Predicted Relative Metabolomic Turnover (PRMT): determining metabolic turnover from a coastal marine metagenomic dataset

Peter E. Larsen[1*], Frank R. Collart[1], Dawn Field[2], Folker Meyer[1], Kevin P. Keegan[1], Christopher S. Henry[1,4],John McGrath[3], John Quinn[3], Jack A. Gilbert[1,5]

[1] Argonne National Laboratory, 9700, S. Cass Ave, Argonne, Illinois, USA
[2] NERC Centre for Ecology and Hydrology, CEH Oxford, Mansfield Road, Oxford, OX1 3SR, UK
[3] School of Biological Science, Queens University, Medical Biology Centre, 97 Lisburn Road. Belfast, BT9 7BL, Northern Ireland, UK
[4] Computation Institute, University of Chicago, 5735 South Ellis Avenue, Chicago, IL 60637, U.S.A
[5] Department of Ecology and Evolution, University of Chicago, 5640 South Ellis Avenue, Chicago, IL 60637, U.S.A

* Corresponding author

Email	PEL:	plarsen@anl.gov
	FRC:	fcollart@anl.goc
	DF:	dfield@ceh.ac.uk
	FM:	folker@mcs.anl.gov
	KPK:	kkeegan@anl.gov
	CSH:	chrisshenry@gmail.com
	JM:	j.mcgrath@qub.ac.uk
	JQ:	j.quinn@qub.ac.uk
	JAG:	gilbertjack@gmail.com

ABSTRACT

Background: The world's oceans are home to a diverse array of microbial life whose metabolic activity helps to drive the earth's biogeochemical cycles. Metagenomic analysis has revolutionized our access to these communities, providing a system-scale perspective of microbial community interactions. However, while metagenome sequencing can provide useful estimates of the relative change in abundance of specific genes and taxa between environments or over time, this does not investigate the relative changes in the production or consumption of different metabolites.

Results: We propose a methodology, Predicted Relative Metabolic Turnover (PRMT) that defines and enables exploration of metabolite-space inferred from the metagenome. Our analysis of metagenomic data from a time–series study in the Western English Channel demonstrated considerable correlations between predicted relative metabolic turnover and seasonal changes in abundance of measured environmental parameters as well as with observed seasonal changes in bacterial population structure.

Conclusions: The PRMT method was successfully applied to metagenomic data to explore the Western English Channel microbial metabalome to generate specific, biologically testable hypotheses. Generated hypotheses linked organic phosphate utilization to *Gammaproteobactaria, Plantcomycetes*, and *Betaproteobacteria*, chitin degradation to *Actinomycetes*, and potential small molecule biosynthesis pathways for *Lentisphaerae*, *Chlamydiae*, and *Crenarchaeota*. The PRMT method can be applied as a general tool for the analysis of additional metagenomic or transcriptomic datasets.

BACKGROUND

Marine biomes dominate the planet's surface and single-celled microorganisms are responsible for up to 98% of the ocean's primary productivity [1]; understanding the nutrient and carbon cycles of the world's oceans has key applications for understanding global ecology. The extremely diverse marine microbial communities mediate the largest active pool of near-surface carbon on the planet [2] and are a dominant force in the planet's biogeochemical cycles [3]. The L4 Station of the Western Channel Observatory (WCO), an oceanographic time-series and marine biodiversity reference site in the Western English Channel (http://www.westernchannelobservatory.org.uk), provides a unique opportunity to study a coastal marine microbial ecosystem. Environmental parameter data from the WCO have been continuously monitored for over a century. More recently, microbial metagenomic data collected from this site have shown that the abundance and relative composition of genes and taxa change over time, demonstrating seasonal structure and predictable community responses to environmental parameters [4, 5, 6, 7].

The seasonal structure in the community composition of both taxa and genes has potential repercussions for the seasonal succession in metabolic potential, which will drive the range and relative abundance of metabolites produced and consumed by a community. Metagenomic analyses explore the functional potential of an ecosystem by describing the changes in the abundance of genes annotated with unique enzyme functions. Here we propose a methodology that alters this paradigm, by describing the metagenome in terms of the relative change in the production or consumption of specific metabolites. Instead of exploring gene-space using an environmental metagenomic analysis, this study explores derived metabolite-space, inferred from the metagenome. The predicted environmental metabolome is the set of all detected unique enzyme functions encoded in a metagenome and all of the metabolites implied by those activities. This network of predicted metabolic reactions represents the theoretical metabolic potential of an environmental metabolome and provides a novel means through which metagenomic observations of seasonal and/or biogeographic trends in microbial communities can be utilized to explore community-wide metabolic dynamics.

Understanding the metabolomic interactions in any marine microbial community is a daunting task. Classical metabolomic techniques such as NMR or GC-MS [8, 9, 10, 11], while powerful, provide measurements for just a fraction of the metabolites predicted to be present in a metabolome. As it is not possible to easily measure every metabolite in an environment, it is important to determine the most pertinent parameters, those that will allow investigators to generate testable biological hypotheses. Currently, methods are available to extract a wide variety of environmental features from metagenomic sequence data [12, 13, 14]. Metagenomics can be used to determine the taxonomic and functional diversity of a microbial community via automated pipelines [15] to curated protein databases such as RefSeq [16], KEGG [17], KEGGnoggs [18], SEED [19], PFAM [20] or TIGRfam [21], and linking environmental conditions with specific metabolic activities inferred from metagenomic data [22].

Here, we propose a novel computational method, Predicted Relative Metabolic Turnover (PRMT), which enables comparative analyses of environmental metabolomes predicted from metagenomic data. To support the methodology with specific biological observations, PRMT was applied to a metagenomic dataset from the time-series study of the bacterial environmental metabolome in the Western English Channel [4]. The PRMT approach also correctly predicted considerable correlations between environmental metabolite concentrations and PRMT metabolomic predictions. It also correctly predicting seasonal variations in the bacterial community primary productivity, and generated specific, testable biological hypotheses for organic phosphate utilization by *Gammaproteobactaria, Plantcomycetes*, and *Betaproteobacteria*; chitin degradation by *Actinomycetes*; and potential small molecule biosynthesis pathways for *Lentisphaerae*, *Chlamydiae*, and *Crenarchaeota*.

MATERIALS AND METHODS

Data used for validation of methodology

Metagenomic sequencing data was acquired from a recent study examining the seasonal structure of functional potential in the Western English Channel [4]. Specific data points for use in this study came from day/night pairs of samples taken on January 28, April 27, and August 27 2008. All data collected on the same day were averaged; previous analyses of these data have shown that the metagenomic functional gene profile is statistically identical between day and night in each of these samples [4]. All metagenomic data were annotated with MG-RAST [15] using parameters previously described [4]. Specifically, nucleic acid sequences were excluded if annotated as rRNA, and all subsequent reads were annotated against the SEED database using MG-RAST (e-value <1×10−3; minimum length of alignment of 50 bp; minimum sequence nucleotide identity of 50%; [15]) to produce an abundance matrix of functional genes and protein-derived predicted taxonomies across the DNA samples. Short read data are available through the European Nucleotide Archive (ENA) short read archive under ERP000118 (http://www.ebi.ac.uk/ena/data/view/ERP000118). All data are available on the CAMERA website under 'Western Channel Observatory Microbial Metagenomic Study' (http://camera.calit2.net). MG-RAST annotations of metagenomic data were collected under IDs 4445064.3, 444077.3, 4445065.3, 4445066.3, 4445068.3, 4444083.3, 4445069.3, and 4445070 (http://metagenomics.anl.gov/). All submissions conform to the minimum information standards (MIxS) of the Genomic Standards Consortium [23].

