Estimating enzyme participation in metabolic pathways for microbial communities from RNA-seq data

F. Rondel[1], R. Hosseini[1], B. Sahoo[1], S. Knyazev[1], I. Mandric[2], Frank Stewart[3], I. I. Măndoiu[4], B. Pasaniuc[5], and A. Zelikovsky[1]

[1] Department of Computer Science, Georgia State University, Atlanta, USA
{frondel1,ahosseini3,bsahoo1,skniazev1}@student.gsu.edu, alexz@gsu.edu
[2] Department of Computer Science, University of California Los Angeles, Los Angeles, CA, USA, imandric@ucla.edu
[3] Department of Microbiology and Immunology, Montana State University, Bozeman, MT, USA, frank.stewart@montana.edu
[4] Computer Science and Engineering Department, University of Connecticut, Storrs, CT, USA, ion@engr.uconn.edu
[5] Bioinformatics Interdepartmental Program, University of California, Los Angeles, California, USA

Abstract. Abstract. Metatranscriptome sequence data analysis is necessary for understanding biochemical changes in the microbial community and their effects. In this paper, we propose a methodology to estimate activities of individual metabolic pathways to better understand the activity of the entire metabolic network. Our novel pipeline includes an expectation-maximization based estimation of enzyme expression and simultaneous estimation of pathway activity level and enzyme participation level in each pathway. We applied our novel pipeline to metatranscriptome data generated from surface water planktonic communities sampled over a day-night cycle in the Northern Gulf of Mexico (Louisiana Shelf). Our results show that estimates of enzyme expression and pathway activity levels are robust and stable. The estimated enzyme and pathway activity levels suggest a 24-hour cycle in microbial community transcription. We also found that as expected, the activity levels of the majority of pathways correlate with environmental parameters. Finally, the estimated enzyme participation levels in each pathway are stable across all data points, implying that our method can quantify the role of each enzyme in each metabolic pathway.

Keywords: NGS, enzyme expression, pathway activity level, microbial community, metatranscriptome, enzyme participation in pathways

1 Introduction

Measuring the functional activity, enrichment, and interaction of metabolic pathways in microbial communities is essential for understanding the biochemical and ecological contributions of microorganisms. Despite many advances in using microbial biomolecules (DNA, RNA, proteins) to assess the biochemical contributions of microbes, it remains challenging to quantify how the expression of individual enzymes contributes to the activity of multi-enzyme metabolic pathways. In this study, we analyze time-series metatranscriptomic (community RNA) data to generate an efficient model for understanding metabolic pathway activity in depth [21, 6, 16, 20]. Even though advances in high-throughput sequencing have aided the exploration of RNA sequencing data, particularly for single organisms, it is often challenging to disentangle community-level data [16, 22, 5], notably as existing pathway analysis tools (e.g., MEGAN4, MetaPathways, MinPath) often yield variable conclusions about the activity of pathways based on RNA data [8, 11, 23, 19]. To overcome the current challenges, we developed a workflow that uses a Maximum Likelihood-based model and annotations based on the KEGG [9] database to estimate transcript frequency, enzyme expression, enzyme participation in pathways, and metabolic pathway activity. In this paper, we test this model using metatranscriptomic data from a marine microbial community. The data span multiple time points with different environmental parameters to elucidate the complex metabolic pathway activity in the microbial community, generally challenging to mimic in the laboratory.

Here we describe our methodology as the first to use a likelihood model to infer the pathway activity considering an enzyme's expression (transcription) and participation coefficient. First, to remove noise caused by unrelated metabolic pathways, we filtered the microbial community-specific metabolic pathways from the KEGG database. Moreover, for the removal of noise due to enzymes, we merged the expression of enzymes sharing the same contigs

and having sequence homologs. We implemented a novel Expectation-Maximization algorithm to estimate the enzyme participation level in each pathway and then used these estimations for more accurate predictions of pathway activity. We validated our results by multiple and marginal linear regression between estimated metabolic pathway activity and environmental parameters. We measured the correlation to understand the enzyme and pathway activity for four different time points and at two depths from the sea surface during a 24-hours diel cycle (three days, 2 24 hour cycles). Our metabolic pathway analysis method, using an advanced Expectation-Maximization algorithm, more accurately estimates the metabolic pathway activity levels from metatranscriptomic data. Our contributions include the following:

- An EM-based algorithm for estimating enzyme expression based NGS read data including grouping homologous enzymes sharing the same contigs
- A novel EM based algorithm for estimating metabolic pathway activity levels using estimation of enzyme participation level in each pathway.
- PubMed-referenced literature-based noise elimination method for extracting the microbial community-specific metabolic pathways
- Validation of enzymes expression and pathways activity using their correlation with the environmental parameters.
- Validation of the enzyme and pathway activity correlation with four different time points and two depths from the sea surface

The rest of the paper is organized as follows. In the next section we describe the pipeline of our software framework and several EM-based algorithms for estimating enzyme expression and metabolic pathway activity in microbial communities. Then we describe our datasets including sequencing data, and extraction of metabolic enzymes and pathways. The next section describes our results of the statistical validation of the proposed pipeline.

2 Methods

We first describe the pipeline containing the previous version of our software and an alternative flow with three new EM algorithms. Then each of these three new EMs are described separately and the global loop for pathway activity level estimation concludes description of our software.

2.1 Pipeline for estimating pathway activity levels

In this section, we describe the procedure of inferring metabolic pathway activity levels from RNA-Seq data for microbiome communities. We also apply differential pathway activity level analysis similar to the non-parametric statistical approach described in [1] which was successfully applied for gene differential expression.

This paper proposes to enhance the pipeline proposed in [13] (see Fig 1) with the inference of enzyme expressions and enzyme participation levels in metabolic pathway repeatedly applying the maximum likelihood model. These models are resolved using the Expectation-Maximization (EM) algorithm. The proposed inferences are highlighted in red (see Fig. 1). The first step is to estimate the abundances of the assembled contigs. The abundances can be inferred by any RNA-seq quantification tool, but we suggest using IsoEM [14] since it is sufficiently fast to handle Illumina Hiseq data and more accurate than Kallisto [2]. We propose to estimate the enzyme expressions based on contig abundances and mapping of contigs onto enzymes (EM for enzyme expression in Fig. 1). The EM for pathway activity levels is based on inferred enzyme expressions and metabolic pathway annotation. Each enzyme is initially assigned a participation level of $1/|w|$, where $|w|$ is the total amount of enzymes in the pathway w. The *Global loop for pathway activity* updates the enzyme participation level by fitting expected enzyme expressions to the expressions estimated by *EM for enzyme expression*.

2.2 EM for enzyme expression estimation

In the microbial community, a single contig can participate in multiple proteins and therefore multiple enzymes. Therefore we need to estimate probability for each contig to be part of each relevant enzyme and use maximum

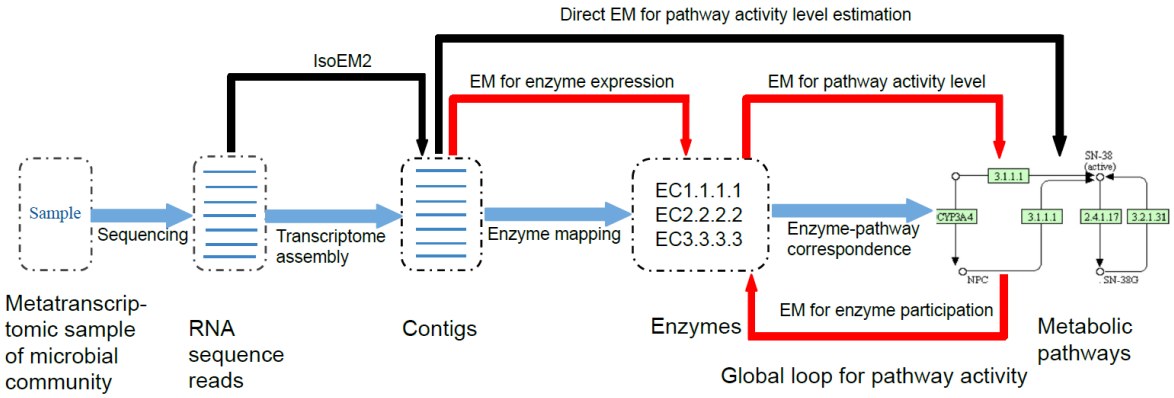

Fig. 1: Pipeline of metabolic pathway analysis for a microbial community sample. The metatranscriptomic data obtained from microbial community samples are sequenced, and raw reads are assembled into contigs. The genes containing obtained contigs are further mapped into the enzyme-pathway database. Contig frequencies are obtained using IsoEM2 [14]. Instead of estimating pathway activity levels using direct EM [12], we propose to first estimate enzyme expression. Then the pathway activity level and enzyme participation coefficients are estimated in a single EM feedback loop.

likelihood (ML) model to estimate enzyme expression. This ML model allows to estimate enzyme expression using the EM algorithm.

