# Ranking Genes by Their Co-expression to Subsets of Pathway Members

**Priit Adler,[a,§] Hedi Peterson,[a,c,§] Phaedra Agius,[b] Jüri Reimand,[b,d] and Jaak Vilo[b,c]**

[a] *Institute of Molecular and Cell Biology, University of Tartu, Tartu, Estonia*

[b] *Institute of Computer Science, University of Tartu, Tartu, Estonia*

[c] *Quretec Ltd., Tartu, Estonia*

[d] *EMBL Outstation, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, United Kingdom*

**Cellular processes are often carried out by intricate systems of interacting genes and proteins. Some of these systems are rather well studied and described in pathway databases, while the roles and functions of the majority of genes are poorly understood. A large compendium of public microarray data is available that covers a variety of conditions, samples, and tissues and provides a rich source for genome-scale information. We focus our study on the analysis of 35 curated biological pathways in the context of gene co-expression over a large variety of biological conditions. By defining a global co-expression similarity rank for each gene and pathway, we perform exhaustive leave-one-out computations to describe existing pathway memberships using other members of the corresponding pathway as reference. We demonstrate that while successful in recovering biological base processes such as metabolism and translation, the global correlation measure fails to detect gene memberships in signaling pathways where co-expression is less evident. Our results also show that pathway membership detection is more effective when using only a subset of corresponding pathway members as reference, supporting the existence of more tightly co-expressed subsets of genes within pathways. Our study assesses the predictive power of global gene expression correlation measures in reconstructing biological systems of various functions and specificity. The developed computational network has immediate applications in detecting dubious pathway members and predicting novel member candidates.**

*Key words:* **pathway reconstruction; gene expression; systems biology**

## Introduction

A gene carries out its function through its products in one or more biological pathways, forming protein complexes or playing an interactive role in metabolism and signaling. One of the most pertinent tasks in systems biology today is the identification of genetic interactions across different biological conditions. The graphical notation of vertices and edges to represent genes and their interactions within a biological network is a common approach adopted by all main pathway databases.[1–3] Although early collections of biological pathway databases predate the human genome sequencing project, only a small fraction of genes are mapped to at least one pathway. The major pathway databases KEGG[1] and Reactome[2] account for only 4,220 and 1,804 interacting genes out of the approximately 23,000 protein-coding genes that comprise the human genome.[4]

High-throughput gene expression data have proved a rich source of genome-scale information ever since the first studies appeared in

Address for correspondence: Jaak Vilo, Institute of Computer Science, University of Tartu, Liivi 2, 50409 Tartu, Estonia. Voice: +3725049365. vilo@ut.ee

§These authors contributed equally and appear in alphabetical order.

The Challenges of Systems Biology: Ann. N.Y. Acad. Sci. 1158: 1–13 (2009).
doi: 10.1111/j.1749-6632.2008.03747.x 

1997.[5,6] Some interacting proteins exhibit a global correlation in gene expression,[7] allowing researchers to reveal potential members of large protein complexes such as proteasome or ribosome.[8] Most interacting proteins, however, vary in expression over different biological conditions and display more evident correlation in only a selection of conditions. For example, Brand and colleagues[9] show that during the erythroid differentiation switch, MafK changes its dimerization partner from Bach1 to p45 to form the NF-E2 complex. This result highlights the importance of exploring possible gene interactions across many meaningful biological conditions. The number of publicly available gene expression datasets has grown rapidly in recent years, and numerous biological conditions can be included in a gene network study.

The first objective of this work is to study the predictive power of gene co-expression to infer gene memberships of known human pathways, using a large number of gene expression datasets. By conducting an exhaustive series of leave-one-out computation experiments for every gene and related pathway, we retrieve a correlation-based gene rank that reflects the co-expression of the gene and the pathway in many biological conditions. Our results show that correlation ranks vary significantly across different types of pathways. Expression correlation clearly distinguishes members of transcriptionally regulated basal biological processes such as metabolism and translation, while failing to recover signaling networks where post-translational modifications and ligand–receptor interactions have a dominant regulatory role.

It has been argued that gene networks contain active subpathways that are more relevant to the expression of a gene than all the pathway members taken together.[10] Genes that code components for the same protein complex need to have similar expression patterns to keep the complex intact. For example, genes in cell cycle minichromosome maintenance (MCM) and origin recognition complex (ORC) form co-expressed pathway subsets that are differentially expressed compared to cell cycle regulators (such as CDK6). Indeed, arguing this from a numerical perspective, if one can identify the most correlated subset of a pathway for a gene, the resulting correlation score would be higher than the same score for all pathway members.

The secondary objective of our work is the identification of members and candidates of known pathways using a selection of pathway members as reference. We perform an exploratory analysis that detects the optimal correlation threshold for every gene, so that the subset of genes in a pathway (that is, a subpathway) corresponding to the threshold results in the best gene rank for the pathway in question. In the process of computing subpathway-based co-expression ranks for members of different pathways, we produce lists of candidate genes that exhibit strong similarity to the subpathways. In some cases, we provide supporting evidence for the membership of these candidates in the form of shared protein–protein interactions with the pathway and enrichments in related Gene Ontology (GO) terms.[11] Our method therefore proves to be a promising new direction in detecting pathway member candidates and assessing dubious annotations.

Previous studies have successfully applied gene expression data to infer certain pathways and thus supported the biological assumption of coregulation, similar functionality, and correlated gene expression profiles.[12–16] Nevertheless it is well known that many gene networks are not recoverable by exclusively using gene expression data. Other approaches have used topological information such as gene connectivity.[15] The method proposed in this paper chooses a subset of the most correlated pathway genes to compare a gene ranking without taking into account the existence of links between genes in the pathway subset. We show that despite computing ranks from optimal subpathways, there are several examples of known pathway members with uninformative expression patterns that prohibit the recovery of such genes from expression data alone. Our method

therefore clearly highlights the shortcomings of exclusively using expression data for pathway recovery, and encourages the application of methods such as that used by Gevaert and colleagues in Reference 17 that infer networks from heterogeneous sources of evidence.