PRMT analysis approach

PRMT scores predict the change in turnover of metabolites (defined as the potential for consumption or production) in an environmental metabolome, given the relative abundance of genes for unique enzyme functions detected in different metagenomes. In this manuscript, we use the term "unique enzyme function" to describe a specific annotation applied to an enzyme, i.e. "Phosphotransferases with an alcohol group as acceptor". We use "enzyme reactions" to refer to metabolite transformations catalyzed by an enzyme function, i.e "ATP + D-Glycerate ↔ ADP + 3-Phospho-D-glycerate". A unique enzyme function may catalyze more than one enzyme reaction and an enzyme reaction may be catalyzed by more than one unique enzyme function. A metabolite is a molecular compound that is a reactant or product in an enzyme reaction. In PRMT, a metabolite is never the protein product of a gene in the metagenome.

This method makes a number of assumptions. First, as with many metagenomic analyses, it assumes that relative abundance of genes for a unique enzyme function in metagenomic sequence is proportionate to relative abundance of expressed functional proteins. Second, PRMT assumes the rate of a reaction is proportionate to the amount of enzyme, and not to the concentrations of reactant or product. Finally, PRMT assumes that the marine metabolome can be modeled as a well-mixed reaction, disregarding compartmentalization of metabolites and activities within individual bacteria. All unique enzyme functions annotated to a set of metagenomes are compared to reference databases of enzyme reactions to infer the set of metabolites present. Below we describe the three main steps to calculating PRMT scores (**Figure 1**):

(1) *Generate Environmental Metabolome Matrix (EMM)*. A network, constructed from annotated unique enzyme functions, enzyme reactions and inferred metabolites, is used to generate the predicted environmental metabolome; this network is expressed as a connectivity matrix, the Environmental Metabolomic Matrix (EMM; Figure 1). The resulting EMM matrix has dimensions *n* x *m*, where *n* is the number of predicted metabolites in the environmental metabolome and *m* is the number of unique enzyme functions detected in the set of analyzed metagenomes. The set of Kyoto Encyclopedia of Genes and Genomes (KEGG) [17] metabolic reactions was used to represent the set of possible enzyme reactions in an environmental metabolome and Enzyme Commission (EC) number annotations for enzyme activity from MG-RAST annotations of metagenomic sequences were used to assign unique enzyme functions to the predicted proteins encoded in a metagenome. KEGG reactions used were selected exclusively from KEGG 'Metabolism' pathways. KEGG pathway maps that are not directly related to specific metabolic activities in microorganisms (e.g. Human Diseases, Drug Development, etc.) were not used in generation of metabolomes. A complete list of the 156 KEGG pathways used in construction of EMM is provided as supplementary data [Additional file 1, Table S1]. As is common in network analysis of biochemical systems [24], the KEGG metabolites water, di- and tri-phosphonucleotides, and all ubiquitous cofactors were excluded from the list of possible reactants as non-specific to particular reactions and metabolic processes. If a reaction is identified as reversible in the KEGG database, then both forward and reverse

reactions are included in the EMM. If a particular metabolite transformation is attributed to more than one enzyme activity, then each transformation reaction is incorporated into the EMM. The stoichiometry of each reaction was not considered as the quantity of metabolites was not considered, just the relative turnover of each.

(2) *Generate normalized Enzyme Activity Counts (EACs).* For each unique enzyme function in each metagenome in the EMM, an Enzyme Activity Count (EAC) is determined by the following equation:

$$EAC_i = \log_2(N_{seq,i})$$

EAC_i is the Enzyme Activity Count for enzymatic function i, and $N_{seq,i}$ is the number of sequence reads in the metagenome annotated with a unique enzyme function i. Collected EACs for a set of metagenomes are normalized (nEAC) by quantiles [25]. Quantile normalization is a technique for making distributions from multiple datasets identical in statistical properties. To quantile-normalize the sets of EACs, the EACs are sorted largest to smallest. A reference distribution is made from the sorted lists such that the highest value in all cases becomes the mean of the highest values, the second highest value becomes the mean of the second highest values, and so on. For each set of EACs, nEAC is generated by assigning the distributions of the reference distribution to the observed EAC distribution. The set of nEACs for a metagenome is expressed as a vector of length n, where n is the total number of unique enzyme functions found in the set of metagenomes.

(3) *Calculate PRMT-scores*. A PRMT-score is calculated for each metabolite in the EMM in a metagenome using the following equation:

$$\vec{c}_{x,y}=M\left(\vec{g}_x-\vec{g}_y\right)$$

$\vec{c}_{x,y}$ is a vector of PRMT-scores of length m, where m is the number of metabolic compounds in the EMM. $\vec{g}_x$ and $\vec{g}_y$ are vectors of normalized enzyme gene counts (nEAC) of length n, where n is the number of unique enzyme functions annotated to metagenomic sequences, and *x* and *y* refer to different metagenome datasets. *M* is the EMM, a matrix of dimensions *m* x *n*. Changes in nEAC for a reversible interaction do not change calculated PRMT-scores for metabolites. For analysis of the L4 WCO environmental metabolome, the reference nEAC was the average of EAC's across all samples from January, April, and August.

Interpretation of PRMT scores

PRMT-scores are unit-less values that represent the change of the turnover of a metabolite in a predicted metabolome relative to a reference metabolome. A PRMT-score is calculated for every metabolite in the EMM for a metagenome. The value and sign (positive or negative) of a PRMT-score provides information about a metabolite's relative turnover. Although a thorough interpretation of a PRMT-score requires that it be considered in the greater context of the complete network, it can be broadly interpreted as follows: A positive PRMT score predicts increased metabolic turnover and relatively greater consumption of a metabolite. A negative PRMT-score predicts decreased turnover and relatively greater accumulation of a metabolite. It is important to note that PRMT-scores do not predict net production or consumption of a metabolite.

Correlations between PRMT-scores and environmental parameters

Numerous methods were considered for exploring the correlative relationships between calculated PRMT scores and measured environmental variables. The non-parametric Spearman's rank correlation was rejected because the resulting coefficients were considered too granular to produce appropriate interpretation of dataset with only 3 time points. Specifically, Spearman's rank correlation returned only 6 possible values of rho for all comparisons in this dataset. Hence, while Spearman's rank may well be suited to correlation coefficient calculation from a dataset with more time points, it was inappropriate for the current study. However, as the PRMT scores generated an approximately normal distribution, it was considered absolutely valid to utilize a parametric test for correlation, namely Pearson's Correlation

Coefficient (PCC). The specific biological parameters considered were environmental metabolites ([4], Table 1, chlorophyll A, Total Organic Nitrogen (TON), Total Organic Carbon (TOC), NO2+NO3, Ammonia, and Soluble Reactive Phosphate (SRP)), bacterial phyla percent abundances [7, Table S2] (only the 23 bacterial phyla and class present at a percent abundance of at least 1% of the total community abundance were used in this analysis.), and number of sequences in metagenomic data annotated to SEED hierarchy I subsystem relative abundances ([4], Figure 7A).

To enable correlation with PRMT scores, the environmental parameters, bacterial phyla percent abundances, and SEED subsystem relative abundances were converted to measures of log relative abundance. Log relative abundance of a parameter was calculated using the following equation:

$$\text{Relative\_Abundance}(x) = \log_2(x/\bar{X})$$

x is a measured experimental parameter and $\bar{X}$ is the average for the parameter across all samples.

PCCs were calculated between four different combinations, namely; all measured environmental metabolites and metagenomic reads annotated to SEED subsystems; environmental metabolites and bacterial phyla; environmental metabolites and PRMT-scores; and PRMT-scores and bacterial phyla. It was considered that statistical significance could not be reliably assigned due to the small number of samples used in the analysis (3 seasonal time points). This was of particular concern given the method's reliance on a multiple-test based procedure. In order to address these concerns, and to provide an informal confidence estimate with which to judge each individual PCC, 10,000 randomized re-samplings of the initial data were used to generate a distribution of PCC-scores. Observed PCC-scores that were in the top or bottom 5th percentile of randomized re-sampled were considered to be a strong correlation.