Let T be a random variable with values from the set of observed contigs, and let E be a random variable whose values belong to the set of relevant metabolic enzymes from the KEGG database. The probability of observing a contig t is given by the following formula: $P(T=t) = \sum_{w \in W} f_w \; P(T=t \mid E=e)$, where f_e stands for the expression of the relevant metabolic enzyme e. Thus, in our model we adopt the following likelihood function:

$$L(f_e) = \prod_{t \in T} \left(\sum_{e \in E} f_e \; P(T=t \mid E=e) \right)^{a_t}$$

where a_t denotes the abundance of t estimated by IsoEM2. The corresponding log-likelihood is

$$l(f_e) = \sum_{t \in T} a_t \; \log \left(\sum_{w \in W} f_e P(T=t|E=e) \right)$$

The probability of assigning the assembled contig t to a enzyme e is proportional to the normalized bit-score of such alignment. It is computed according to the following formula.

$$P(T=t|E=e) = p_{te} = \frac{b_{te}}{\sum_{t' \in e} b_{t'e}}$$

where b_{te} is the best bit-score obtained from the alignment of t to the protein that have a function of the enzyme e.

The details of the EM for enzyme expression are as follows. We initialize estimates for each enzyme with a random number $f_e \in [0,1], \; e \in E$. Then, we iterate the following two steps until a convergence criteria is satisfied:

The E-step. We first compute the expected number of reads n_e emitted by each enzyme e through the following formula:

$$n_e = \sum_{t \in T} a_t \cdot \frac{p_{te} f_e}{\sum_{e' \in E} p_{te'} f_{e'}}$$

The M-step. The new estimates are provided based on a standard normalization step:

$$f_e^{new} = \frac{n_e}{\sum_{e' \in E} n_{e'}}$$

The algorithm halts when the change in estimates between iterations is small enough: $||f_E^{new} - f_E|| \leq \epsilon$, where $\epsilon \ll 1$

2.3 EM for Estimation of pathway activity level

Here, we apply the EM algorithm for estimating pathway activity levels $f_W = \{f_w | w \in W\}$ based on frequencies of enzymes $f_E = \{f_e | e \in E\}$. Following [12] we use the uniform probability distribution over the set of enzymes/enzyme groups participating in each pathway. We count an enzyme in a pathway only if this enzyme/enzyme group has a non-zero frequency. This means the following:

$$P(E = e|W = w) = p_{ew} = \begin{cases} \frac{1}{|w|}, & \text{if } e \in w \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

We initialize each of the abundance estimates for each pathway with a random number $f_{w_j} = f_j \in [0, 1],\ w_j \in W$. Then, we iterate the following two steps until a convergence criteria is satisfied:

The E-step. Compute the expected amount of enzymes n_w emitted by each pathway w through the following formula:

$$n_w = \sum_{e \in E} f_e \cdot \frac{p_{ew} f_w}{\sum_{w' \in W} p_{ew'} f_{w'}}$$

The M-step. The new estimates are provided based on a standard normalization step:

$$f_w^{new} = \frac{n_w}{\sum_{w' \in W} n_{w'}}$$

The algorithm halts when the change in estimates between iterations is small enough: $||f_W^{new} - f_W|| \leq \epsilon$, where $\epsilon \ll 1$.

2.4 Global loop for pathway activity level estimation

The initial estimate (1) of the participation level of enzyme e in the pathway w can be very far from reality. More accurate estimates of the enzyme participation levels can lead to more accurate estimates for the pathway activity levels. The algorithm below estimates pathway activity levels Steps (1-3) and then checks how well the the computed activities f_w's fit the enzyme expressions (step (4)). If the fit is not good enough, then EM-based algorithm is applied to update the enzyme participation levels p_{ew}'s (Steps (5-6)) and then f_w's are recomputed according to updated p_{ew}'s in Step (3).

