Metabolic analysis of metatranscriptomic data from planktonic communities

Igor Mandric[1], Sergey Knyazev[1], Cory Padilla[2], Frank Stewart[2], Ion I. Măndoiu[3], and Alex Zelikovsky[1]

[1] Department of Computer Science, Georgia State University, Atlanta, GA, USA
imandric1@student.gsu.edu, skniazev1@student.gsu.edu, alexz@cs.gsu.edu
[2] School of Biological Sciences, Georgia Institute of Technology, Atlanta, GA, USA
frank.stewart@biology.gatech.edu, cpadilla7@gatech.edu
[3] Computer Science and Engineering Department, University of Connecticut, Storrs, CT, USA,
ion@engr.uconn.edu

Abstract. This paper describes an enhanced method for analyzing microbial metatranscriptomic (community RNA-seq) data using Expectation - Maximization (EM)-based differentiation and quantification of predicted gene, enzyme, and metabolic pathway activity. Here, we demonstrate the method by analyzing the metatranscriptome of planktonic communities in surface waters from the Northern Louisiana Shelf (Gulf of Mexico) during contrasting light and dark conditions. The analysis reveals that the level of transcripts encoding proteins of oxidative phosphorylation varys little between day and night. In contrast, transcripts of pyrimidine metabolism are significantly more abundant at night, whereas those of carbon fixation by photosynthetic organisms increase 2-fold in abundance from night to day.

1 Introduction

RNA-seq is a standard method for comparative analysis of gene transcription across different conditions. It supplanted a widely used microarray approach, enabling analysis of a much larger number of genes, including those represented in pools of transcripts from complex multi-species communities (metatranscriptomes). RNA-seq allows researchers to determine and compare gene transcription levels, as well as the transcriptional activity of distinct metabolic pathways. Diverse bioinformatic tools have been developed to facilitate comparisons of RNA-seq data [1–10]. Such tools include web-based services with automated pipelines that allow assessment of the metabolic properties represented in RNA-seq datasets. For example, the MAP platform [11] predicts genes expressed in samples, while also provides information about gene classification into orthology groups (see figure 1). Unfortunately, such pipelines fail to quantify transcripts in concert with the annotation step. We therefore propose an enhanced pipeline that combines the biochemical annotation with quantification analysis. For this purposes, we propose to use an expectation-maximization (EM) technique similar to one from IsoEM2 [12]. We tested our algorithm using metatranscriptome data from marine bacterioplankton sampled during both the day and nighttime, and therefore likely exhibiting predictable variation in community transcription patterns.

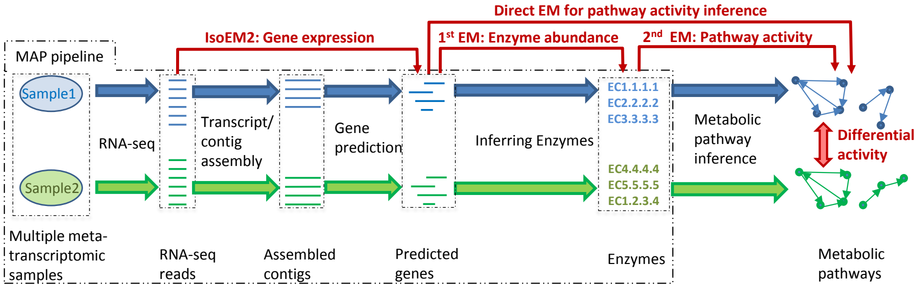

Fig. 1: The pipeline MAP and the enhanced pipeline for quantification and differential analysis of the metabolic pathway activity. The quantification enhancements are drawn in red.

2 Methods

In this section we describe the procedure of inferring metabolic pathway activity levels from RNA-Seq data for naturally occurring microbial communities. We also apply differential pathway activity level analysis similar to the non-parametric statistical approach described in [13], which was successfully applied for gene differential expression.