A graphical representation of observed correlations was performed using ‘Cytoscape’ [51] to generate a network in which experimental measurements are represented as nodes and strong correlations between relative abundance of measurements were represented as edges.

RESULTS AND DISCUSSION

Correlation networks between environmental parameters, bacterial phyla, and SEED subsystems

To demonstrate why PRMT demonstrates a significant advance on existing metagenomic analytical tools, e.g. exploring changes in the relative abundance of predicted gene functions, it was necessary to perform correlative network analysis of the relationships between the relative abundance of taxa, annotated genes and measured environmental parameters (**Figure 2**). Although there may be other environmental causal factors at play, this figure identifies correlations between changes in metabolic functions and environmental conditions or bacterial relative abundance. The most abundant taxa in this ecosystem, *Alphaproteobacteria*, had a strong positive correlation (in the 5th percentile of randomized resamples) with the relative abundance of metagenomic reads annotated to SEED subsystems 'carbohydrate metabolism' and 'cell division'. Additionally, 'cell division' had a strong positive correlation with the relative abundance of TON. This suggests a relationship between TON availability and the growth of *Alphaproteobacterial* populations. Additionally, there is a strong positive correlation between the *Roseobacteriales* order of class *Alphaproteobacteria* and the availability of TON; both peak in the summer. The second most abundant phylum, *Bacteroidetes*, had a strong positive correlation with the SEED subsystem 'Phosphorus metabolism', which, as with TON and *Alphaproteobacteria*, could implicate Phosphorus as a limiting nutrient for *Bacteroidetes*. Interestingly, less abundant taxa were more frequently characterized by strong negative correlations with measurements of nutrients in their environment. The relative abundances of taxa *Chlamydiae*, *Crenarchaeota,* and *Epsilonpreoteobacteria* demonstrated a strong negative correlation with TON; *Deferribacteres* and *Fusobacteria* had a strong negative correlation with NO_2+NO_3; the phylum *Actinobacteria* had a strong negative correlation with ammonia. Conversely, rare phyla more frequently had strong positive correlations with SEED subsystems, *e.g.* 'Cofactors', 'vitamins', 'prosthetic groups, and pigments', 'Clustering based subsystems', 'Virulence', and 'Motility and chemotaxis'; further exploration of these may yield additional avenues of discovery.

A strong negative correlation was observed between the relative abundance of the SEED subsystem 'Photosynthesis metabolism' and the concentration of chlorophyll A. This is similar to previous observations [4] that indicate the relative abundance of genes with homology to components of the cyanobacterial photosynthesis pathway peak in winter, while chlorophyll A concentrations, which relate to Eukaryotic phytoplankton abundance, peak in summer. There is also a strong negative correlation between the relative change in concentration of SRP and the SEED subsystem for 'Phosphate metabolism'. This suggests that when there are more genes for phosphate metabolism there is a greater predicted likelihood for SRP consumption. As a caveat for all of these relationships, correlation does not prove causality; they do however, identify relationships that invite further investigation.

The analysis of the relationships between phyla, environmental parameters, and functional genes indicates that changes in taxonomic diversity affect environmental metabolomic functional potential. To complement this gene-centric perspective, we present a metabolite-centric tool, PRMT, which infers metabolic activity from unique enzyme functions.

The Environmental Metabolomic Matrix from the Western English Channel

The predicted L4 Western English Channel metabolome consists of 2281 predicted metabolites and 4152 enzyme reactions for 990 unique enzyme activities (Figure 3). In the EMM, 70% of enzyme reactions are reversible. The largest connected subnetwork in the network has 1257 metabolites and 3197 enzyme reactions. The second largest subnetwork contains 30 predicted metabolites and 51 enzymatic reactions. There are 194 subnetworks that consist of only two metabolites each. Approximately 63% of the enzyme activities in the KEGG global metabolism map (KEGG map01100) are present in the predicted environmental metabolome. Individual unique enzyme activities in map01100 were represented an average of 128 counts per metagenomic dataset. An average of 81% of total unique enzyme counts in each metagenome could be mapped to KEGG pathway map01100. The representation of core metabolic pathways in annotated metagenomes is especially evident within specific sub-groups; for example energy metabolism is well represented, with 95% of citrate cycle (map00020), 80% of glycolysis/gluconeogenesis (map00010), and 92% of photosynthesis metabolism (map 00195) enzyme activities detected. Cell membrane metabolism is detected as present in the predicted environmental metabolome, with 72% of fatty acid metabolism (map00071) and 65% of fatty acid biosynthesis (map00061). For the DNA metabolism, 65% of pyrimidine metabolism (map00240) and 67% of purine metabolism (map00230) are detected. For protein metabolism, 79% of glycine, serine, threonine metabolism (map00260), 89% of valine, leucine, and isoleucine biosynthesis (map00290), and 90% of phenylalanine, tyrosine, and tryptophan biosynthesis (map00400) enzyme activities are represented. For a bacteria-dominated system, 100% coverage of core metabolic KEGG pathways is not anticipated, as the database also contains plant and animal specific reactions.

PRMT scores calculated from the Western English Channel EMM

PRMT scores were calculated for each predicted metabolite in the EMM at each of three seasonal time points (January 28, April 27, and August 27, 2008). Correlations between PRMT-scores and relative abundance of environmental parameters were generated (Table 1). The complete set of PRMT-scores is available as supplemental data (Additional file 2, Table S2). A strong negative correlation between PRMT scores and relative changes in the concentration of an environmental parameter demonstrates that the metabolomic capacity for *synthesis* of a parameter increased when the parameter is at a relatively higher concentration. Conversely, a strong positive correlation occurs if metabolomic capacity for the *consumption* of a metabolite increases when the environmental parameter is at a relatively higher concentration. Only three of the six correlations were considered strong (i.e. in the top or bottom 5th percentile of randomized resamples). However, using a calculation for the cumulative normal distribution

(CND), given the distribution of correlations between relative abundances of environmental parameters and all calculated PRMT-scores, the probability that all six PCC-scores would be between -1 and -0.8 was <1x 10^{-256} (lowest is -0.81; Table 1). This probability indicates that, even though some individual correlations are not strong (based on the given criteria), it is extremely unlikely that all six correlations could all be between -1 and -0.8 by chance, assuming a normal distribution of correlation coefficients.

Correlations between PRMT-scores and the relative abundances of bacterial phyla can be used to generate hypotheses regarding the potential of different taxonomic groups to produce specific metabolites in subnetworks in the EMM. For example, *Actinobacteria* play an important environmental role in the decomposition of cellulose [26]. The relative abundance of *Actinobacteria* had a strong negative correlation with PRMT-scores for the predicted metabolite cellulose. This strong negative correlation indicates that when the *Actinobacteria* are relatively more abundant, the predicted tendency for cellulose to be consumed in the environmental metabolome increases. In another example, the *Planctomycetes* was the only phyla whose change in relative abundance demonstrated a strong positive correlation with PRMT-scores for L-Glutamate. This was expected, as the cell wall scaffolds of most *Plancotomycetes* are made up of glycoproteins rich in glutamate instead of murein, as is common for other bacterial taxa [27]. Of greater interest than investigating the predicted metabolome for individual metabolites is to seek connected metabolic subnetworks in the EMM that have strong correlations with changes in relative abundance of specific taxa. Connected subnetworks imply not just single metabolites, but partial or complete biochemical pathways. These connected subnetworks were generated by identifying all edges in the network that connect pairs of metabolite nodes where both nodes' PRMT-scores significantly correlate with the relative abundance of a specific bacterial phylum (**Figure 3**). The largest connected subnetwork for interactions that correlated with specific phyla was comprised of 28 metabolites and 70 enzyme reactions, and was comprised of metabolic interactions associated with the metabolism of phosphonoacetate. PRMT-scores for metabolites in this largest subnetwork strongly correlated with relative abundances of *Planctomycetes* and *Gammaproteobacteria*. In the set of connected subnetworks were 5 small subnetworks of 5 or 6 nodes, each with PRMT-scores that exclusively correlated to the relative abundance of *Lentisphaerae*. Predicted metabolites in these subnetworks were associated with flavanoids and anthocyanins. Flavonoids have been observed to have antimicrobial effects on marine microorganisms [28] and flavonoid biosynthesis can be hypothesized to either have defense-related functions in *Lentispharae* or else other marine microorganisms possess the capacity to synthesize flavonoids in response to increased relative abundances of *Lentispharae*. Biosynthesis of terpanoids from the precursor Farnesyl pyrophosphate correlated strongly, not only with the relative abundances of *Crenarchaeota* and *Chlamydiae*, but also with the relative change in concentration of TOC. Terpenoids are a functionally diverse set of molecules whose synthesis by marine bacteria has been previously reported [29, 30].