## Materials and Methods

### Gene Expression Data and Pathways

In this study, we use the human gene expression atlas compiled by Lukk and colleagues.[18] The compendium covers 6,108 manually curated and quality controlled samples from different biological conditions that have been jointly normalized using the RMA algorithm.[19] The entire sample collection originates from the popular AffyMetrix HG_U133A microchip platform that includes about 12,000 genes that we consider the genome.

Note that a significant fraction of the genes on the HG_U133A platform have multiple microarray probe sets corresponding to a single gene. In our analysis, we select the most favorable probe set for each gene in order to get the highest correlation score between a pair of genes.

This study covers all 35 human pathways from the pathway database Reactome[2] version 23. Reactome includes processes like translation and DNA repair, various signaling pathways, and several metabolic pathways. Pathway associations and gene-probe set mappings are provided by the g:Profiler software.[20]

### Query Genes and Training Genes

First, let us denote a pathway as a set of genes $P = \{p_1, \ldots, p_m\}$ and the set of genes in the genome outside the pathway as $G = \{g_1, \ldots, g_n\}$. Each gene in the pathway is used in turn as a *query gene*, denoted as $p^*$. We refer to the remaining *m-1* genes in the pathway as the *training genes*, the set denoted as $P'$. We derive a rank $r^*$ for the query gene $p^*$ by computing the average co-expression score between $p^*$ and the training genes $P'$, and comparing it to the co-expression scores between $P'$ and the rest of the genes in G.

As we compute the rank for every gene in a selected pathway, we assess the predictive power of gene co-expression in the context of this pathway. As a side result, we obtain a list of candidate genes that are not annotated to the pathway but show strong similarity to its expression patterns, frequently outweighing the co-expression values of annotated members.

### Correlation Scores and Gene Ranks

Our goal is to compute a rank $r^*$ for a selected query gene $p^*$ and a pathway $P$. The rank measures the query gene's co-expression with the genes within the pathway, in contrast to the pathway members' co-expression with all other genes in the genome. We first define a *correlation-based similarity score* to measure the average co-expression between a gene $g$ and the members of a pathway $P = \{p_1, \ldots, p_m\}$. Given that $g$ is not part of $P$ and $m$ is the number of members of $P$, we compute the score as follows:

$$score\ (g, P) = \frac{1}{m} \sum\nolimits_{p \in P} corr(g, p).$$

In order to compute rank $r^*$, we split the pathway of interest into the query gene $p^*$ and a set of training genes $P' = P \setminus p^*$. In other words, we leave the query gene out of the pathway, regarding it to be without annotation. We then calculate the scores $s_i = score(g_i, P')$ for all the genes $g_i$ of $G = \{g_1, \ldots, g_n\}$ outside the pathway, as well as the score $s^* = score(p^*, P')$ for our query gene $p^*$. The *gene rank* $r^*$ of $p^*$ is the position of the score $s^*$ after sorting the list of scores $(s^*, s_1, \ldots, s_n)$ in descending order.

The rank $r^*$ varies within the range of genes on the microarray platform. Intuitively, $r^* = 1$ delivers the fact that, on average, the gene $p^*$ is more similar to the pathway than any other unrelated gene. Increasing rank reflects decreasing expression similarity between $p^*$ and the pathway, meaning that there are other genes in the genome that actually exhibit stronger co-expression to the pathway.

### Correlation and Absolute Correlation

Throughout this study, we apply the standard Pearson correlation method to evaluate expression pattern similarity. Although Pearson correlation admittedly fails to detect certain types of variable dependencies, the comparison between other similarity metrics such as Euclidean distance and mutual information criteria remains beyond the scope of this study.

While studies have found Pearson's correlation to be an effective method for measuring co-expression, it does not account for all interactions. Most importantly, there are well-known instances where pathway members behave as suppressors and have an anticorrelated pattern in their expression profiles. In light of this, we performed a second run of our algorithm using the absolute Pearson measure that equally captures anticorrelation and thus models the activating and inhibiting regulation within the pathway in a more meaningful way.

### Subpathways and Gene Ranks

The gene rank computation algorithm described above uses all the genes in the pathway $P$ as training genes to assess the pathway co-expression of query gene $p^*$. However, it is generally known that pathways may contain subsystems and compartments that show higher co-expression values due to common regulatory mechanisms. We use the term *subpathway* to refer to the optimal pathway subset. It should be noted that in the context of this analysis, co-expression and pathway annotation are the only criteria for constructing subpathways; topological information such as edges and paths between members are not taken into account.

In order to incorporate subpathway associations into our analysis, we extend the ranking method to choose an optimal subset from the *m-1* training genes as reference of $p^*$, rather than automatically measuring the similarity between $p^*$ and all the training genes in $P'$. The subpathway is considered optimal if it includes only those training genes that incur a lowest $r^*$. To this end, we define a threshold $t$ so that the correlation scores $score(g, P)$ are computed using only correlations $corr(g, p) \geq t$. It is the choice of $t$ that matters here; a small $t$ results in all $m-1$ training genes participating in the rank, while an overly stringent $t$ excludes most or all genes from the calculation. The choice of $t$ could be made using no prior information about $p^*$, and a predictive model may be derived. However, this is more of an explorative study that investigates the impact of subpathway considerations on the gene rank. Therefore, we perform an exhaustive search by varying the threshold $t$ over the correlation search space of 0 to 1 in 0.02 steps, and choose the threshold that gives the lowest rank for $p^*$. It should be noted that while absolute Pearson correlations range from 0 to 1, Pearson correlations range from $-1$ to 1; therefore we normalize the Pearson correlations to a one-unit space in order to keep the search space fixed.