1. Find expression $f(e)$ of each enzyme e running EM from Section 2.2.
2. According to (1), initialize $p_{ew} = \frac{1}{|w|}$ for $e \in w$ and $p_{ew} = 0$, otherwise.
3. Find activity levels f_w for each pathway $w \in W$ running EM from Section 2.3.
4. Find expected frequency of each enzyme e according to formula $f_e^{exp} = \sum_{w \in W} p_{ew} f_w$ If expected and observed enzymes frequencies are close to each other: $||f_{e \in E}^{exp} - f_{e \in E}|| = \sum_{e \in E} (f_e^{exp} - f_e)^2 < \epsilon \ll 1$, then exit, i.e. go to step 7.
5. Find better fitted p'_{ew}'s by using the following EM algorithm:
   The E-step. Compute expected p_{ew}^{exp}'s that will make $f_e = f_e^{exp}$ for each $e \in E, w \in W$,

$$p_{ew}^{exp} = p_{ew} \times \frac{f_e}{f_e^{exp}}$$

The M-step. Provide the new estimates by normalization for each $e \in E, w \in W$,

$$p_{ew}^{new} = \frac{p_{ew}^{exp}}{\sum_{e \in E} p_{ew}^{exp}}$$

The algorithm halts when the change in estimates between iterations is small enough: $||p^{new} - p|| = \sum_{e \in E, w \in W} (p_{ew}^{new} - p_{ew})^2 \leq \epsilon \ll 1$

6. For each $e \in E, w \in W$, update $p_{ew} \leftarrow p'_{ew}$ and go to step 3
7. Output $\{f_w | w \in W\}$ and $\{p_{ew} | e \in E, w \in W\}$

3 Datasets

Samples. The study uses metatranscriptome data from 26 samples (see Table 1) collected from surface water (depths of 2 and 18 m) on the Louisiana Shelf in the Gulf of Mexico. These samples were collected via Niskin water at the same site (28.867N -90.476W) over a 3-day period in July 2015. Furthermore, six environmental parameters - including PAR (photosynthetic active radiation) and seawater dissolved oxygen concentration, density, salinity, temperature, and chlorophyll concentration - were measured for each sample. One liter of seawater was pumped onto a 0.22 um Sterivex filter. Filtered biomass was then preserved using 1.8 ml of RNA-later and flash frozen. Filters were stored at -80 C until RNA extraction. RNA was extracted via the mirVana™ Total RNA Isolation kit, with residual DNA removed via DNase treatment. RNA samples were then sequenced via the Illumina HiSeq 2500 1TB sequencing protocol following cDNA preparation at the Department of Energy Joint Genome Institute (DOE-JGI). All datasets are publicly available through the JGI Genomes Online (GOLD) database via GOLD ID Gs0110190. Out of 26 samples (see Table 1) three samples (Day1, 12:00, 18m; Day 2, 20:00, 2m; Day 3, 08:00, 2m) were discarded as they did not contain enough reads to assemble transcripts for our pipeline

Samples						
Depth	18 meters			2 meters		
Time \ Day	Day 1	Day 2	Day 3	Day 1	Day 2	Day 3
00:00		✓	✓		✓	✓
04:00		✓	✓		✓	✓
08:00		✓	✓		✓	✗
12:00	✗	✓	✓	✓	✓	✓
16:00	✓	✓		✓	✓	
20:00	✓	✓		✓	✗	

Table 1: The 26 RNA-seq samples of microbial communities drawn from the Northern Louisiana Shelf during contrasting light and dark conditions during 3 consecutive days at two depths 2m and 18m. The x's denote samples which were dropped because of very small number of reads.

Microbial-specific metabolic pathway identification The KEGG pathway database has information on all metabolic pathways that occur in the living organisms. However, the scope of the current tool is to analyse metabolic pathways in microbial communities. We extracted metabolic pathways that play a significant role in microbial communities which is confirmed by literature referenced in PUBMED. Furthermore, we remove from consideration the high-level

metabolic pathways including ec01100, ec01110, ec01120, and ec01130. As a result, we extracted 69 microorganism-relevant pathways out of 152 metabolic pathways. The reduced number of pathways increased the efficiency and performance of the algorithm.

Metabolic enzyme dataset identification and modification We restrict ourselves to enzymes that belong to microbial metabolic pathways and remove the unlikely enzyme matches. Since the same set of contigs assembled from reads can match multiple metabolic enzymes, the EM for enzyme expression cannot differentiate between them. Therefore, we identified the enzymes sharing the same set of contigs and grouped them. For detecting such groups of enzymes, we use an essential property that the individual enzyme expression can vary across randomly initialized EM runs, while the sum of the expression of all enzymes in the group does not change. We collapsed the enzymes belonging to a single group and rerun EM to get an accurate and stable enzyme expression. After applying the above method, we obtain expressions of 1446 enzymes and enzyme groups for the metabolic pathway activity analysis.