Specific Examples of PRMT Analysis: primary production, phosphonate metabolism, and chitin catabolism

Primary Production: The chemical equation for average photosynthesis [31] provides a framework for analysis of PRMT-scores in the context of primary production:

$$106\ CO_2 + 16\ HNO_3 + H3PO_4 + 122\ H_2O \rightarrow C_{106}H_{263}O_{110}N_{16}P + 138\ O_2$$

We used PRMT-scores for the metabolite 'alpha-amino acids' to represent biomass in the above chemical equation. 'Alpha-amino acids' is a useful proxy for biomass as the single largest fraction of the dry-weight of prokaryote cells is protein [32]. The PRMT-scores for relevant metabolites in this equation indicate that, as expected, when CO_2, SRP, and NO_3 are consumed, more biomass (alpha-amino acids) is synthesized (**Figure 4**).

The PRMT scores for CO_2 were most positive in August (indicating consumption) and most negative in January (indicating synthesis) (**Figure 4**). Also, PRMT scores indicated increased production of alpha amino acids (biomass) in summer, and active consumption in the winter **(Figure 4)**. This matches expectations, as both bacterial and eukaryotic cellular biomasses are higher in summer [7]. When compared to investigating the abundance of sequences with homology to photosynthetic genes, which predict that the photosynthetic apparatus is more abundant in winter [4], PRMT scores reveal another level of potentially more useful information about the productivity of the ecosystem. While generated from the same data, consideration of sequence abundances and PRMT scores provide two unique perspectives. The largest PRMT-score shift in primary production was observed for nitrate. PRMT-scores for nitrite correlated strongly with the observed relative abundance of chlorophyll A. PRMT-scores for orthophosphate and alpha-amino acid all had strong correlations with PRMT-scores for CO_2. As previously noted (**Table 1**), relative abundance of biological measurements of chlorophyll A had a strong correlation with PRMT-scores for chlorophyll A, linking the molecular components for photosynthesis with the PRMT-predicted capacity to synthesize the mechanism of photosynthesis.

One key limiting metabolite for primary productivity-derived biomass is iron [33, 34]. Increased concentrations of Fe^{2+} correlates with increased primary productivity in ocean microbial populations [35, 36, 37]. The PRMT-scores for Fe^{2+} demonstrated a strong positive correlation with PRMT-scores for alpha amino-acids, the EMM proxy for biomass, as well as the relative abundance of measured TON. PRMT analysis identified a sound positive correlation between Fe^{2+} and production of bacterial biomass.

Phosphonoacetate metabolism: Soluble Reactive Phosphorus (SRP) is considered to be a limited nutrient in marine ecosystems, although this is probably not the case in the Western English Channel [38]**.** However, it was interesting that PRMT scores for inorganic phosphate (orthophosphate) were strongly negatively correlated (PCC -0.93) with changes in the relative concentration SRP (**Table 1**). One reason for this discrepancy is that SRP is comprised of both inorganic and organic phosphate compounds. One group of organic phosphate compounds, the phosphonates, were considered recalcitrant to biological life for many years, but recent evidence from the Western English Channel suggests that they are used by a wide variety of bacteria and constitute a large fraction of the available phosphorus pool [38]. Currently, it is not technically feasible to measure specific phosphonate concentrations in marine systems, and this makes computational approaches like PRMT potentially valuable for investigating their relative turnover.

Strikingly, 92% of the enzyme activities in the phosphonate and phosphinate metabolism KEGG pathway (map00440) were represented in the predicted EMM, indicating that the metabolic capacity to utilize phosphonates was present in the metagenomes.

The PRMT-scores for phosphonate compound, 3-Phosphonopyruvate, had strong correlations with PRMT-scores for CO_2, and with biological measurements of chlorophyll A, suggesting a potential relationship between phosphonopyruvate metabolism and primary production. This is not totally unexpected, as phosphonopyruvate is a key intermediate step in the biosynthesis of all known natural phosphonates [39]. Another phosphonate of potential interest is phosphonoacetate, as the presence of phosphonoacetate as a natural product has been suspected but not confirmed [40]. Measured phosphonoacetate was strongly correlated with relative abundances of *Planctomycetes*, *Gamaproteobacteria*, and *Betaproteobacteria*. This insight into metabolic pathways associated with a metabolite difficult to measure in the environment led to a study to screen for a phosphonoacetate-oxidising activity in marine microorganisms; such an activity has recently been identified in cell-extracts of *Roseovarius nubinhibens*, and is now being characterized [unpublished data].

Chitin catabolism: The highly insoluble, nitrogen-containing compound chitin is the most abundant polymer in the ocean [41, 42, 43]. Yet, it is so rapidly degraded to fructose-6-phosphate, acetate, and ammonia that traces of chitin are only found in marine sediments [44]. Marine bacteria were primarily responsible for this rapid turnover, and chitin catabolism in marine microorganisms has been well studied [45, 46, 47, 48]. Therefore, PRMT analysis was used to test the hypothesis that chitin degradation is a likely source of ammonia in the Western English Channel. Chitin is degraded via the following pathway (KEGG amino sugar and nucleotide sugar metabolism pathway (map00520)):

Chitin → Chitobiose → GlcNAc → GlcN → GlcN-6P → Fructose-6P + Acetate + NH_3

These reactions were all represented in the Western English Channel metagenomes. In the KEGG metabolic pathways, Fructose-6P metabolism leads to KEGG pathways for Glycolysis/gluconeogenesis (map00010) and Fructose and mannose metabolism (map00051), where both acetate and ammonia are generated.

When considering PRMT-scores for the metabolites on the pathway for chitin degradation (**Figure 5**), a pattern for chitin catabolism was evident. Relative turnover of chitin was highest in April, concurrent with the spring bloom of chitin-synthesizing diatoms in the Western English Channel [49]. Increased turnover of chitin in spring leads to a predicted relative increase in the synthesis of chitobiose and increased consumption of GlcNac, ultimately yielding increased relative synthesis of chitin's breakdown products, Fructose-6P, acetate, and ammonia. Relative abundance of measured concentrations of ammonia had a strong negative correlation with PRMT-scores for GlcNAc. This indicates that when the environmental capacity for the consumption of GlcNAc was increased, the relative synthesis of ammonia was also increased. This was consistent with the reactants and products in the chemical equation for Figure 5. The PRMT-scores for GlcNAc had a strong positive correlation with the relative abundance of the Actinobacteria, members of which are involved in decomposition of organic materials such as chitin [50]. The PRMT-scores predicted an increase in the turnover of chitin, concurrent with spring diatom bloom and with increase in relative abundance of bacterial Phyla predicted to possess the capacity for chitin degradation. These PRMT scores and taxonomic correlations support the hypothesis that measured environmental concentrations of ammonia can be linked to bacteria catabolism of available large environmental pools of chitin.

