# Metabolic pathway activity estimation from RNA-Seq data

Yvette Temate-Tiagueu[1], Meril Mathew[2], Igor Mandric[1], Qiong Cheng[3], Olga Glebova[1], Nicole Beth Lopanik[2], Ion Măndoiu[4], and Alex Zelikovsky[1]

[1] Department of Computer Science, Georgia State University, Atlanta GA
[2] Department of Biology, Georgia State University, Atlanta, GA
[3] Department of Pharmacology, University of Miami, Miami FL
[4] Department of Computer Science and Engineering, University of Connecticut, Storrs, CT

**Abstract.**
An interesting approach to study the metabolic differences between species is metabolic pathway. In this study, we characterize pathways activity levels of two samples. We applied our proposed methods on RNA-Seq *Bugula neritina* metagenomics data. We successfully identified several differential pathway activity and we selected 3 of them for qPCR validation.

**Keywords:** RNA-Seq reads, KEGG, metabolic pathways, ortholog groups

## 1 Introduction

For the past several years, transcriptome sequencing through deep sequencing technologies or RNA-Seq, has revolutionized sequencing technologies with the many advantages it provides. Because of RNA-Seq, it is easier to characterize transcripts and their isoforms, to detect genes without need of prior information in the form of probe, also RNA-Seq can estimate expression level of transcript over a wide range with good precision.
The problem of transcriptome quantification has been recently shown extremely important since an estimated rate of 84% of protein level variation can be explained by transcription alone without taking in account variation in translation and degradation [1, 2] while the rate drops to only 73% for microarray data.
This paper primary goal is to develop highly accurate algorithms for metabolic pathway activity level estimation and testing differential pathway activity. Activity levels will be inferred using expectation maximization algorithms applied to novel uniform binary and maximum likelihood models while robust testing of pathway significance will be achieved by employing a novel graph-based approach.
In contrast to array-based methods, pathway analysis based on RNA-seq data does not measure gene expression directly but allows inference based on total RNA content. When applied to metatranscriptome data, the first challenge of pathway analysis is to decide which metabolic pathways are active in the sampled community (i.e., pathway activity detection). Recent software tools (*MEGAN4* [3] and *MetaPathways* [4] using SEED and KEGG [5] annotations) enable the organization of transcripts into

ortholog groups and pathways by collecting all pathways represented by at least one ortholog group. The parsimonious approach *MinPath* [6] identifies the smallest family of pathways covering all expressed ortholog groups. A more elaborate MCMC approach takes into account the co-occurrences of genes in more than one pathway for analyzing metagenomic data [7]. Following pathway detection, the second major challenge of pathway analysis is to infer pathway activity levels to enable detection of differential expression. Few existing tools incorporate this step, a major focus of this paper.

Methods that treat pathways as simple gene sets [8, 9] are popular even though they do not use all information available. In recent years, a number of pathway analysis methods have been developed that combine knowledge of pathway topology (e.g., gene position on the pathway, gene-gene interactions, etc) with gene expression data based on comparative analyses (reviewed in [10]). Such methods have been applied primarily to experimental studies of single organisms. Despite the inherent pathway architecture of microbial biochemical function, relatively few analyses of complex metatranscriptomic datasets incorporate pathway-level inference of metabolic activity. We explore this new analysis techniques on a particular metagenomics RNA-Seq data from *Bugula neritina*.

In this paper, we represent metabolic pathways as graphs that use nodes to represent biochemical compounds, with enzymes associated with edges describing biochemical reactions. Ideally, a comprehensive pathway analysis method would be able to take into consideration the position and role of each gene in a pathway, the efficiency with which a certain reaction is carried out, and the rate limiting conditions. With genome data, it is possible to consider pathways size, gene length and overlap in gene content among pathways [7] to compute the relative abundance of pathways and pathway ranking, but this approach might not work with RNA-Seq data.

Another representation of pathways we will use in this paper is to view the ortholog group as a set cover. We will use a binary ortholog group expression model to tell if there is or not RNA-seq evidence for the expression of a given ortholog group in a given sample.

The validation step of these methods consist of extracting the proteins involve in our estimated differential pathways activity levels, and analyzing their expression levels or transcript frequency estimation. We expect to see the differential pathway activity confirmed at the protein and contigs level. We carry this final analysis through the novel bootstrapping tool IsoDE [11].