4 Results

Our results consist of empirical and statistical validation of estimated enzyme expression, enzyme participation levels, and pathway activity level estimations. We first analyze the stability of enzyme participation levels and list persistently active metabolic pathways. Then we check how many enzyme expressions and pathway activities correlate with environmental parameters. Finally, we verify whether the enzyme expression and pathway activity agrees with the 24-hour cycle.

ec00020	D1:12	D1:16	D1:20	D2:00	D2:04	D2:08	D2:12	D2:16	D3:00	D3:04	D3:12	AVE	STDEV
EC:1.2.4.1	12.82	21.68	20.64	33.71	35.76	30.38	21.78	23.71	32.40	28.07	21.98	25.72	6.60
EC:1.2.7.1	0.51	6.18	15.43	6.69	4.97	9.32	13.14	9.61	7.87	12.95	2.54	8.11	4.37
EC:1.2.7.3	13.99	21.46	20.32	26.74	28.96	24.87	21.26	22.22	27.08	24.44	26.70	23.46	4.02
EC:1.8.1.4	7.61	12.92	11.24	16.94	16.65	14.39	12.93	16.92	19.16	14.03	22.16	15.00	3.78
EC:2.3.1.12	12.82	21.68	20.64	33.71	35.76	30.38	21.78	23.71	32.40	28.07	21.98	25.72	6.60
EC:4.1.1.32	12.82	21.68	20.64	33.71	35.76	30.38	21.78	23.71	32.40	28.07	21.98	25.72	6.60
EC:4.1.1.49	14.78	23.66	23.38	32.19	36.13	37.34	26.62	28.41	35.90	33.66	25.61	28.88	6.60
EC:1.1.1.37	18.14	19.76	26.62	17.90	18.93	30.78	20.27	20.43	22.97	22.13	44.21	23.83	7.43
EC:1.1.1.41	72.88	72.85	70.78	71.20	68.42	38.66	45.68	60.11	62.77	61.29	27.09	59.25	14.74
EC:1.1.1.42	19.96	24.06	22.58	21.52	23.68	19.95	22.48	22.32	22.95	21.92	42.38	23.98	5.95
EC:1.1.5.4	0.00	0.00	0.00	29.35	0.00	0.00	0.00	20.53	0.00	0.00	0.00	24.94	4.41
EC:1.2.4.2	10.10	13.02	10.76	11.91	10.91	11.72	12.75	14.08	14.74	10.13	25.75	13.26	4.21
EC:1.3.5.1	21.35	27.74	28.74	34.65	39.51	30.74	29.40	29.56	36.38	33.32	46.73	32.56	6.43
EC:2.3.1.61	10.10	13.02	10.76	11.91	10.91	11.72	12.75	14.08	14.74	10.13	25.75	13.26	4.21
EC:2.3.3.1	86.31	41.26	66.16	28.14	39.20	260.41	208.96	93.27	70.39	107.86	96.40	99.85	68.92
EC:2.3.3.8	19.96	24.06	22.58	21.52	23.68	19.95	22.48	22.32	22.95	21.92	42.38	23.98	5.95
EC:4.2.1.2	14.54	18.81	19.68	23.77	28.00	20.30	19.67	20.16	24.74	22.70	32.79	22.29	4.72
EC:4.2.1.3	33.31	29.83	34.13	23.43	28.96	41.10	44.43	37.46	35.39	38.11	69.02	37.74	11.35
EC:6.2.1.4	19.96	24.06	22.58	21.52	23.68	19.95	22.48	22.32	22.95	21.92	42.38	23.98	5.95
EC:6.4.1.1	14.54	18.81	19.68	23.77	28.00	20.30	19.67	20.16	24.74	22.70	32.79	22.29	4.72

Table 2: Enzyme participation levels for all enzymes across all data points for 2m depth in the metabolic pathway ec00020. Two rightmost columns are means and standard deviations of enzyme participation levels.