CONCLUSIONS

We have presented a novel computational approach, Predictive Relative Metabolic Turnover (PRMT), for making predictions of the relative change in the production or consumption of a metabolite by a microbial community. The set of all PRMT-scores for all the metabolites predicted to be present in an environmental metabolome provides a system-scale representation of the environmental metabolome derived from metagenomic sequence analysis that changes in response to fluctuations in the structure of its bacterial community. The approach was supported by comparing predicted relative turnover of metabolites to relative abundance of the biological measurements for those same metabolites. PRMT also provided the expected relationships between CO_2, Iron, orthophosphate, nitrate, and chlorophyll with marine primary production. Specific, testable, biological hypotheses regarding the utilization of organic phosphorus and chitin were made, some of which are currently being actively investigated. The predictions made by PRMT are consistent with the hypothesis that bacterial population diversity is linked to the metabolic capacity of the community. While we have restricted our analysis to the metagenome, PRMT calculations are equally applicable to metatranscriptomic data. Ongoing environmental monitoring projects such as the Global Ocean Survey, TARA Oceans, Hawaiian Ocean Time Series, Bermudan Ocean Time Series, the Long Term Ecological Research sites, NEON, and the extended application of the Earth Microbiome Project (www.earthmicrobiome.org) are generating vast amounts of metagenomic and experimental metadata that can readily be investigated by further PRMT analyses.

ABBREVIATIONS

Predicted Relative Metabolomic Turnover (PRMT), Environmental Metabolomic Matrix (EMM), Enzyme Activity Count (EAC), Normalized Enzyme Activity Count (nEAC), Total Organic Carbon (TOC), Total Organic Nitrogen (TON), Soluble Reactive/Phosphate (SRP), Pearson Correlation Coefficient (PCC).

COMPETING INTERESTS

The authors declare that they have no competing interests.

AUTHOR CONTRIBUTIONS

Conceived and designed the experiments: PEL, JAG. Performed the experiments: JAG, DF, PEL. Analyzed the data: PEL, JM, JQ, JAG. Contributed reagents/materials/analysis tools: FRC, FM, JAG.

Wrote the paper: PEL, FRC, JAG, KPK, CSH, JM, JQ, DF. All authors read and approved the final manuscript

ACKNOWLEDGEMENTS

The work was funded by the US Department of Energy, Office of Science, Office of Biological and Environmental Research, Biological Systems Science Division under contract DE-AC02-06CH11357 at Argonne, IL. Sequence data was produced with a grant from the Natural Environmental Research Council (NERC - NE/F00138X/1).

REFERENCES

1. Jørgensen BB, Boetius A: **Feast and famine--microbial life in the deep-sea bed.** *Nat Rev Microbiol*, 10:770-81.
2. Buesseler KO, Lamborg CH, Boyd PW, Lam PJ, Trull TW, Bidigare RR, Bishop JK, Casciotti KL, Dehairs F, Elskens M, Honda M, Karl DM, Siegel DA, Silver MW, Steinberg DK, Valdes J, Van Mooy B, Wilson S: **Revisiting carbon flux through the ocean's twilight zone.** *Science* 2008316(5824):567-70.
3. Paul G. Falkowski, Tom Fenchel, and Edward F. Delong: **The Microbial Engines That Drive Earth's Biogeochemical Cycles.** *Science* 23 May 2008: Vol. 320 no. 5879 pp. 1034-1039.
4. Gilbert JA, D Field, P Swift, S Thomas, D Cummings, B Temperton, K Weynberg, S Huse, M Hughes, I Joint, PJ Somerfield, and M Mühling: **The Taxonomic and Functional Diversity of Microbes at a Temperate Coastal Site: A 'Multi-Omic' Study of Seasonal and Diel Temporal Variation.** *PLoS ONE* 2010, 5(11): e15545. doi:10.1371/journal.pone.0015545.
5. Gilbert JA, Meyer F, Joint I, Mühling M, Field D: **Short Metagenome Report: Metagenomes and metatranscriptomes from the L4 long term coastal monitoring station in the Western English Channel.** SIGS. 2010, Vol 3 (2).
6. Mitra S, Gilbert JA, Field D and Huson DH:. **Comparison of multiple metagenomes using phylogenetic networks based on ecological indices**. *ISME J*. 2010, 4, 1236-1242.
7. Gilbert JA, Field D, Swift P, Newbold L, Oliver A, Smyth T, Somerfield P, Huse S, Joint I: **Seasonal succession of microbial communities in the Western English Channel using 16S rDNA-tag pyrosequencing.** *Env. Microb*, 2009. 11(12), 3132–3139.
8. Bundy JG, M. P. Davey, M. R. Viant: **Environmental metabolomics: A critical review and future perspectives.** *Metabolomics* 2009, 5, 3-21.
9. Viant MR: **Metabolomics of aquatic organisms: the new omics on the block**. *Mar. Ecol. Prog. Series* 2007, 332, 301-306.
10. Viant MR: **Recent developments in environmental metabolomics**. *Molecular Biosystems* 2008, 4, 980-986.
11. Lin C-Y, Viant MR and Tjeerdema RS: **Metabolomics: Methodologies and Applications in the Environmental Sciences**. *J. Pestic. Sci*. 2006, 31, 245-251.

12. Wooley JC, Godzik A, Friedberg I: **A Primer on Metagenomics**. *PLoS Comput Biol* 2010, 6(2): e1000667. doi:10.1371/journal.pcbi.1000667.

13. Heidelberg KB, Gilbert JA and Joint I: **Review: The revolution in genomic approaches to describe environmental microbial diversity.** Microb Biotechnol. 2010 September; 3(5): 531–543.

14. Gilbert JA and Dupont C. 2011. Marine Metagenomics. Annual Review of Marine Science. Volume 3: 347-371.

15. Meyer F, D Paarmann, M D'Souza R Olson EM Glass M Kubal T Paczian, A Rodriguez, R Stevens, A Wilke, J Wilkening and RA Edwards: **The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes**. *BMC Bioinformatics* 2008, 9, 386.

16. Pruitt KD, Tatusova, T, Maglott DR: **NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins**. *Nucleic Acids Res* 2007, 35(Database issue):D61-5.

17. Kanehisa M. and Goto S: **KEGG: Kyoto Encyclopedia of Genes and Genomes.** *Nucleic Acids Res*. 2000, 28, 27-30.

18. Muller J, Szklarczyk D, Julien P, Letunic I, Roth A, Kuhn M, Powell S, von Mering C, Doerks T, Jensen LJ, Bork P: **KeggNOG v2.0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations.** *Nucleic Acids Res*. 2010, (Database issue):D190-5.

19. Overbeek R, T Begley, RM Butler,0 Jomuna, V Choudhuri, H Chuang, M Cohoon, V de Crécy-Lagard, N Diaz, T Disz, R Edwards, M Fonstein, ED Frank, S Gerdes, EM Glass, A Goesmann, A Hanson, D Iwata-Reuyl, R Jensen, N Jamshidi, L Krause, M Kubal, N Larsen, B Linke, AC McHardy, F Meyer, H Neuweger, G Olsen, R Olson, A Osterman, V Portnoy, GD Pusch, DA Rodionov, C Rückert, J Steiner, R Stevens, I Thiele, O Vassieva, Y Ye, O Zagnitko, and V Vonstein: **The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes**. *Nucleic Acids Res*. 2005, 17:5691-702.

20. Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, Hotz HR, Ceric G, Forslund K, Eddy SR, Sonnhammer EL, Bateman A: **The Pfam protein families database**. *Nucleic Acids Res*. 2008, (Database issue):D281-8.

21. Selengut JD, Haft DH, Davidsen T, Ganapathy A, Gwinn-Giglio M, Nelson WC, Richter AR, White O: **TIGRFAMs and Genome Properties: tools for the assignment of molecular function and biological process in prokaryotic genomes**. *Nucleic Acids Res*. 2007, 35(Database issue):D260-4.