A general meta-omic pipeline is described on Figure 1. Several metatranscriptomic samples are sequenced on an Illumina Hi-Seq (2x150 bp) and the resulting reads are assembled into a set of contigs. Genes detected on the contigs are mapped against protein databases and enzymatic functions are inferred. Finally, the representation of metabolic pathways is inferred based on the presence/absence of enzymes within each pathway. The above generic pipeline has been described in [11]. This paper proposes to enhance the above pipeline with the inference of metabolic pathway activity levels using repeated maximum likelihood inference and resolution by the Expectation - Maximization (EM) algorithm. The proposed inferences are depicted in red on Figure 1.

Inference of pathway activity levels The first step is to estimate the abundances of the assembled contigs. The abundances can be inferred by any RNA-seq quantification tool. Here, we suggest using IsoEM2 [12], as this method is sufficiently fast to handle Illumina Hiseq data and more accurate than kallisto [14]. The next proposed step is to estimate the abundance of enzymes based on contig abundances. For this step we propose so-called *1-st EM*. The *2-nd EM* is used to infer metabolic pathway activity levels based on inferred enzyme abundances and databases of metabolic pathways. The 1-st and the 2-nd EM's can be also integrated into a single *direct EM* that directly infers pathway activity levels from contig abundances. All componentsm (1-st EM, 2-nd EM and direct EM) are built with similarities to IsoEM2 methodology.

Differential analysis of pathway activity Using the estimates of pathway activity levels in the differential pathway activity analysis requires estimating uncertainty. The extension of our bootstrapping approach introduced in [15] is useful for the direct maximum likelihood model since the pathway activity levels are inferred directly from RNA-seq reads that can be resampled. The current version of IsoEM2 allows the user to generate bootstrapped samples from the RNA-Seq reads and to infer abundance estimates, based on Fragments Per Kilobase of transcript per Million mapped reads (FPKM). We estimate pathway activity level for each of the bootstrapped samples and then run a differential expression (DE) analysis similar to the one described in [13].

3 Results

In this section we apply our analysis pipeline to two conditions (day. night) of a planktonic marine microbial community. We describe a subset of the most abundant pathways and conduct a differential pathway activity level analysis that highlights statistically significant functional features from the repertoire of metabolic processes occurring in the community.

Datasets. The samples were collected from surface waters (2 m depth) at 12:30 and 23:55 (local time) at a station on the Northern Louisiana Shelf (Gulf of Mexico) in July 2015. Seawater (1 L) was pumped directly onto a 0.22 um Sterivex filter, preserved in 1.8 ml of RNA-later and flash frozen. Samples were stored a -80 C until extraction. RNA was isolated from the samples by a phenol-chloroform method following the Mirvana RNA kit protocol. Samples were treated with DNase to remove residual DNA signal from the metatranscriptome. The RNA-Seq data were generated via Illumina HiSeq 2500 sequencing at the Department of Energy Joint Genome Institute (DOE-JGI). Detailed information about the two samples is provided in the Table 1.

Sample				Reads			Contigs	
Name	Depth	Code	Time	Length	Count	Insert size	Total	Total length
Day	2m	177_2m	12:30 PM	2× 151 bp	89.4 M	195±49	94.7 k	58.3 MB
Night	2m	240_2m	11:55 PM	2× 151 bp	91.4 M	187±49	108 k	68.1 MB

Table 1: Dataset description

MAP pipeline. A preliminary annotation of RNA-seq data was obtained using the DOE-JGI Metagenome Annotation Pipeline (MAP v.4) (JGI portal) [11]. The MAP processing consists of feature prediction including identification of protein-coding genes. In this pipeline, the MEGAHIT metagenome assembler is used to first assemble RNA-Seq reads into scaffolds. Further, several software suites (GeneMark.hmm, MetaGeneAnnotator, Prodigal, FragGeneScan) are used to predict genes on assembled scaffolds. The MAP pipeline also annotates genes according to EC numbers, which are a necessary input in our maximum likelihood model. The annotations are obtained via homology searches (using USEARCH) against a non-redundant proteins sequence database (maxhits=50, e-value=0.1) where each protein is assigned to a KEGG Orthology group (KO). The top 5 hits for each KO, with the condition that the identity score