Enzyme Participation Levels. We estimate the participation level of each enzyme in each pathway separately for each data point. Table 2 presents the participation level of all expressed enzymes in the pathway ec00020. We can see that the participation level does not significantly change from one data point to another, i.e.. the standard deviation is significantly smaller than the mean for all enzymes. Note that if an enzyme is not expressed in a sample, then the

participation is not defined and the participation level is reported as 0. This means that we need to take in account only data points with non-zero participation levels when computing mean and standard deviation over all data points.

Persistently Active Metabolic Pathways. Amino acids are precursors of the synthesis of many metabolites inside the cell aids in growth and other biological processes. Amino acid biosynthesis is strongly associated with carbon, nitrogen, and sulfur metabolism. As expected and consistent with the literature in the microbial community, amino acid metabolism pathways lysine, phenylalanine, tyrosine, and tryptophan biosynthesis pathways are persistently enriched across the samples and appeared among the top five in our result [3, 17, 15, 18].

Correlation with Environmental Parameters. The goal of regression-based validation is to check our hypothesis that the expression of a large number of enzymes and the activity levels of many metabolic pathways correlate with each environmental parameter. For each environmental parameter, we check whether it significantly correlates ($P < 5\%$) with each enzyme across 11 data points (see Table 3, 2m). Since the upper bound of 95% CI for salinity is 190 (row 2), we conclude that there is no evidence of enzymes significantly correlated with salinity. We also report the enzyme that correlates the most with salinity, i.e. EC 1.2.1.59. From Table 3 we see that most parameters do not correlate well with enzymes, except perhaps PAR.

Table 4 is the same as Table 3 but reports correlation significance of pathway activities instead of enzyme expressions. In contrast to enzymes it is clear that the many metabolic pathways correlate with each environmental parameter and this correlation is not by chance. Indeed, pathway activity is supposed to be more stable than enzyme expression since generally metabolism is much less affected by the current. For each environmental parameter, we also cross-check the PubMed database whether the most correlated pathway is known to depend on this parameter. For instance, fatty acid degradation is well correlated with salinity, and several studies reported that fatty acid degradation is often altered by salinity at sea surface environments [7, 10, 4].

Environmental Parameter	Salinity	Temp	Oxygen	Chl	PAR	Density	Multiple
1. # of significantly correlated enzymes	146	110	117	93	97	138	156
2. # of randomized correlated (95% CI)	(80-190)	(79-114)	(62-94)	(58-92)	(36-63)	(82-123)	(70-107)
3. The most correlated enzyme	EC1.2.1.59	EC2.6.1.1	EC3.1.3.11	EC 2.2.1.7	EC 3.5.1.16	EC 2.4.1.16	EC 1.1.1.136

Table 3: 1. The number of enzymes significantly correlated with each of 6 environmental parameters and correlated via multiple linear regression. 2. The number of enzymes strongly correlated with randomly permuted parameter values (95% CI). 3. The ID of the metabolic enzyme which is the most strongly correlated with the corresponding parameter.

Environmental Parameter	Salinity	Temp	Oxygen	Chl	PAR	Density	Multiple
1. # of significantly correlated pathways	31	22	19	18	14	30	22
2. # of randomized correlated (95% CI)	(1-8)	(0-8)	(0-6)	(0-6)	(0-6)	(1-8)	(0-7)
3. The most correlated pathway	ec00071	ec00195	ec00622	ec00460	ec00360	ec00071	ec00626

Table 4: 1. The number of pathways significantly correlated with each of 6 environmental parameters and correlated via multiple linear regression. 2. The number of pathways strongly correlated with randomly permuted parameter values (95% CI). 3. The ID of the metabolic pathway which is the most strongly correlated with the corresponding parameter.

24-hours Cycle of Enzyme Expressions and Pathway Activity Levels. We hypothesize that we will be able to observe the cyclic changes in enzyme expression and pathway activity level during 24 hours from 00:00 am on day 2 until 00:00 am on day 3. The cyclic changes should manifest themselves as a higher similarity between two midnight

data sample which are 24h apart than the similarity between two data samples that are 12h apart. We measure similarity between two data points by the correlation between all enzyme expressions and all pathway activity levels estimated for these two data points. Our correlation analysis confirms that enzymes and pathways activity agree with the 24-hour cycle (see Fig.1). We conclude, correlations between enzymes expression are less than pathways activity level.Also correlations between night and day (12h) and day and night are less than correlation between same time points at nights (24h).