22. Gianoulis TA, Raes J, Patel PV, Bjornson R, Korbel JO, Letunic I, Yamada T, Paccanaro A, Jensen LJ, Snyder M, Bork P, Gerstein MB: **Quantifying environmental adaptation of metabolic pathways in metagenomics**. *Proc Natl Acad Sci U S A*. 2009, 106(5):1374-9. Epub 2009 Jan 22.

23. Field D, Joint I, Gilbert JA et al: **Towards a richer description of our complete collection of genomes and metagenomes: the "Minimum Information about a Genome Sequence" (MIGS) specification.** *Nat. Biotech*. 2008, 6(5):541-7.

24. Ma HW, Aeng AP: **Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms**. *Bioinformatics* 2003, 19(2), 270-277.

25. Rao Y, Lee Y, Jarjoura D, Ruppert AS, Liu CG, Hsu JC, Hagan JP: **A comparison of normalization techniques for microRNA microarray data.** *Stat Appl Genet Mol Biol*. 2008;7(1):Article22. Epub 2008 Jul 21.

26. Petković H, Cullum J, Hranueli D, Hunter IS, Perić-Concha N, Pigac J, Thamchaipenet A, Vujaklija D, Long PF: **Genetics of Streptomyces rimosus, the oxytetracycline producer**. *Microbiol Mol Biol Rev*. 2006 Sep;70(3):704-28.

27. Glöckner FO, Kube M, Bauer M, Teeling H, Lombardot T, Ludwig W, Gade D, Beck A, Borzym K, Heitmann K, Rabus R, Schlesner H, Amann R, and Reinhardt R: **Complete genome sequence of the marine planctomycete Pirellula sp. strain 1**. *PNAS* 2003, 100:14 8298-8303

28. Sivakumar PM, Prabhawathi V, Doble M: **Antibacterial activity and QSAR of chalcones against biofilm-producing bacteria isolated from marine waters**. *SAR QSAR Environ Res*. 2010 Apr;21(3-4):247-63.
29. Cho JY, Kwon HC, Williams PG, Jensen PR, Fenical W: **Azamerone, A Terpenoid Phthalazinone from a Marine-Derived Bacterium Related to the Genus Streptomyces** *Actinomycetales*. 2006 June 8; 8(12): 2471–2474.
30. Motohashi K, Irie K, Toda T, Matsuo Y, Kasai H, Sue M, Furihata K, Seto H: **Studies on terpenoids produced by actinomycetes. 5-dimethylallylindole-3-carboxylic Acid and A80915G-8"-acid produced by marine-derived Streptomyces sp. MS239**. *J Antibiot (Tokyo).* 2008 Feb;61(2):75-80.
31. Paytan, A and McLaughlin, K: **The Oceanic Phosphorus Cycle**. *Chem. Rev*. 2007, 107, 563-576.
32. Mikhail VS, Fuchs BM, Eilers H, Burkill PH, Amann R: **Determination of Total Protein Content of Bacterial Cells by SYPRO Staining and Flow Cytometry**. *Appl Env Microbiol*. July 1999, p. 3251–3257 Vol. 65, No. 7.
33. Martin, J. H.; Fitzwater, S. E: **Iron-deficiency limits phytoplankton growth in the Northeast Pacific Subarctic**. *Nature* 1988, 331: 341–343.
34. Martin JH, Coale KH, Johnson KS, Fitzwater SE, Gordon RM, Tanner SJ, Hunter CN, Elrod VA, Nowicki JL, Coley TL, Barber RT, Lindley S, Watson AJ, Van Scoy K, Law CS, Liddicoatp MI, Ling R, Stanton T, Stockel J, Collins C, Anderson A, Bidigare R, Ondrusek M, Latasa M, Millerostar FJ, Leestar K, Yao W, Zhangstar JZ, Friederich G, Sakamoto C, Chavez F, Buck K, Kolber Z, Greene R, Falkowski P, Chisholm SW, Hoge F, Swift R, Yungel J, Nightingale SP, Hatton A, Liss P and Tindale NW: **Testing the iron hypothesis in ecosystems of the equatorial Pacific Ocean**. Nature 2002, 371, 123 – 129
35. Martin JH: **Glacial-interglacial CO2 change: The iron hypothesis.** *Paleoceanography* 1990, 5, 1–13.
36. Street JH, Paytan A: **Iron, phytoplankton growth, and the carbon cycle**. *Met Ions Biol Syst*. 2005, 43:153-93.
37. Blain S, et al.: **Effect of natural iron fertilization on carbon sequestration in the Southern Ocean**. *Nature*. 2007, 446, 1070–1074.
38. Gilbert JA, Thomas S, Cooley NA, Kulakova A, Field D, Booth T, McGrath JW, Quinn JP, Joint I: **Potential for Phosphonate utilisation by marine bacteria in temperate coastal waters**. *Environmental microbiology* 2009, 11(1):111-25.
39. Kulakova AN, Wisdom GB, Kulakov LA, Quinn JP: **The purification and characterization of phosphonopyruvate hydrolase, a novel carbon-phosphorus bond cleavage enzyme from Variovorax sp Pal2.** *J Biol Chem*. 2003 Jun 27;278(26):23426-31. Epub 2003 Apr 15.
40. Quinn JP, Kulakova AN, Cooley NA and McGrath JW: **New ways to break an old bond: the bacterial carbon–phosphorus hydrolases and their role in biogeochemical phosphorus cycling.** *Environmental Microbiology* 2007, 9(10), 2392–2400.
41. Aluwihare, LI, Repeta DJ, Pantoja S, and Johnson CG: **Two chemically distinct pools of organic nitrogen accumulate in the ocean**. *Science* 2005, 308:1007–1010.
42. Gooday, GW: **The ecology of chitin degradation**, p. 387–429. In K. C. Marshall (ed.), *Advances in microbial ecology* 1990, vol. 11. Plenum Press, Inc., New York, NY.
43. Jeuniaux C, and Vossfoucart, MF: **Chitin biomass and production in the marine environment**. *Biochem. Syst. Ecol*. 1991, 19:347–356.
44. Keyhani NO, Roseman S. **Physiological aspects of chitin catabolism in marine bacteria**. *Biochim Biophys Acta*. 1999, 1473(1):108-22.
45. Cottrell MT, Kirchman DL: **Natural assemblages of marine proteobacteria and members of the Cytophaga-Flavobacter cluster consuming low- and high-molecular-weight dissolved organic matter**. *Appl Environ Microbiol*. 2000 Apr;66(4):1692-7.

46. Cottrell MT, Wood DN, Yu L, Kirchman DL: **Selected chitinase genes in cultured and uncultured marine bacteria in the alpha- and gamma-subclasses of the proteobacteria**. *Appl Environ Microbiol*. 2000 Mar;66(3):1195-201.
47. Cottrell MT, Moore JA, Kirchman DL: **Chitinases from uncultured marine microorganisms**. *Appl Environ Microbiol*. 1999 Jun;65(6):2553-7.
48. Svitil AL, Kirchman DL: **A chitin-binding domain in a marine bacterial chitinase and other microbial chitinases: implications for the ecology and evolution of 1,4-beta-glycanases**. *Microbiology*. 1998 May;144 (Pt 5):1299-308.
49. Southward AJ, Langmead O, Hardman-Mountford NJ, Aiken J, Boalch GT, Dando PR, Genner MJ, Joint I, Kendall MA, Halliday NC, Harris RP, Leaper R, Mieszkowska N, Pingree RD, Richardson AJ, Sims DW, Smith T, Walne AW, Hawkins SJ: **Long-term oceanographic and ecological research in the Western English Channel**. *Adv Mar Biol*. 2005;47:1-105.
50. Petković H, Cullum J, Hranueli D, Hunter IS, Perić-Concha N, Pigac J, Thamchaipenet A, Vujaklija D, Long PF: **Genetics of Streptomyces rimosus, the oxytetracycline producer**. *Microbiol Mol Biol Rev*. 2006 Sep;70(3):704-28.
51. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: a software environment for integrated models of biomolecular interaction networks**. *Genome Res*. 2003, 13(11):2498-504.