Our experimental study was made on Bugula neritina RNA-Seq data. Using the two novel computational approaches we implemented, we were able to find differentially expressed pathways from the data. This result is been validated by quantitative PCR (qPCR) conducted using a housekeeping gene also identified in the data. Since the qPCR experiment is time consuming and expensive, the in-vitro analysis is limited to the following pathways: K04369, K05087 and K16332 selected from our results.

## 2 Methods

### 2.1 Binary model of pathway activity

In this section we present an EM-based algorithms for inferring pathway activity levels based on metatranscriptome sequence data. By *de novo* co-assembly of RNA-seq data

and BLAST-ing resulting contigs against protein databases, with a certain confidence, we can infer the ortholog groups expressed in the sample. From the pathway databases we can easily extract the enzyme information associated with each pathway. Let $w$ be a pathway that is considered to be a set of enzymes represented by their ortholog groups $w = \{p_1, \ldots, p_k\}$. Since an ortholog group can have multiple functions and participate in multiple pathways, the pathways can be viewed as a family of subsets $W$ of the set of all ortholog groups $P$. Below we first introduce a uniform binary pathway activity model based on a discrete ortholog group expression model.

The uniform binary pathway activity model is based on the assumptions of *uniformity*, namely that each molecule from an ortholog group participates in each active pathway with the same probability (i.e., in equal proportions) and of *binary activity*, which postulates that a pathway is active if the level of ortholog group activity exceeds a certain (possibly pathway dependent) threshold. Formally, let $\delta(w)$ be a binary variable indicating the *activity status* of $w$, i.e., $\delta(w) = 1$ if $w$ is active and $\delta(w) = 0$, otherwise. Also let the *activity level* of pathway $w$ be the summation over constituent ortholog groups $g$ of their participation $g_w$ in $w$. Since we assume that each ortholog group $g$ is equally likely to participate in each pathway containing it, it follows that $g_w = (1 + \sum_{w' \ni p, w' \neq w} \delta(w'))^{-1}$ and the activity level $f_w$ of pathway $w$ is given by

$$f_w = \sum_{g \in w} g_w = \sum_{g \in w} \frac{1}{1 + \sum_{w' \ni g, w' \neq w} \delta(w')} \tag{1}$$

The binary activity status of $w$ is computed from its activity level $f_w$ and the threshold $T_w$ as follows

$$\delta(w) = \begin{cases} 0 & \text{if } f_w < T_w \\ 1 & \text{if } f_w \geq T_w \end{cases} \tag{2}$$

The uniform binary model described by equations (1)-(2) can be solved using a simple iterative algorithm. The algorithm starts with assigning activity status $\delta(w) = 1$ to each pathway $w \in W$, i.e., $\Delta^0(W) = \{\delta^0(w) | w \in W\} \leftarrow 1$ and then repeatedly updates the activity level according to (1) and the activity status according to (2). The procedure terminates when the status sequence $\Delta^0(W) = 1, \Delta^1(W), \Delta^2(W), \ldots$ starts to oscilate $\Delta^{n+k}(W) = \Delta^n(W)$. In all our preliminary experiments, an oscilation with period $k = 2$ is achieved in at most 10 iterations. Also the threshold $T_w$ does not significantly change the order of pathways sorted with respect to their activity levels estimated as the mean $f_w$ after convergence. This model is better explain by the right side of Figure 1.

### 2.2 Graph-based estimation of pathway significance

In our second approach, each pathway can be viewed as a network of enzymes also called EC numbers (Enzyme Commission number). In this paper, we convert pathways to graphs - vertice and nodes components - to compute their statistical significance. We also propose to distinguish active pathways using a permutation model for finding significant pathway alignments and motifs [12].