We demonstrate in the following section that the optimized procedure yields a lower and thus more informative rank for $p^*$, when compared to the rank derived from all training genes. These results reinforce the notion that the consideration of subpathways can greatly enhance pathway inference techniques.

## Results

In this section we present the results of co-expression ranks for the genes in 35 Reactome pathways (see Table 1). We highlight the advantage of our subpathway method, which selects an optimal set of training genes to calculate the rank for the query gene, compare the Pearson correlation measure with absolute Pearson correlation, and discuss the resulting differences in gene ranks and underlying causes. We also present plots that demonstrate the proportion of high-ranking uninformative genes within different pathways, and discuss the possibility of using our methods to uncover potential pathway genes and verify dubious annotations.

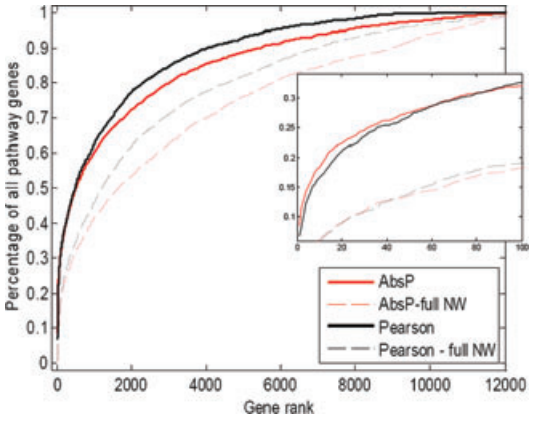


**FIGURE 1.** Receiver operating characteristic (ROC) curves indicating the percentage of pathway genes that have a rank of less than indicated on the x-axis. Large areas under curve (AUC) indicate better performance.

## The Subpathway Advantage

We study the performance of leave-one-out computations with four different experimental settings (see Fig. 1). First, we use either all the training genes in the given pathway to infer the gene rank ("FullNW" in figure) or apply the method of finding an optimal subpathway. Second, we use either Pearson correlation ("Pearson") or absolute Pearson correlation ("AbsP") to measure gene co-expression.

Figure 1 shows the percentage of all pathway genes (y-axis) that have a rank less than or equal to the x-axis value. Such graphical displays, known as receiver operating characteristic (ROC) curves, have proven reliable for evaluating the quality of a classification function.[21] The area under curve (AUC) gives an idea of the performance of the method. Curves that hug the upper left corner are optimal and have an AUC close to 1.0; in contrast, curves of random methods approach the plot diagonal and result in an AUC of 0.5. In case of absolute correlation, the AUC of our optimal subpathway method is 0.865, which is nearly a 15% advance over the full pathway AUC of 0.755. We see similar values for standard correlation: the ROC curve for the subpathway method has an AUC of 0.894, almost 13% more than the AUC of the full pathway approach.

## Standard and Absolute Pearson Correlation

Regarding the comparison of standard and absolute Pearson correlation, the ROC curves in Figure 1 display an AUC difference of 3% to the advantage of standard Pearson correlation, mainly covering genes with ranks above 1,000. A possible explanation is that remotely related genes in the pathway involve a larger number of low correlations, and the absolute score is more pronounced since it involves both positive and negative correlation. On the one hand, the absolute correlation generally includes more genes and therefore produces a higher and potentially noisier rank than the more selective standard correlation. On the other hand, Pearson correlation recovers a higher number of pathway genes with ranks less than 100 (see Fig. 1), suggesting that absolute correlation better captures closely correlated and anticorrelated genes. Although these differences are interesting, there is insufficient statistical support to decide upon the advantage of one method over another.

However, while using absolute Pearson correlation, we uncovered a strong example of RHOQ gene in the pathway *Signaling by Rho GTPases*, where absolute correlation ranks the gene first, while the standard correlation rank is almost 500. The gene with highest anticorrelation with RHOQ is RHOT2. The pair are known to be induced during myogenic and neuronal differentiation and have differential expression in muscle and brain.[22] This may account for the anticorrelation found in our study.

## Low-Ranking Pathway Genes

Earlier we established that our subpathway method results in lower gene ranks. Here we examine in detail the ranks of pathway genes below selected thresholds. In Figure 2, we show the percentage of pathway genes (y-axis) with lower ranks than defined thresholds (x-axis). The thresholds are derived using the Pearson