FIGURES

Figure 1.
Example of generating an EMM from metagenomic data. This figure is an example of generating a simple EMM with hypothetical data. Letters a-f represent unique enzyme functions identified in the annotation of a hypothetical set of metagenomes. In (A), the set of all enzyme reactions for enzyme functions a-f between compounds C1-C5 from a database of possible reactions is listed. In (B), a metabolome is constructed from the reactions identified in A. (C) Shows the connectivity matrix of the network in B. (D) Is the complete EMM for metagenomic annotated enzyme functions a-f, normalizing values in C such that the sum of all inputs to a compound is 1 and the sum of all outputs from a compound is -1.

Figure 2.
Strong correlations between environmental metabolites, metabolic subsystems, and bacterial population structure. This network is a graphical representation of strong (i.e. in the top or bottom 5th percentile of randomized resamples) correlations between relative abundance of measured environmental metabolites (diamonds), relative abundance of metagenomic reads annotated to metagenomic SEED subsystems (hexagons), and relative abundance of bacterial taxa (circles) across seasonal variation for the Western English Channel L4 station. Strong positive correlations are represented by solid lines and strong negative correlations by dashed lines.

Figure 3.
L4 Environmental Metabolome. In the figure, edges represent enzyme functions identified in annotated metagenomes. Nodes are predicted metabolites, inferred by the reactions catalyzed by detected enzyme functions. Nodes are highlighted if calculated PRMT scores for seasonal metagenomes correlate strongly (i.e. in the top or bottom 5th percentile of randomized resamples) with relative abundance of measured environmental parameters (Red for Total Organic Carbon, blue for Total Organic Nitrogen, and gold for Soluble Reactive Phosphorus). Edges are highlighted in one of 23 colors if they connect nodes that correlate with relative abundance of a bacterial phylum. Figure was generated using Cytoscape v2.6.1 [51]. The network and calculated PRMT-scores in this figure is available for download as Additional file 3, figure S1.

Figure 4.
PRMT-scores for photosynthesis pathway. PRMT-scores for metabolites along the photosynthesis pathway are grouped by time point of metagenomic sample. In this figure, the marine biomass in equation for average photosynthesis is represented by PRMT-score for alpha-amino acids. A positive PRMT-scores indicate a predicted increase in relative catabolism, and negative PRMT-scores indicate a predicted increase in synthesis of a metabolite.

Figure 5.
PRMT-scores for chitin catabolism subnetwork. PRMT-scores for metabolites along the chitin catabolism pathway are grouped by time point of metagenomic sample. Positive PRMT-scores indicate a predicted increase in relative catabolism, and negative PRMT-scores indicate a predicted increase in synthesis of a metabolite.

TABLES

Table 1. Correlations of calculated PRMT scores with relative abundance of selected environmental parameter measurements.

Parameter	**PRMT metabolite**	**PCC**
Chlorophyll A	Chlorophyll A	**-0.98**
Total Organic Nitrogen	alpha-Amino acid	**-0.99**
Total Organic Carbon	Starch	-0.98
NO_2+NO_3	Nitrite	**-0.98**
NH_3	NH3	-0.81
Soluble Reactive Phosphorus	Orthophosphate	-0.93

Parameters are relative concentrations of biological measurements from Western English Channel L4 station and PRMT-metabolites are those predicted metabolites that are representative of measured parameters. PCC-scores between measured parameters and PRMT-scores are highlighted in **bold** if a strong correlation (i.e. in the top or bottom 5th percentile of randomized resamples) was observed.

Additional file 1, Table S1
This file contains a list of all enzymatic reactions in the predicted environmental metabolome for the Western L4 Station. Each line in the tab-separated file is a metabolic reaction in the format: Reactant, unique enzymatic function, product.

Additional file 2, Table S2
This file contains all calculated PRMT-scores, all PCC-scores between PRMT and relative abundances of Taxa, and between PRMT and relative abundances of measured environmental metabolites for the Western English Channel L4 Station environmental metabolome.

Additional file S3, Figure S1
This file contains a list of all enzymatic reactions in the predicted environmental metabolome for the Western L4 Station. Each line in the tab-separated file is a metabolic reaction in the format: Reactant, unique enzymatic function, product. It is in the format of a 'Cytoscape' .sif file (http://www.cytoscape.org).

A

$$C1 \xrightarrow{a} C2$$

$$C1 \xrightarrow{a} C3$$

$$C3 \xrightarrow{b} C1$$

$$C2 \xrightarrow{c} C3$$

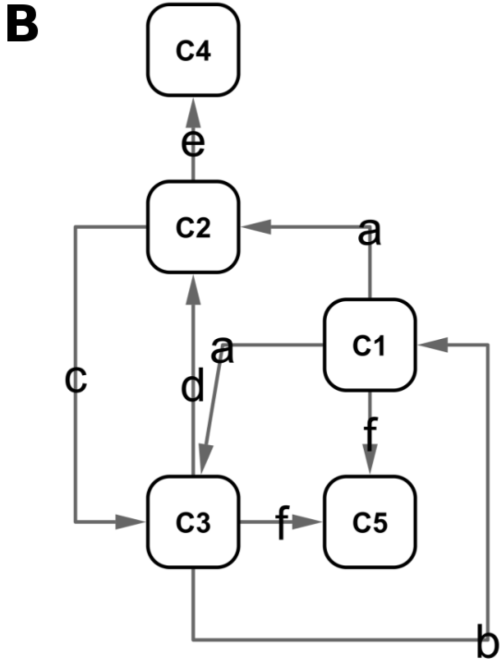
$$C3 \xrightarrow{d} C2$$

$$C2 \xrightarrow{e} C4$$

$$C1 \xrightarrow{f} C5$$

$$C3 \xrightarrow{f} C5$$

B



C

	v_a	v_b	v_c	v_d	v_e	v_f
C1	-2	1	0	0	0	-1
C2	1	0	-1	1	-1	0
C3	1	-1	1	-1	0	-1
C4	0	0	0	0	1	0
C5	0	0	0	0	0	2

D

	v_a	v_b	v_c	v_d	v_e	v_f
C1	-0.66	1	0	0	0	-0.33
C2	0.5	0	-0.5	0.5	-0.5	0
C3	0.5	-0.33	0.5	-0.33	0	-0.33
C4	0	0	0	0	1	0
C5	0	0	0	0	0	1