This model assumes that the subset of expressed enzymes in an active annotated pathway should be connected. The enzyme permutation model finds the average vertex degree in the subgraph induced by expressed enzymes. Then the same parameter is computed for sufficiently many random permutations of enzyme labels. The statistically significant match should have density higher than in 95% of permutations. Specific characteristics of the graph taken into account in our analysis are: (1) Number of nodes. A node represents a protein that got mapped during BLAST. In KEGG, their color is green as shown in Figure **??**, (2) Number of green connected components, (3) Largest Number of nodes in connected component and (4) Largest Number edges in connected component.
Using these metrics, we compute the density of the induced graph composed by only mapped proteins. We obtain the names of those proteins through EC numbers on the graph. We also compute the density without mapping, assuming all proteins were detected in the organism. Below, we present two graph-based models, the vertex label swapping and the edge swapping, to analyze pathways. This model is better explain by the left side of Figure 1.

**Model 1: Vertex label swapping**
In this model, we keep the same topology but we allow swapping of labels between two vertices. One known issue of this approach is, vertex with high degree always get connected. This might lead to too many significant matches.

**Model 2: Edge swapping**
Because of the bias in the vertex label swapping model, we will also implement the edge swapping. Here, the plan is to keep the in-degree and out-degree of each node the same, swapping nodes only if these value does not change. We keep vertex labels the same.

## 3 Results and Discussion

### 3.1 Results

We used KEGG to generate pathways from Trinity contigs and proteins form BLASTX as input. Then we extracted all pathways along with all mapped protein. KEGG represents proteins as ko number and we also follow this representation. The next step was to download all KGML - KGML is an exchange format of KEGG pathway maps- files associated with the pathways using the API provided by KEGG. To convert KGML files to graph of node and vertices, we implemented and ran a novel tool called KGML-Pathway2Graph. Mapping the output of KGMLPathway2Graph with ko numbers from KEGG analysis of our data, allowed us to compute pathway significance through $p\text{-}value$.
This analysis was made between sample 1 and sample 2 of the *Bugula neritina* data. Sample 1 contains the Symbiont bacteria while this symbioisis relationship is not present in the Bugula from sample 2. Results are presented in Table 1.
In the auxiliary materials, we present the following results: (1) Transcripts differential expression (DE) analysis results: DE contigs using IsoDE [11] and Fisher's exact test

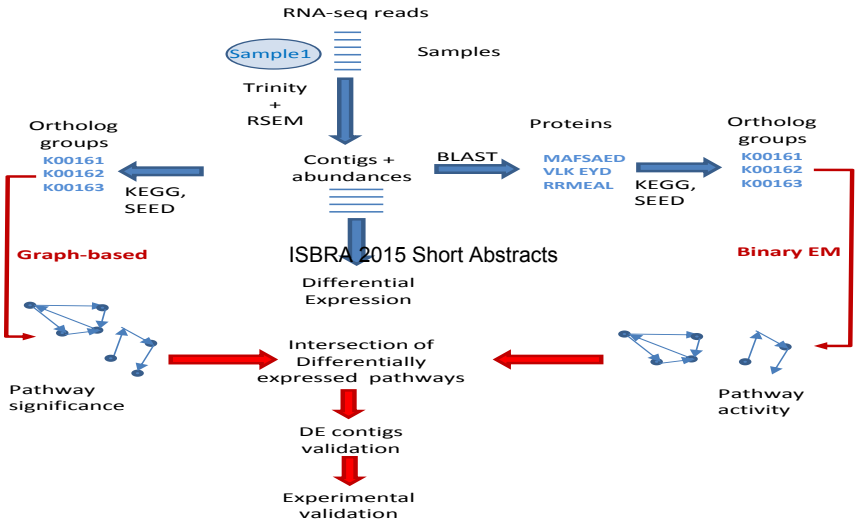


Fig. 1: Proposed metatranscriptomic analysis flows.

with housekeeping gene), (2) Pathway activity estimation results and finally (3) Differential Analysis of pathway expression.