Pathway		Abundance reads $\times 10^3$	
Code	Description	Day	Night
ko00190	Oxidative phosphorylation (Energy metabolism)	2260	2700
ko00710	Carbon fixation in photosynthetic organisms (Energy metabolism)	837	422
ko00240	Pyrimidine metabolism (Nucleotide metabolism)	644	1110
ko00270	Cysteine and methionine metabolism (Amino acid metabolism)	568	176
ko00020	Citrate cycle - TCA cycle (Carbohydrate metabolism)	525	411
ko00900	Terpenoid backbone biosynthesis (Metabolism of terpenoids and polyketides)	508	261
ko01230	Biosynthesis of amino acids	333	471
ko00195	Photosynthesis (Energy metabolism)	327	63
ko00230	Purine metabolism (Nucleotide metabolism)	318	618
ko00630	Glyoxylate and dicarboxylate metabolism (Carbohydrate metabolism)	299	530
ko00061	Fatty acid biosynthesis (Lipid metabolism)	37	179

Table 2: 10 most abundant pathways in the Day and Night samples.

is at least 30% and 70% of the protein length is matched, are used. The KO IDs are translated into EC numbers using KEGG KO to EC mapping.

The enhanced quantification pipeline. Our enhanced pipeline is depicted in red on Figure 1. We start our analysis from the RNA-Seq metatranscriptomic reads. First, we find the abundance estimates (frequencies) for each metatranscriptomic gene/transcript by applying Maximum Likelihood abundance estimation. For this purpose we use IsoEM2. The custom GTF annotation file needed for supplying each run of IsoEM2 was prepared by using the fastaToGTF script from the same software suite. Next, we use FPKM estimates as the weights of each transcript for inferring abundances of each EC number. We use transcripts to EC notation alignments as provided by the MAP pipeline.

Highly active pathways. Table 2 shows the 10 most active pathways in the Day sample sorted in descending order of their activity level, i.e., the number of reads attributed by the proposed maximum likelihood model. The 11th pathway listed (ko0061) is among the 10 most active at night but is not among the 10 most active in the day. Similarly, the pathway ko00195 is among the most 10 active at night but is not among the 10 most active in the day. All other 9 pathways are among the most active during both night and day.

Differential pathway analysis. In Table 3 there is a list of all metabolic pathways which are up-regulated at noon with at least 1.7 fold change, 95% confidence and at least 1000 reads assigned by EM. The values of abundances are given at 95% confidence interval upper boundary (therefore, they are slightly greater than in the Table 2). In Table 4 there is a list of all metabolic pathways which are up-regulated at noon with at least 1.7 fold change, 95% confidence and at least 1000 reads assigned by EM.

Discussion. The results in Tables 2-4 are reflective of planktonic microbial communities driven by a diurnal cycle. During the daytime, pathways mediating photosynthesis, carbon fixation, and the building blocks for amino acid biosynthesis are the most abundant. At night there is an increase in nucleotide and lipid generation, probably for new cell production. In general, the community appears to be gaining energy and substrates during the day and expending them at night by generating crucial cellular components. This is supported by the differential expression between the day and night transcript pools, with energy (photosynthesis) and small organic molecule synthesis (e.g, fructose, glutamine-glutamate, glycosaminoglycan, etc.) being up-regulated during the day and the synthesis of larger biomolecules at night (e.g. lipid metabolism, amino acids, and carotenoids). There is a clear shift in energy sources between day and night. While oxidative phosphorylation is highly transcribed at both time points, it is clear that photosynthesis elevates some of this energy requirement. This is evidenced by a slight decrease of oxidative phosphorylation and increase of TCA-related transcripts during the day, potentially replenishing the NADH/NADPH reserves for the use of the electron transport chain at night. As predcited, these results indicate a community undergoing diel cycling, thereby providing validation of our proposed EM-based pipeline and suggesting this method as an valuable tool for coupled annotation and quantification of metabolic pathways in community RNA-seq data.