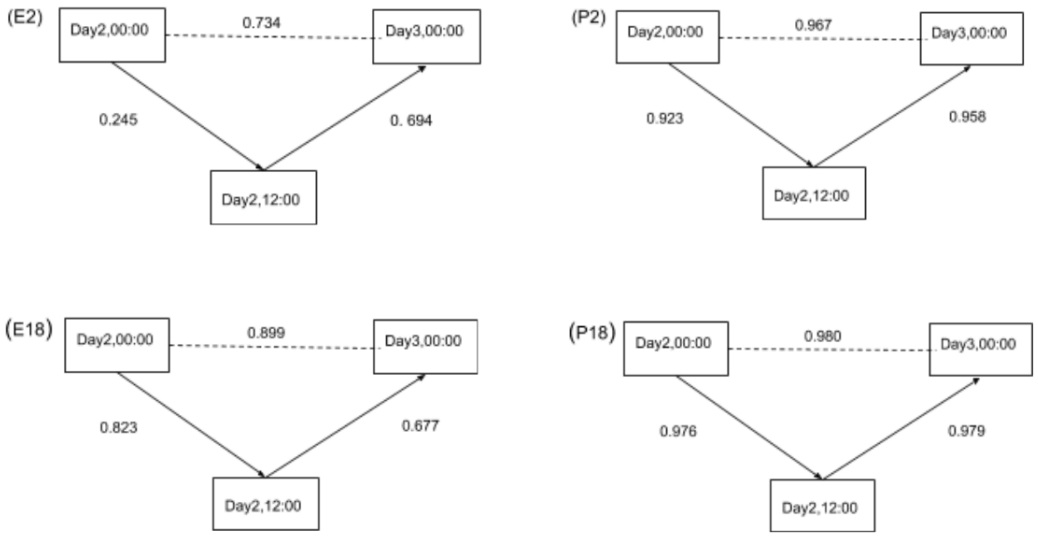

Fig. 2: (E2) correlation between enzyme expressions for 3 samples at 2m-depth: midnight (00:00) of day 2 and day 3 and noon (12:00) of day 2. (P2) correlation between pathway activity levels for 3 samples at 2m-depth: midnight (00:00) of day 2 and day 3 and noon (12:00) of day 2. (E18)correlation between enzyme expressions for 3 samples at 18m-depth: midnight (00:00) of day 2 and day 3 and noon (12:00) of day 2. (P18)correlation between pathway activity levels for 3 samples at 18m-depth: midnight (00:00) of day 2 and day 3 and noon (12:00) of day 2.

5 Discussion

This paper proposes a maximum likelihood model for the estimation of metabolic pathway activity in the microbial community using the KEGG pathway database. Specifically, the proposed model uses an advanced EM-based pipeline to estimate enzyme expression, enzyme participation levels in pathways, and metabolic pathway activity from metatranscriptomic data. The proposed metabolic pathway analysis was applied to the metatranscriptomic data of 26 samples collected with different environmental parameters. The key findings of the study are as follows:

- The participation levels of enzymes in pathways do not significantly vary across the data samples.
- Amino acid metabolism pathways activity such as lysine, phenylalanine, tyrosine, and tryptophan biosynthesis pathways consistently active across all the microbial community samples.
- The enzyme expression and metabolic pathway activities were validated using regression with each environmental parameter: salinity, temperature, oxygen, chlorophyll, and PAR. In contrast to enzyme expressions, pathway activity levels significantly correlate with environmental parameters, e.g. 31 out of 61 metabolic pathways significantly correlate with salinity.
- Enzyme expressions and pathway activity levels agree with the 24-hour cycle.