**TABLE 1.** Summary table of leave-one-out analysis for 35 reactome pathways

| | **abspear** | | | | **abspear_fixed** | | | | **pear** | | | | **pear_fixed** | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | T1 | T10 | T100 | T1000 | T1 | T10 | T100 | T1000 | T1 | T10 | T100 | T1000 | T1 | T10 | T100 | T1000 |
| Apoptosis (36) | 0 | 0.03 | 0.222 | 0.333 | 0 | 0 | 0.028 | 0.167 | 0 | 0.03 | 0.167 | 0.472 | 0 | 0 | 0.056 | 0.361 |
| Cell cycle checkpoints (64) | 0.17 | 0.28 | 0.422 | 0.656 | 0 | 0.11 | 0.312 | 0.656 | 0.09 | 0.28 | 0.422 | 0.688 | 0 | 0.03 | 0.328 | 0.672 |
| Cell cycle, mitotic (102) | 0.12 | 0.22 | 0.392 | 0.657 | 0 | 0.07 | 0.343 | 0.627 | 0.08 | 0.22 | 0.392 | 0.667 | 0 | 0.05 | 0.333 | 0.637 |
| DNA repair (85) | 0.02 | 0.08 | 0.129 | 0.447 | 0 | 0.02 | 0.082 | 0.365 | 0.05 | 0.08 | 0.118 | 0.459 | 0 | 0.01 | 0.094 | 0.4 |
| DNA replication (49) | 0.18 | 0.29 | 0.49 | 0.776 | 0 | 0.16 | 0.429 | 0.714 | 0.18 | 0.33 | 0.49 | 0.776 | 0 | 0.08 | 0.429 | 0.714 |
| **Electron transport chain (64)** | **0.11** | **0.33** | **0.688** | **0.922** | **0** | **0.13** | **0.594** | **0.875** | **0.09** | **0.33** | **0.703** | **0.938** | **0** | **0.14** | **0.594** | **0.891** |
| Gene expression (203) | 0.09 | 0.3 | 0.448 | 0.698 | 0 | 0.06 | 0.26 | 0.552 | 0.03 | 0.2 | 0.423 | 0.758 | 0 | 0.06 | 0.253 | 0.609 |
| Hemostasis (96) | 0.09 | 0.16 | 0.26 | 0.646 | 0 | 0.01 | 0.031 | 0.198 | 0.03 | 0.14 | 0.292 | 0.635 | 0.01 | 0.01 | 0.031 | 0.344 |
| HIV infection (134) | 0.01 | 0.04 | 0.146 | 0.504 | 0 | 0 | 0.058 | 0.328 | 0.01 | 0.02 | 0.175 | 0.489 | 0 | 0 | 0.088 | 0.314 |
| Influenza infection (68) | 0.16 | 0.27 | 0.368 | 0.691 | 0.01 | 0.09 | 0.294 | 0.588 | 0.13 | 0.24 | 0.353 | 0.706 | 0.03 | 0.1 | 0.294 | 0.588 |
| Integration of energy metabolism (37) | 0 | 0.05 | 0.135 | 0.405 | 0 | 0.03 | 0.027 | 0.135 | 0 | 0.03 | 0.162 | 0.378 | 0 | 0 | 0.027 | 0.108 |
| Lipid and lipoprotein metabolism (118) | 0.03 | 0.08 | 0.203 | 0.483 | 0 | 0 | 0.025 | 0.102 | 0.03 | 0.09 | 0.212 | 0.534 | 0 | 0.01 | 0.042 | 0.263 |
| Membrane trafficking (9) | 0.11 | 0.11 | 0.333 | 0.778 | 0 | 0.11 | 0.222 | 0.778 | 0.11 | 0.11 | 0.333 | 0.778 | 0 | 0.11 | 0.222 | 0.778 |
| Metabolism of amino acids (65) | 0.02 | 0.05 | 0.185 | 0.554 | 0 | 0.02 | 0.031 | 0.2 | 0 | 0.05 | 0.215 | 0.6 | 0 | 0.02 | 0.108 | 0.369 |
| Metabolism of carbohydrates (70) | 0.03 | 0.11 | 0.214 | 0.586 | 0 | 0 | 0.029 | 0.214 | 0.03 | 0.1 | 0.271 | 0.629 | 0 | 0 | 0.057 | 0.214 |
| Metabolism of noncoding RNA (17) | 0.24 | 0.35 | 0.471 | 0.824 | 0 | 0.18 | 0.353 | 0.824 | 0.24 | 0.35 | 0.412 | 0.824 | 0 | 0.12 | 0.294 | 0.824 |
| Metabolism of vitamins and cofactors (39) | 0.05 | 0.08 | 0.154 | 0.333 | 0 | 0.03 | 0.051 | 0.103 | 0.08 | 0.1 | 0.103 | 0.359 | 0 | 0 | 0.051 | 0.128 |
| Metabolism of xenobiotics (59) | 0.07 | 0.2 | 0.305 | 0.407 | 0 | 0.05 | 0.169 | 0.305 | 0.05 | 0.2 | 0.288 | 0.508 | 0 | 0.12 | 0.203 | 0.407 |
| mRNA processing (116) | 0.09 | 0.16 | 0.284 | 0.681 | 0.01 | 0.04 | 0.216 | 0.534 | 0.09 | 0.14 | 0.302 | 0.681 | 0.01 | 0.04 | 0.224 | 0.543 |
| Nucleotide metabolism (78) | 0.03 | 0.05 | 0.231 | 0.59 | 0.01 | 0.01 | 0.141 | 0.385 | 0.01 | 0.04 | 0.244 | 0.603 | 0 | 0.01 | 0.09 | 0.397 |
| Porphyrin metabolism (10) | 0 | 0.3 | 0.4 | 0.5 | 0 | 0.3 | 0.3 | 0.4 | 0.1 | 0.3 | 0.4 | 0.6 | 0.1 | 0.2 | 0.3 | 0.4 |