Acknowledgements

IM, SK and AZ were partially supported from NSF Grants 1564899 and 16119110, IM and SK were partially supported by GSU Molecular Basis of Disease Fellowship, IIM was partially supported from NSF Grants 1564936 and 1618347, CP and FS were partially supported by NSF Grants 1151698, 1558916, and 1564559, and Simons Foundation award 346253.

References

1. Donato, M., Xu, Z., Tomoiaga, A., Granneman, J.G., MacKenzie, R.G., Bao, R., Than, N.G., Westfall, P.H., Romero, R., Draghici, S.: Analysis and correction of crosstalk effects in pathway analysis. Genome research **23**(11), 1885–1893 (2013)
2. Efron, B., Tibshirani, R.: On testing the significance of sets of genes. The annals of applied statistics, 107–129 (2007)
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. 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)
6. Sharon, I., Bercovici, S., Pinter, R.Y., Shlomi, T.: Pathway-based functional analysis of metagenomes. Journal of Computational Biology **18**(3), 495–505 (2011)
7. 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)

Pathway		reads in 10^3	
Code	Description	Day	Night
ko00051	Fructose and mannose metabolism (Carbohydrate metabolism)	326	34.1
ko00195	Photosynthesis (Energy metabolism)	488	93.1
ko00261	Monobactam biosynthesis (Biosynthesis of other secondary metabolites)	237	44.5
ko00410	beta-Alanine metabolism (Metabolism of other amino acids)	10.0	0.01
ko00471	D-Glutamine and D-glutamate metabolism	6.79	0
ko00532	Glycosaminoglycan biosynthesis - chondroitin sulfate / dermatan sulfate	28.8	3.65
ko00533	Glycosaminoglycan biosynthesis - keratan sulfate	22.9	0.609
ko00604	Glycosphingolipid biosynthesis - ganglio series	4.17	0
ko00660	C5-Branched dibasic acid metabolism (Carbohydrate metabolism)	4.39	0.01
ko00930	Caprolactam degradation (Xenobiotics biodegradation and metabolism)	3.80	0.883
ko00332	Carbapenem biosynthesis (Biosynthesis of other secondary metabolites)	10.3	1.54
ko00565	Ether lipid metabolism (Lipid metabolism)	10.4	0.682
ko00590	Arachidonic acid metabolism (Lipid metabolism)	51.8	19.4
ko00270	Cysteine and methionine metabolism (Amino acid metabolism)	787	246
ko00514	Other types of O-glycan biosynthesis (Glycan biosynthesis and metabolism)	7.75	2.96
ko00450	Selenocompound metabolism (Metabolism of other amino acids)	201	80.2
ko00710	Carbon fixation in photosynthetic organisms(Energy metabolism)	1000	487
ko00983	Drug metabolism - other enzymes (Xenobiotics biodegradation & metabolism)	58.3	16.5
ko00520	Amino sugar and nucleotide sugar metabolism (Carbohydrate metabolism)	265	123

Table 3: Up-regulated pathways in the Day sample

8. Tarca, A.L., Draghici, S., Bhatti, G., Romero, R.: Down-weighting overlapping genes improves gene set analysis. BMC bioinformatics **13**(1), 136 (2012)
9. Temate-Tiagueu, Y., Seesi, S.A., Mathew, M., Mandric, I., Rodriguez, A., Bean, K., Cheng, Q., Glebova, O., Măndoiu, I., Lopanik, N.B., Zelikovsky, A.: Inferring metabolic pathway activity levels from rna-seq data. BMC Genomics **17**(5), 542 (2016). doi:10.1186/s12864-016-2823-y
10. Ye, Y., Doak, T.G.: A parsimony approach to biological pathway reconstruction/inference for genomes and metagenomes. PLoS computational biology **5**(8), 1000465 (2009)
11. Huntemann, M., Ivanova, N.N., Mavromatis, K., Tripp, H.J., Paez-Espino, D., Tennessen, K., Palaniappan, K., Szeto, E., Pillay, M., Chen, I.-M.A., *et al.*: The standard operating procedure of the doe-jgi metagenome annotation pipeline (map v. 4). Standards in genomic sciences **11**(1), 17 (2016)
12. Mandric, I., Temate-Tiagueu, Y., Shcheglova, T., Seesi, S.A., Zelikovsky, A., Mandoiu, I.: Fast bootstrapping-based estimation of confidence intervals of expression levels and differential expression from rna-seq data. Bioinformatics (to appear)
13. 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**(8), 2 (2014)
14. Bray, N.L., Pimentel, H., Melsted, P., Pachter, L.: Near-optimal probabilistic rna-seq quantification. Nature biotechnology **34**(5), 525–527 (2016)
15. Al Seesi, S., Mangul, S., Caciula, A., Zelikovsky, A., Măndoiu, I.: Transcriptome reconstruction and quantification from rna sequencing data. Genome Analysis: Current Procedures and Applications, 39 (2014)

