasion (Rodrigues et al. 2005). For this pathway, the impact analysis yields a raw $P$-value of 0.005, which remains significant even after the FDR correction ($P = 0.048$), at the 5% level. In contrast, the ORA analysis using the hypergeometric model yields a raw $P$-value of 0.155 (FDR corrected to 0.627), while the GSEA analysis yields a raw $P$-value of 0.16 (FDR corrected to 0.384). For both techniques, not even the raw $P$-values are significant at the usual levels of 5% or 10%. This is not a mere accident but an illustration of the intrinsic limitations of the classical approaches. These approaches completely ignore the position of the genes on the given pathways, and therefore, they are not able to identify this pathway as being highly impacted in this condition. Note that any ORA approach will yield the same results for this pathway for any set of four differentially expressed genes from the set of genes on this pathway. Similarly, GSEA will yield the same results for any other set of four genes with similar expression values (yielding similar correlations with the phenotype). Both techniques are unable to distinguish between a situation in which these genes are upstream, potentially commandeering the entire pathway as in this example, or randomly distributed throughout the pathway.

The third pathway as ranked by the impact analysis is Wnt signaling (FDR corrected $P = 0.055$, significant at 10%). The importance of this pathway is well supported by independent research. At least three mechanisms for the activation of Wnt signaling pathway in lung cancers have been recently identified: (1) over-expression of Wnt effectors such as Dvl, (2) activation of a non-canonical pathway involving MAPK (previously known as JNK), and (3) repression of Wnt antagonists such as WIF (Mazieres et al. 2005). Mazieres and colleagues also argue that the blockade of Wnt pathway may lead to new treatment strategies in lung cancer.

In the same data set, Huntington’s disease, Parkinson’s disease, prion disease, and Alzheimer’s disease have low $P$s (corrected $P$-values of >0.20), correctly indicating that they are unlikely to be relevant in lung adenocarcinomas.

A second data set includes genes identified as being associated with poor prognosis in breast cancer (van’t Veer et al. 2002).
Figure 4 shows a comparison among the classical hypergeometric approach, GSEA, and the pathway impact analysis. Based on these data, GSEA finds no significantly impacted pathways at any of the usual 5% or 10% levels. In fact, the only FDR-corrected value below 0.25, in the entire data set is 0.11, corresponding to the ubiquitin mediated proteolysis pathway. Furthermore, GSEA’s ranking does not appear to be useful for these data, with none of the cancer-related pathways being ranked toward the top. The most significant signaling pathway according to the hypergeometric analysis, cell cycle, is also the most significant in the impact analysis. However, the agreement between the two approaches stops here. In terms of statistical power, according to the classical hypergeometric model, there are no other significant pathways at either 5% or 10% significance on the corrected P-values. If we were to ignore the usual significance thresholds and only consider the ranking, the third highest pathway according to the hypergeometric model is Parkinson’s disease. In fact, based on current knowledge, Parkinson’s disease is unlikely to be related to rapid metastasis in breast cancer. At the same time, the impact analysis finds several other pathways as significant. For instance, focal adhesion is significant with an FDR-corrected P-value of 0.03. In fact, a link between focal adhesion and breast cancer has been previously established (Golubovskaya et al. 2002; van Nimwegen and van de Water 2006). In particular, PTK2 (previously known as FAK), a central gene on the focal adhesion pathway, has been found to contribute to cellular adhesion and survival pathways in breast cancer cells, which are not required for survival in nonmalignant breast epithelial cell (Beviglia et al. 2003). Recently, it has also been shown that Doxorubicin, an anti-cancer drug, caused the formation of well-defined focal adhesions and stress fibers in mammary adenocarcinoma MTLn3 cells early after treatment (van Nimwegen et al. 2006). Consequently, the PTK2/PI-3 kinase/PKB signaling route within the focal adhesion pathway has been recently proposed as the mechanism through which Doxorubicin triggers the onset of apoptosis (van Nimwegen et al. 2006).