*Continued.*

**TABLE 1.** *Continued*

| | abspear | | | | abspear_fixed | | | | pear | | | | pear_fixed | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | T1 | T10 | T100 | T1000 | T1 | T10 | T100 | T1000 | T1 | T10 | T100 | T1000 | T1 | T10 | T100 | T1000 |
| Post-translational protein modification (32) | 0.03 | 0.06 | 0.156 | 0.531 | 0 | 0 | 0 | 0.125 | 0.03 | 0.06 | 0.25 | 0.5 | 0 | 0 | 0.062 | 0.125 |
| **Pyruvate metabolism and TCA cycle (24)** | **0.17** | **0.29** | **0.708** | **0.792** | **0** | **0.17** | **0.458** | **0.75** | **0.21** | **0.33** | **0.708** | **0.833** | **0.08** | **0.21** | **0.5** | **0.792** |
| Signaling by EGFR (6) | 0 | 0 | 0.167 | 0.167 | 0 | 0 | 0 | 0.167 | 0 | 0 | 0.167 | 0.333 | 0 | 0 | 0 | 0.167 |
| Signaling by FGFR (20) | 0 | 0 | 0.1 | 0.4 | 0 | 0 | 0.05 | 0.15 | 0 | 0.05 | 0.1 | 0.4 | 0 | 0.05 | 0.05 | 0.15 |
| Signaling by insulin receptor (32) | 0.03 | 0.09 | 0.281 | 0.531 | 0 | 0 | 0 | 0.031 | 0.03 | 0.06 | 0.188 | 0.5 | 0 | 0 | 0 | 0.094 |
| **Signaling by NGF (9)** | **0** | **0** | **0.111** | **0.333** | **0** | **0** | **0** | **0.111** | **0** | **0** | **0.111** | **0.333** | **0** | **0** | **0** | **0.111** |
| Signaling by Notch (13) | 0 | 0 | 0.077 | 0.462 | 0 | 0 | 0 | 0.308 | 0 | 0 | 0.231 | 0.538 | 0 | 0 | 0.077 | 0.385 |
| Signaling by Rho GTPases (105) | 0.03 | 0.09 | 0.171 | 0.419 | 0.01 | 0.02 | 0.067 | 0.257 | 0 | 0.07 | 0.19 | 0.438 | 0 | 0.01 | 0.057 | 0.19 |
| Signaling by TGF beta (11) | 0 | 0 | 0.091 | 0.636 | 0 | 0 | 0.091 | 0.364 | 0 | 0.09 | 0.091 | 0.727 | 0 | 0 | 0.091 | 0.455 |
| Signaling by Wnt (10) | 0 | 0 | 0.1 | 0.5 | 0 | 0 | 0.1 | 0.4 | 0 | 0 | 0.1 | 0.5 | 0 | 0 | 0.1 | 0.2 |
| Signaling in immune system (136) | 0.13 | 0.24 | 0.471 | 0.699 | 0 | 0.02 | 0.074 | 0.434 | 0.13 | 0.24 | 0.456 | 0.743 | 0.01 | 0.03 | 0.103 | 0.537 |
| Telomere maintenance (40) | 0.08 | 0.18 | 0.225 | 0.625 | 0 | 0.03 | 0.175 | 0.375 | 0.08 | 0.2 | 0.275 | 0.575 | 0 | 0.03 | 0.1 | 0.425 |
| Transcription (136) | 0.03 | 0.07 | 0.147 | 0.471 | 0 | 0 | 0.074 | 0.324 | 0.03 | 0.07 | 0.176 | 0.493 | 0 | 0 | 0.074 | 0.338 |
| **Translation (137)** | **0.31** | **0.58** | **0.693** | **0.839** | **0.08** | **0.42** | **0.599** | **0.73** | **0.28** | **0.53** | **0.708** | **0.854** | **0.06** | **0.43** | **0.628** | **0.788** |

T($x$) represents the recovery rate of pathway genes in top $x$ candidates. F for example, T10 shows the percentage of genes that were recovered within rank 10 or less for the given pathway. The comparison between standard (pear) and absolute Pearson (abspear) correlation is performed for standard leave-one-out using full pathways (pear_fixed, abspear_fixed), as well as using the optimal subpathway (pear, abspear) method. Pathways discussed in the text are marked in bold.

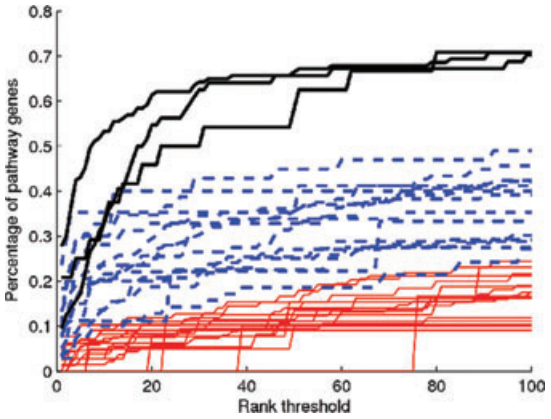


**FIGURE 2.** ROC curves indicating the percentage of pathway genes that have a rank of less than indicated on the x-axis. Three pathways have more than 50% of the genes with a rank below 100 and are drawn in solid black lines. Pathways where less than 50% of the genes have ranks below 100 are drawn in dotted blue lines, and pathways where less than 25% genes have ranks below 100 are drawn in red lines.

correlation and vary from 1 to 100 for each pathway. Pathways with more than 50% of their gene ranks below the highest threshold of 100 are indicated in black, pathways with more than 25% and less than 50% of genes are in blue, and pathways with up to 25% are in red. There are three pathways in black with about 70% of their genes having ranks smaller than 100: *Electron Transport Chain*, *Pyruvate metabolism and TCA cycle*, and *Translation*.

## The Best Predicted Pathway—Translation

Our methods recovered a good number of genes with the best possible rank of 1. Rank 1 was assigned to 21% of the genes in *Pyruvate metabolism and TCA cycle*, 24% of *Metabolism of noncoding RNA*, and 31% of *Translation*. The *Translation* pathway consists mainly of ribosomal genes, translation initiation factors, and elongation factors. Ribosome itself is a complex of ~80 proteins that need to be highly co-expressed to form the functional cellular machinery. We identified seven genes (RPS3A, RPS26P10, RPS12, RPL31, RPL10, RPS8, and RPL3L) that have been shown to be part of either a small or large subunit of the ribosome complex, but whose expression is not as similar to other ribosomal genes as compared to the rest of the protein complex components. Extraribosomal activity has been shown for at least three of the above seven genes. For example, RPS3A takes part in apoptosis,[23] RPL10 is known to interact with GDNF and to be active in neurite growth,[24] and RPL31 (also known as chondromodulin-3) participates in synthesis of chondrocytes.[25] Such biological insights suggest a reason for the different expression profiles that distinguish those genes from the other ribosome components.