Pathway		reads in 10^3	
Code	Description	Day	Night
ko00053	Ascorbate and aldarate metabolism (Carbohydrate metabolism)	0	1.88
ko00061	Fatty acid biosynthesis (Lipid metabolism)	55.9	270
ko00120	Primary bile acid biosynthesis (Lipid metabolism)	2.75	116
ko00140	Steroid hormone biosynthesis (Lipid metabolism)	0	4.11
ko00232	Caffeine metabolism (Biosynthesis of other secondary metabolites)	0	1.05
ko00260	Glycine, serine and threonine metabolism (Amino acid metabolism)	49.3	227
ko00311	Penicillin and cephalosporin biosynthesis	0	2.74
ko00365	Furfural degradation (Xenobiotics biodegradation and metabolism)	0	2.12
ko00430	Taurine and hypotaurine metabolism (Metabolism of other amino acids)	3.19	62.3
ko00472	D-Arginine and D-ornithine metabolism (Metabolism of other amino acids)	0	1.25
ko00780	Biotin metabolism (Metabolism of cofactors and vitamins)	7.05	48.6
ko00906	Carotenoid biosynthesis (Metabolism of terpenoids and polyketides)	0	26.2
ko00984	Steroid degradation (Xenobiotics biodegradation and metabolism)	0	2.07
ko00362	Benzoate degradation (Xenobiotics biodegradation and metabolism)	3.58	16.7
ko00592	alpha-Linolenic acid metabolism (Lipid metabolism)	0.19	2.89
ko00072	Synthesis and degradation of ketone bodies (Lipid metabolism)	2.67	11.6
ko00364	Fluorobenzoate degradation (Xenobiotics biodegradation and metabolism)	0.180	2.96
ko01051	Biosynthesis of ansamycins (Metabolism of terpenoids and polyketides)	0	3.38
ko00760	Nicotinate and nicotinamide metabolism (Mcofactors and vitamins)	30.2	103
ko00281	Geraniol degradation (Metabolism of terpenoids and polyketides)	1.57	170
ko00627	Aminobenzoate degradation (Xenobiotics biodegradation and metabolism)	0.949	4.06
ko00730	Thiamine metabolism (Metabolism of cofactors and vitamins)	10.4	35.4
ko00643	Styrene degradation (Xenobiotics biodegradation and metabolism)	0.958	22.6
ko01200	Carbon metabolism	13.7	86.9
ko00220	Arginine biosynthesis (Amino acid metabolism)	3.53	11.0
ko00440	Phosphonate and phosphinate metabolism	1.30	5.33
ko00905	Brassinosteroid biosynthesis (Metabolism of terpenoids and polyketides)	2.00	35.6
ko00941	Flavonoid biosynthesis (Biosynthesis of other secondary metabolites)	2.84	6.03
ko00720	Carbon fixation pathways in prokaryotes (Energy metabolism)	1.36	15.9
ko00290	Valine, leucine and isoleucine biosynthesis (Amino acid metabolism)	68.0	193
ko00403	Indole diterpene alkaloid biosynthesis	0	2.68
ko01053	Biosynthesis of siderophore group nonribosomal peptides	0	1.16
ko00920	Sulfur metabolism (Energy metabolism)	47.7	135
ko00625	Chloroalkane and chloroalkene degradation	24.3	51.8

Table 4: Up-regulated pathways in the Night sample