References

1. Sahar Al Seesi, Yvette Temate Tiagueu, Alexander Zelikovsky, and Ion I Măndoiu. Bootstrap-based differential gene expression analysis for RNA-Seq data with and without replicates. *BMC Genomics*, 15 Suppl 8:S2, November 2014.
2. Nicolas L Bray, Harold Pimentel, Páll Melsted, and Lior Pachter. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.*, 34(5):525–527, May 2016.
3. Mariusz Bromke. Amino acid biosynthesis pathways in diatoms, 2013.
4. Carla de Carvalho, Carla de Carvalho, and Maria Caramujo. The various roles of fatty acids, 2018.
5. Michele Donato, Zhonghui Xu, Alin Tomoiaga, James G Granneman, Robert G Mackenzie, Riyue Bao, Nandor Gabor Than, Peter H Westfall, Roberto Romero, and Sorin Draghici. Analysis and correction of crosstalk effects in pathway analysis. *Genome Res.*, 23(11):1885–1893, November 2013.
6. Bradley Efron and Robert Tibshirani. On testing the significance of sets of genes, 2007.
7. Sandra M Heinzelmann, David Chivall, Daniela M'Boule, Danielle Sinke-Schoen, Laura Villanueva, Jaap S Sinninghe Damsté, Stefan Schouten, and Marcel T J van der Meer. Comparison of the effect of salinity on the D/H ratio of fatty acids of heterotrophic and photoautotrophic microorganisms, 2015.
8. Daniel H Huson, Suparna Mitra, Hans-Joachim Ruscheweyh, Nico Weber, and Stephan C Schuster. Integrative analysis of environmental sequences using MEGAN4. *Genome Res.*, 21(9):1552–1560, September 2011.
9. M Kanehisa. KEGG: Kyoto encyclopedia of genes and genomes, 2000.
10. J Z Kaye. Halomonas neptunia sp. nov., halomonas sulfidaeris sp. nov., halomonas axialensis sp. nov. and halomonas hydrothermalis sp. nov.: halophilic bacteria isolated from deep-sea hydrothermal-vent environments, 2004.
11. Kishori M Konwar, Niels W Hanson, Antoine P Pagé, and Steven J Hallam. MetaPathways: a modular pipeline for constructing pathway/genome databases from environmental sequence information. *BMC Bioinformatics*, 14:202, June 2013.
12. Igor Mandric, Sergey Knyazev, Cory Padilla, Frank Stewart, Ion I. Măndoiu, and Alex Zelikovsky. Metabolic analysis of metatranscriptomic data from planktonic communities. In Zhipeng Cai, Ovidiu Daescu, and Min Li, editors, *Bioinformatics Research and Applications*, pages 396–402, Cham, 2017. Springer International Publishing.
13. Igor Mandric, Sergey Knyazev, Cory Padilla, Frank Stewart, Ion I Măndoiu, and Alex Zelikovsky. Metabolic analysis of metatranscriptomic data from planktonic communities. In *Bioinformatics Research and Applications*, pages 396–402. Springer International Publishing, 2017.
14. Igor Mandric, Yvette Temate-Tiagueu, Tatiana Shcheglova, Sahar Al Seesi, Alex Zelikovsky, and Ion I Mandoiu. Fast bootstrapping-based estimation of confidence intervals of expression levels and differential expression from RNA-Seq data. *Bioinformatics*, 33(20):3302–3304, October 2017.
15. V Martin-Jézéquel, S A Poulet, R P Harris, J Moal, and J F Samain. Interspecific and intraspecific composition and variation of free amino acids in marine phytoplankton, 1988.
16. Cristina Mitrea, Zeinab Taghavi, Behzad Bokanizad, Samer Hanoudi, Rebecca Tagett, Michele Donato, Călin Voichi ta, and Sorin Drăghici. Methods and approaches in the topology-based analysis of biological pathways, 2013.
17. Amanda C Northrop, Rachel K Brooks, Aaron M Ellison, Nicholas J Gotelli, and Bryan A Ballif. Environmental proteomics reveals taxonomic and functional changes in an enriched aquatic ecosystem, 2017.
18. B Palenik and F M M Morel. Amino acid utilization by marine phytoplankton: A novel mechanism, 1990.
19. Itai Sharon, Sivan Bercovici, Ron Y Pinter, and Tomer Shlomi. Pathway-based functional analysis of metagenomes. *J. Comput. Biol.*, 18(3):495–505, March 2011.
20. Mengyuan Shen, Qi Li, Minglei Ren, Yan Lin, Juanping Wang, Li Chen, Tao Li, and Jindong Zhao. Trophic status is associated with community structure and metabolic potential of planktonic microbiota in plateau lakes. *Front. Microbiol.*, 10:2560, November 2019.
21. Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, and Jill P Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.*, 102(43):15545–15550, October 2005.
22. Adi Laurentiu Tarca, Sorin Draghici, Gaurav Bhatti, and Roberto Romero. Down-weighting overlapping genes improves gene set analysis. *BMC Bioinformatics*, 13:136, June 2012.
23. Yuzhen Ye and Thomas G Doak. A parsimony approach to biological pathway reconstruction/inference for genomes and metagenomes. *PLoS Comput. Biol.*, 5(8):e1000465, August 2009.