## The Worst Predicted Pathway—Signaling by NGF

The genes in signaling pathways appear to have the worst ranks in our results, which is an unsurprising observation given the nature of signaling. Signaling pathway components do not necessarily exhibit co-expression since most of the signaling does not involve stable protein complexes or transcriptional control; instead post-translational modifications and temporal ligand–receptor bindings mediate the processes. Using pathway-specific perturbation data would probably give better results in recovering signaling genes and identifying good candidates.

The *nerve growth factor* (NGF) pathway consists of nine genes showing low co-expression values and unsatisfactory ranks. Interestingly, our methods detect genes that are not annotated to the pathway but nevertheless show high co-expression to pathway genes. We have identified five such genes as potential pathway candidates. One of the genes, Dynamine 1 (DNM1) is a nerve terminal phosphoprotein known to be upregulated after NGF induction of PC12 cell differentiation into neurons.[26,27] The other four candidates (SMPD1, NR1D1, PPP3CC, and L3MBTL) are known to be expressed in the brain (see for example GeneCards[28]).

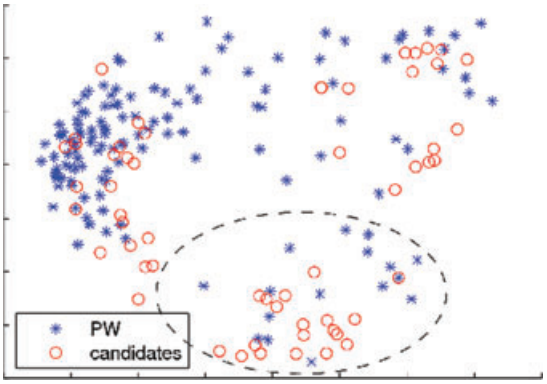


**FIGURE 3.** A two-dimensional plot of the *Translation* pathway genes and candidates. The figure depicts the result of principal component analysis (PCA) of gene correlation scores. Note the three subgroups that form in the figure. The genes represented by stars in the ellipse are initiation factors.

## Low-Ranking Non-pathway Genes

When a rank $r^*$ for a query gene $p^*$ in the pathway is greater than 1, this inherently implies that the method recovered $r^* - 1$ lower-ranked genes that are not annotated to the pathway. Here we investigate the possibility of these genes playing a functional role in the pathway, and we refer to them as candidate genes. We compile a list of 54 candidate genes with rank $\leq 30$ for the *Translation* pathway and compute absolute correlation scores between the candidates and all the 137 pathway genes.

Upon further investigation, we found that five of these candidate genes had an average rank of less than 10: P4HA1, a ribosomal protein homolog PD-1; RACK1, a gene that is localized in the 40S subunit[29] and regulates translation through the RACK1-PKC complex; ENSG00000214925, a novel Ensembl transcript prediction that represents the ribosomal protein L44E; EEF1B2, elongation factor 1-beta; and FEN1, the only well-ranking candidate that lacks such a well-defined connection to the pathway.

Using the scores as features, we apply principal component analysis (PCA) to project the data into two dimensions and observe the proximity of pathway genes and candidates (Fig. 3). The PCA plot also suggests that pathway genes and candidates can be partitioned into three main groups. In fact, the pathway in question consists of three main types of genes: ribosomal genes, elongation factors, and initiation factors. We indicate the initiation factors by drawing an ellipse that encapsulates the pathway genes (indicated by the blue stars in Fig. 3) and a number of candidate genes (indicated by the red circles in Fig. 3).

Interestingly, genes inside the ellipse in Figure 3 fall into two main categories: translation initiation factor complex genes and spindle complex genes. We found a reference for *Xenopus laevis* where a gene named Maskin (TACC3 in *H. sapiens*) is shown to associate both with the spindle assembly[30] and the translation initiation factor complex eIF4F.[31] Although no transcriptional control has been described between these associates, it is possible that Maskin plays a crucial role in the formation and regulation of both complexes.

## Gene Ontology Analysis of Candidate Genes

We investigated pathway candidates through a functional enrichment analysis of Gene Ontology (GO) terms,[11] KEGG[1] and Reactome[2] pathways, miRBASE microRNA (miRNA) target sites[32] and TRANSFAC regulatory motifs.[33] The functional analysis was performed using the g:Profiler software.[20] The numbers of relevant categories are gathered in Table 2. We found several statistically significant examples of candidates sharing common GO annotations with pathway genes. It is worth noting that the candidates of the pathway *Signaling in Immune System* are enriched in immune response (GO:0006955) and immune system process (GO:0002376) with *P*-values below $1.0^{-25}$. Another example is the *Mitotic Cell Cycle* pathway where 13 candidate genes with an average rank $<30$ are mapped to various GO terms related to mitotic cell cycle (such as GO:0000278) with a *P*-value of $8.78^{-18}$. These results indicate that our method