From our statistical analysis, We got some pathways which were found differentially expressed by all methods. The next step is to experimentally validate these results. The following housekeeping gene - contig code m.4423 - was experimentally detected and will be used to validate our results. Since the qPCR experiment is time consuming and expensive, then we will limit the in-vitro analysis to the following pathways: K04369, K05087 and K16332 .

| **Pathway** | **P-value1** | **P-value2** | **Expression1** | **Expression2** | **Diff. Sign.** | **Diff. Act.** | **Inter** |
|---|---|---|---|---|---|---|---|
| KO04068 | 8542 | 7871 | 6094 | 14160 | yes | yes | yes |
| KO01230 | 8458 | 7766 | 5975 | 14097 | yes | yes | yes |
| KO04020 | 8204 | 7592 | 5841 | 13713 | yes | yes | yes |
| KO04145 | 8088 | 7472 | 5766 | 13567 | yes | no | yes |
| KO05012 | 7906 | 22C | A 23s. | 22C | yes. | yes | yes |

Table 1: P-valuel1 and P-value2 are respectively from Vertex label and Edge swapping model. Expression1 and Expression2 represent the expression of the pathway activity. This table presents the most significant divergence in pathway results, using the criteria described in section 2, they are declared differentially significant.

### 3.2 Discussion

Although all the EM and the graph-based methods worked on the same data generated by KEGG, the input to each approach were very different. Example, running trinity output of sample1 on KEGG generate about 179 pathways. All of these pathways were considered for EM methods while only a small subset of 80 was used as input to each of the graph-based model. Different factor contributed to only about one third of the pathways to be analyzed in the edge/vertex swapping model: (1) we were not able to extract the KGML of all pathways from from Kegg; (2) we were not able to convert all KGML to actual graph and (3) some graph didn't carry enough mapping to be be significant (we excluded pathways with less than 3 ortholog group mapped).
Consequently, the graph-based approaches yield considerable less results than EM methods although results from both models in the graph-based approaches were very consistent. Also, the graph-based analysis appears to be more stringent selecting only the pathways which are the farthest apart according to our statistic criteria.

## 4 Conclusions

Biomelcular interactions through tool like KEGG provide huge data and enable us to have a better understanding of metabolic pathways. Two approaches were designed to estimate pathway activity as well as pathway significance, a graph-based and an expectation maximization approach. Our experimental comparisons on *Bugula neritina* RNA-seq data is able to show at the protein level, the difference in the pathways activity of two samples. Our results will undergo a final validation through qPCR analysis.

## References

1. Li, J.J., Bickel, P.J., Biggin, M.D.: System wide analyses have underestimated protein abundances and the importance of transcription in mammals. PeerJ **2**, 270 (2014)
2. Li, J.J., Biggin, M.D.: Statistics requantitates the central dogma. Science **347**(6226), 1066–1067 (2015)
3. Huson, D.H., Mitra, S., Ruscheweyh, H.-J., Weber, N., Schuster, S.C.: Integrative analysis of environmental sequences using MEGAN4. Genome research **21**(9), 1552–1560 (2011)
4. Konwar, K.M., Hanson, N.W., Pagé, A.P., Hallam, S.J.: Metapathways: a modular pipeline for constructing pathway/genome databases from environmental sequence information. BMC bioinformatics **14**(1), 202 (2013)
5. Kanehisa, M., Goto, S.: Kegg: kyoto encyclopedia of genes and genomes. Nucleic acids research **28**(1), 27–30 (2000)
6. Ye, Y., Doak, T.G.: A parsimony approach to biological pathway reconstruction/inference for genomes and metagenomes. PLoS computational biology **5**(8), 1000465 (2009)
7. Sharon, I., Bercovici, S., Pinter, R.Y., Shlomi, T.: Pathway-based functional analysis of metagenomes. Journal of Computational Biology **18**(3), 495–505 (2011)
8. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., *et al.*: Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America **102**(43), 15545–15550 (2005)

9. Efron, B., Tibshirani, R.: On testing the significance of sets of genes. The annals of applied statistics, 107–129 (2007)
10. Mitrea, C., Taghavi, Z., Bokanizad, B., Hanoudi, S., Tagett, R., Donato, M., Voichita, C., Drăghici, S.: Methods and approaches in the topology-based analysis of biological pathways. Frontiers in physiology **4** (2013)
11. Al Seesi, S., Tiagueu, Y.T., Zelikovsky, A., Măndoiu, I.I.: Bootstrap-based differential gene expression analysis for rna-seq data with and without replicates. BMC genomics **15**(Suppl 8), 2 (2014)
12. Cheng, Q., Zelikovsky, A.: Combinatorial optimization algorithms for metabolic networks alignments and their applications. IJKDB **2**(1), 1–23 (2011)