TGF-beta signaling (P = 0.032) and MAPK (P = 0.064) are also significant. Both fit well with previous research results. TGF-beta1, the main ligand for the TGF-beta signaling pathway, is known as a marker of invasiveness and metastatic capacity of breast cancer cells (Todorovic-Rakovic 2005). Furthermore, plasma levels of TGF-beta1 have been used to identify low-risk postmenopausal metastatic breast cancer patients (Nikolic-Vukosavljevic et al. 2004). Finally, MAPK has been shown to be connected not only to cancer in general but to this particular type of cancer. For instance the proliferative response to progesterin and estrogen was shown to interplay between estrogen receptors and EBRB2 (previously known as HER-2; human epidermal growth factor receptor 2) (Todorovic-Rakovic 2005). Furthermore, plasma levels of TGF-beta1 have been used to identify low-risk postmenopausal metastatic breast cancer patients (Nikolic-Vukosavljevic et al. 2004). Finally, MAPK has been shown to be connected not only to cancer in general but to this particular type of cancer. For instance the proliferative response to progesterin and estrogen was shown...
to be inhibited in mammary cells microinjected with inhibitors of MAP kinase pathway (Chen et al. 2001). Also, it is worth noting the gap between the P-values for regulation of actin cytoskeleton ($P = 0.111$), which may be relevant in cancer, and the next pathway, Parkinson’s disease ($P = 0.239$), which is irrelevant in this condition.

A third data set involves a set of differentially expressed genes obtained by studying the response of a hepatic cell line when treated with palmitate (Swagell et al. 2005). Figure 5 shows the comparison between the classical statistical analysis (ORA) and the pathway impact analysis. (The GSEA analysis requires expression values for all genes; since this experiment was performed with a custom array and not all values are publicly available, GSEA could not be applied here.) The classical statistical analysis yields three pathways significant at the 5% level: complement and coagulation cascade, focal adhesion, and MAPK. The impact analysis agrees on all these but also identifies several additional pathways. The top four pathways identified by the impact analysis are well supported by the existing literature. There are several studies that support the existence of a relationship between different coagulation factors, present in the complement and coagulation cascades pathway, and palmitate. Sanders et al. (1999), for instance, demonstrated that a high palmitate intake affects factor VII coagulant ($\text{FVIIc}$) activity. Interestingly, Figure 1 shows not only that this pathway has a higher than expected proportion of differentially expressed genes but also that six out of seven such genes are involved in the same region of the pathway, suggesting a coherently propagated perturbation. The focal adhesion and tight junction pathways involve cytoskeletal genes. Swagell et al. (2005) considered the presence of the cytoskeletal genes among the differentially expressed genes as very interesting and hypothesized that the down-regulation of these cytoskeletal genes indicates that palmitate decreases cell growth. Finally, the link between MAPK and the palmitate was established by Susztak et al. (2005), who showed that $p38$ MAP kinase is a key player in the palmitate-induced apoptosis.

**Conclusions**

A statistical approach using various models is commonly used in order to identify the most relevant pathways in a given experiment. This approach is based on the set of genes involved in each pathway. We identified a number of additional factors that may be important in the description and analysis of a given biological pathway. Based on these, we developed a novel impact analysis method that uses a systems biology approach in order to identify pathways that are significantly impacted in any condition monitored through a high-throughput gene expression technique. The impact analysis incorporates the classical probabilistic component but also includes important biological factors that are not captured by the existing techniques: the magnitude of the expression changes of each gene, the position of the differentially expressed genes on the given pathways, the topology of the pathway that describes how these genes interact, and the type of signaling interactions between them. The results obtained on several independent data sets show that the proposed approach is very promising. This analysis method has been implemented as a Web-based tool, Pathway-Express, freely available as part of the Onto-Tools (http://vortex.cs.wayne.edu).

**Acknowledgments**

This material is based upon work supported by the following grants: NSF DBI-0234806, CCF-0438970, 1R01HG003491-01A1.
An impact analysis of gene signaling pathways


Received December 11, 2006; accepted in revised form June 28, 2007.
A systems biology approach for pathway level analysis

Sorin Draghici, Purvesh Khatri, Adi Laurentiu Tarca, et al.

*Genome Res.* 2007 17: 1537-1545 originally published online September 4, 2007
Access the most recent version at doi:10.1101/gr.6202607

**Supplemental Material**
http://genome.cshlp.org/content/supp/2007/09/05/gr.6202607.DC1

**References**
This article cites 48 articles, 9 of which can be accessed free at:
http://genome.cshlp.org/content/17/10/1537.full.html#ref-list-1

**License**

**Email Alerting Service**
Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or click here.

To subscribe to *Genome Research* go to:
http://genome.cshlp.org/subscriptions