**TABLE 2.** Protein–protein interactions and GO analysis for candidate genes

| | Candidate genes | Protein–protein interactions | Number of relevant GO categories | Mean number of GO candidate genes |
|---|---|---|---|---|
| Apoptosis (36) | 396 | 34 | 1 | 12 |
| Cell cycle checkpoints (64) | 137 | 82 | 25 | 10.44 |
| Cell cycle, mitotic (102) | 165 | 91 | 20 | 12 |
| DNA repair (85) | 273 | 103 | 18 | 11 |
| DNA replication (49) | 131 | 46 | 9 | 8.56 |
| **Electron transport chain (64)** | **290** | **12** | **8** | **9.38** |
| Gene expression (203) | 592 | 162 | 1 | 21 |
| Hemostasis (96) | 412 | 88 | 14 | 13.29 |
| HIV infection (134) | 515 | 190 | 11 | 9.55 |
| Influenza infection (68) | 325 | 0 | 6 | 13.83 |
| Integration of energy metabolism (37) | 351 | 0 | N/A | N/A |
| Lipid and lipoprotein metabolism (118) | 354 | 37 | 7 | 14.14 |
| Membrane trafficking (9) | 127 | 2 | 2 | 9 |
| Metabolism of amino acids (65) | 400 | 6 | 3 | 9.67 |
| Metabolism of carbohydrates (70) | 465 | 15 | N/A | N/A |
| Metabolism of noncoding RNA (17) | 177 | 75 | 14 | 10.86 |
| Metabolism of vitamins and cofactors (39) | 453 | 4 | N/A | N/A |
| Metabolism of xenobiotics (59) | 264 | 3 | 7 | 12.57 |
| mRNA processing (116) | 371 | 78 | N/A | N/A |
| Nucleotide metabolism (78) | 322 | 18 | 3 | 13.67 |
| Porphyrin metabolism (10) | 142 | 0 | 1 | 4 |
| Post-translational protein modification (32) | 255 | 0 | 7 | 14 |
| **Pyruvate metabolism and TCA cycle (24)** | **208** | **0** | **2** | **5** |
| Signaling by EGFR (6) | 51 | 0 | N/A | N/A |
| Signaling by FGFR (20) | 241 | 1 | N/A | N/A |
| Signaling by insulin receptor (32) | 381 | 67 | 1 | 34 |
| **Signaling by NGF (9)** | **109** | **0** | **N/A** | **N/A** |
| Signaling by Notch (13) | 156 | 1 | N/A | N/A |
| Signaling by Rho GTPases (105) | 460 | 33 | 4 | 14.75 |
| Signaling by TGF beta (11) | 186 | 0 | N/A | N/A |
| Signaling by Wnt (10) | 151 | 2 | N/A | N/A |
| Signaling in immune system (136) | 380 | 0 | 27 | 13.33 |
| Telomere maintenance (40) | 202 | 55 | 19 | 13.16 |
| Transcription (136) | 453 | 214 | 5 | 14.4 |
| **Translation (137)** | **524** | **62** | **N/A** | **N/A** |

The total number of resulting candidate genes for each pathway is shown (column 2), together with the number of protein–protein interactions between pathway members and candidates (column 3). The last two columns show the number of relevant GO categories for each pathway and the average number of candidate genes corresponding to a category. Pathways discussed in the text are marked in bold.

successfully suggests candidate genes that share characteristic GO terms with related pathways.

## Protein–Protein Interactions

We further studied the discovered candidate genes by exploring corresponding protein–protein interactions (PPI) from the databases Intact,[34] HPRD,[35] and ID_SERVE,[36] using the GraphWeb software.[37] We found that pathway genes have interactions with candidate genes in the majority of pathways (see Table 2). Using this approach, we detected a candidate gene TNFRSF10C for the *Apoptosis* pathway. This gene has protein–protein interactions to four

pathway genes and is annotated to the *Apoptosis* pathway in KEGG[1] and to various apoptosis-related terms in GO. TNFRSF10C was not part of the *Apoptosis* pathway in Reactome version 23 used in this study, but it appears in the newest version 24, released in March 2008.

## Conclusions

We present a computational study that assesses the power of microarray gene expression measurements to detect complex interactions of genes and proteins of known pathways under various biological conditions. We define a gene rank-based similarity measure that takes into account the co-expression correlation of genes across many microarray samples. We then apply the measure to predict the members of 35 human pathways from the Reactome database, using more than 6,000 microarray samples. We perform exhaustive leave-one-out experiments for all the genes in the pathways, deriving gene ranks based on their expression similarity to other members of the corresponding pathways.

Our study leads to several biologically sound conclusions. Ranking of pathway components shows significant improvement when only a subset of pathway genes is used for expression correlation reference. Such effect refers to the existence of closely related subsets of genes within pathways that display stronger similarity to the gene in question and potentially relate to common function and regulation.

The reliability of gene expression measurements in pathway analysis shows significant variation depending on the biological role of the pathway in question. Our approach of assessing the limitations of gene expression analysis through an extensive study of curated human pathways is novel in the field. Our work establishes this in a strong way because, despite choosing the optimal subpathway and rank, some pathway genes persistently achieve insignificant ranks and remain unrecoverable in a pathway. Genes involved in essential base processes such as metabolism and translation have smaller, informative ranks and are generally well predicted by our method. Signaling pathways, on the other hand, exhibit larger uninformative ranks due to lower levels of co-expression. These conclusions can be explained by different mechanisms of pathway regulation, where genes in the former processes are governed by consistent transcriptional regulation and exhibit strong co-expression, while the post-translational and cell signaling regulation dominant in the latter processes is less related to the mRNA expression levels measured with microarrays.

Although gene expression alone cannot explain the complexity of biological pathways, our computational framework is immediately applicable to a number of interesting problems. Expression similarity ranks can be used to detect candidate genes that are co-expressed with known pathway components. Gene expression data from other species can be incorporated via gene orthology mapping in order to improve pathway inference through conserved co-expression patterns. Protein–protein interactions and Gene Ontology analysis can be applied for filtering candidate genes resulting from our methods. Candidates can then be tested through experimental approaches and, given sufficient evidence, incorporated to curated pathways. A similar strategy can be applied for verifying dubious pathway annotations.

### Acknowledgments

This work has been supported by the EU FP6 grants ENFIN LSHG-CT-2005-518254 and Estonian Science Foundation grant ETF7437. Jüri Reimand acknowledges funding from EITSA Tiger University and Marie Curie Biostar. We wish to thank Margus Lukk for providing the dataset for the study, Sven Laur for proofreading and useful comments, and anonymous reviewers for helpful comments.

### Conflicts of Interest

The authors declare no conflicts of interest.

## References


1. Kanehisa, M. *et al*. 2008. KEGG for linking genomes to life and the environment. *Nucleic Acids Res.* **D36:** D480–D484.
2. Västik, I. *et al*. 2007. Reactome: a knowledge base of biologic pathways and processes. *Genome Biol.* **8:** R39.
3. Caspi, R. *et al*. 2008. The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res.* **D36:** D623–D631.
4. Flicek, P. *et al*. 2008. Ensembl 2008. *Nucleic Acids Res.* **D36:** D707–D714.
5. DeRisi, J.L., V.R. Iyer & P.O. Brown. 1997. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278:** 680–686.
6. Lockhart, D.J. *et al*. 1996. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.* **14:** 1675–1680.
7. Kemmeren, P. *et al*. 2002. Protein interaction verification and functional annotation by integrated analysis of genome-scale data. *Mol. Cell.* **9:** 1133–1143.
8. Stuart, J.M. *et al*. 2003. A gene-coexpression network for global discovery of conserved genetic modules. *Science* **302:** 249–255.
9. Brand, M. *et al*. 2004. Dynamic changes in transcription factor complexes during erythroid differentiation revealed by quantitative proteomics. *Nat. Struct. Mol. Biol.* **11:** 73–80.
10. Chen, C. *et al*. 2007. A search engine to identify pathway genes from expression data on multiple organisms. *BMC Syst. Biol.* **1:** 20.
11. Ashburner, M. *et al*. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*. **25:** 25–29.
12. D'haeseleer P., S. Liang & R. Somogyi. 2000. Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics* **16:** 707–726.
13. Segal, E., H. Wang & D. Koller. 2003. Discovering molecular pathways from protein interaction and gene expression data. *Bioinformatics*. **19:** i264–i272.
14. Bansai, M. *et al*. 2007. How to infer gene networks from expression profiles. *Mol. Syst. Biol.* **3:** 78.
15. Ulitsky, I. & R. Shamir. 2007. Identification of functional modules using network topology and high-throughput data. *BMC Syst. Biol.* **1:** 8.
16. Pe'er, D. *et al*. 2001. Inferring subnetworks from perturbed expression profiles. *Bioinformatics* **17:** S215–S224.
17. Gevaert, O., S. Van Vooren & B. De Moor. 2007. A framework for elucidating regulatory networks based on prior information and expression data. *Ann. N. Y. Acad. Sci.* **1115:** 240–248.
18. Lukk, M. *et al*. 2008. Application of text mining to create human gene expression atlas from public data. *Manuscript in preparation*. Data set accession number in ArrayExpress: E-MTAB-62.
19. Irizarry, R.A. *et al*. 2003. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4:** 249–264. 2
20. Reimand, J. *et al*. 2007. g:Profiler—a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Res.* **D35:** W193–W200.
21. Egan, J. P. 1975. Signal detection theory and ROC analysis. In *Series in cognition and perception*. Scientific Press. New York.
22. Tomoyuki, A. *et al*. 2003. Small GTPase Tc10 and its homologue RhoT induce N-WASP-mediated long process formation and neurite outgrowth. *J. Cell Sci.* **116:** 155–168.
23. Naora, H. 1999. Involvement of ribosomal proteins in regulating cell growth and apoptosis: Translational modulation or recruitment for extraribosomal activity. *Immunol. Cell Biol.* **77:** 197–205.
24. Park, S. & D.G. Jeong. 2006. Ribosomal protein L10 interacts with the SH3 domain and regulates GDNF-induced neurite growth in SH-SY-5y cells. *J. Cell. Biochem.* **99:** 624–634.
25. Hiraki, Y. *et al*. 1996. A novel growth-promoting factor derived from fetal bovine cartilage, chondromodulin III. *J. Biol. Chem.* **37:** 22657–22662.
26. Scaife, R. & R.L. Margolis. 1990. Biochemical and immunochemical analysis of rat brain dynamin interaction with microtubules and organelles in vivo and in vitro. *J. Cell Biol.* **111:** 3023–3033.
27. Liu, J.P. *et al*. 1994. Dynamin I is a Ca(2+)-sensitive phospholipid-binding protein with very high affinity for protein kinase C. *J. Biol. Chem.* **269:** 21043–21050.
28. Safran, M. *et al*. 2002. GeneCards™ 2002: towards a complete, object-oriented, human gene compendium. *Bioinformatics* **11:** 1542–1543. 18.
29. Nilsson, J. *et al*. 2004. Regulation of eukaryotic translation by the RACK1 protein: a platform for signalling molecules on the ribosome. *EMBO Rep.* **5:** 1137–1141.
30. O'Brien, L.. 2005. The xenopus TACC homologue, maskin, functions in mitotic spindle assembly. *Mol. Biol. Cell* **16:** 2836–2847.
31. Cao, Q., J.H. Kim & J.D. Richter. 2006. CDK1 and calcineurin regulate Maskin association with eIF4E and translational control of cell cycle progression. *Nat. Struct. Mol. Biol.* **13:** 1128–1134.

32. Griffits-Jones, S. 2006. miRBASE: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res.* **34:** D140–D144.
33. Matys, V.. 2006. TRANSFAC® and its module TRANSCompel®: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.* **34:** D108–D110.
34. Kerrien, S. *et al*. 2007. IntAct – open source resource for molecular interaction data. *Nucleic Acids Res.* **D35:** D561–D565.
35. Peri, S. *et al*. 2003. Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res.* **13:** 2363–2371.
36. Ramani, A.K. *et al*. 2005. Consolidating the set of known human protein-protein interactions in preparation for large-scale mapping of the human interactome. *Genome Biol.* **6:** R40.
37. Reimand, J. *et al*. 2008. GraphWeb: mining heterogeneous biological networks for gene modules with functional significance. *Nucleic Acids Res.* **36:** W452–W459.