Figure 1

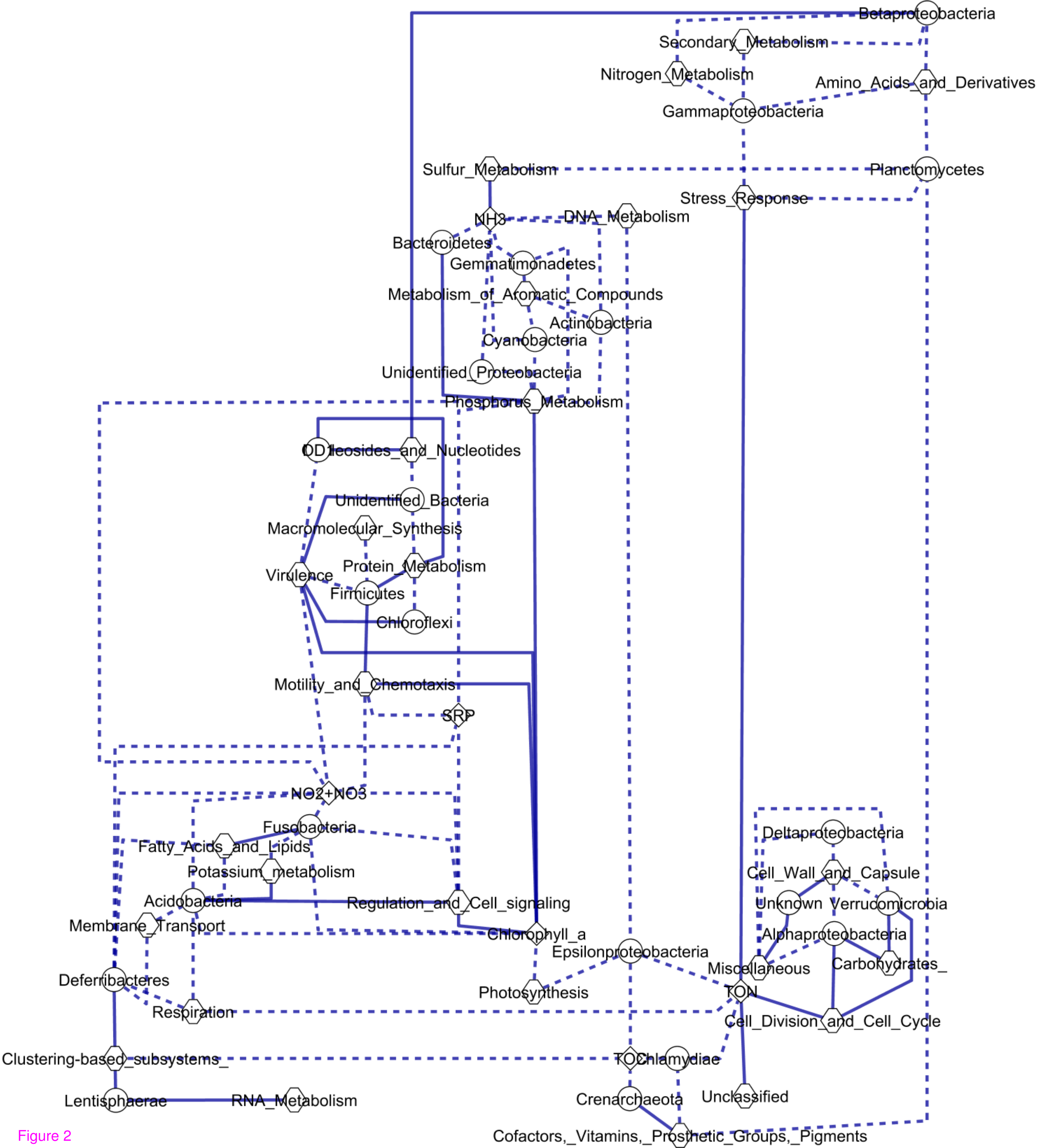

Betaproteobacteria
Secondary_Metabolism
Nitrogen_Metabolism
Amino_Acids_and_Derivatives
Gammaproteobacteria
Sulfur_Metabolism
Planctomycetes
Stress_Response
NH3
DNA_Metabolism
Bacteroidetes
Gemmatimonadetes
Metabolism_of_Aromatic_Compounds
Actinobacteria
Cyanobacteria
Unidentified_Proteobacteria
Phosphorus_Metabolism
DO
Nucleosides_and_Nucleotides
Unidentified_Bacteria
Macromolecular_Synthesis
Virulence
Protein_Metabolism
Firmicutes
Chloroflexi
Motility_and_Chemotaxis
SRP
NO2+NO3
Fusobacteria
Fatty_Acids_and_Lipids
Potassium_metabolism
Acidobacteria
Regulation_and_Cell_signaling
Membrane_Transport
Chlorophyll_a
Deltaproteobacteria
Cell_Wall_and_Capsule
Unknown
Verrucomicrobia
Alphaproteobacteria
Epsilonproteobacteria
Miscellaneous
Carbohydrates_
Deferribacteres
TON
Photosynthesis
Respiration
Cell_Division_and_Cell_Cycle
Clustering-based_subsystems_
TOC
Chlamydiae
Lentisphaerae
RNA_Metabolism
Crenarchaeota
Unclassified
Cofactors,_Vitamins,_Prosthetic_Groups,_Pigments

Figure 2

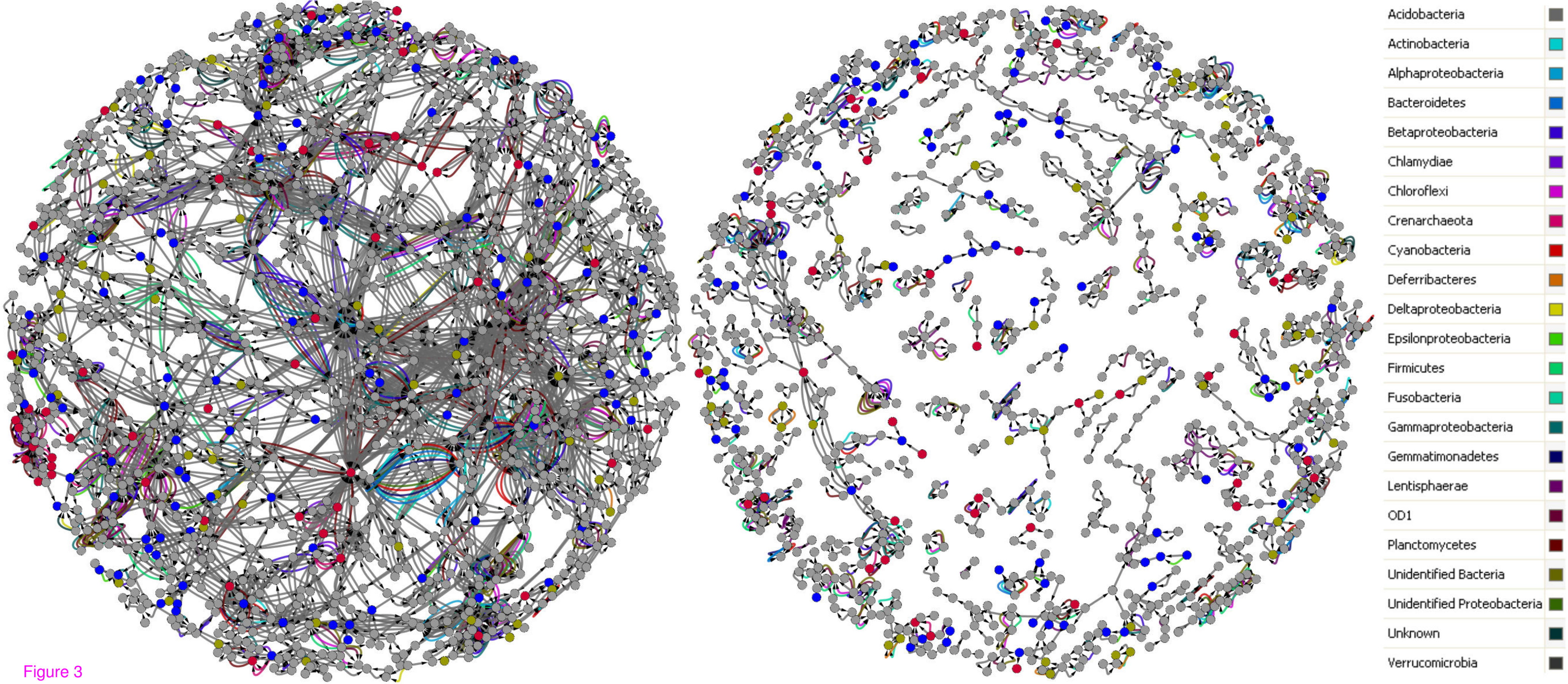

Acidobacteria
Actinobacteria
Alphaproteobacteria
Bacteroidetes
Betaproteobacteria
Chlamydiae
Chloroflexi
Crenarchaeota
Cyanobacteria
Deferribacteres
Deltaproteobacteria
Epsilonproteobacteria
Firmicutes
Fusobacteria
Gammaproteobacteria
Gemmatimonadetes
Lentisphaerae
OD1
Planctomycetes
Unidentified Bacteria
Unidentified Proteobacteria
Unknown
Verrucomicrobia

Figure 3

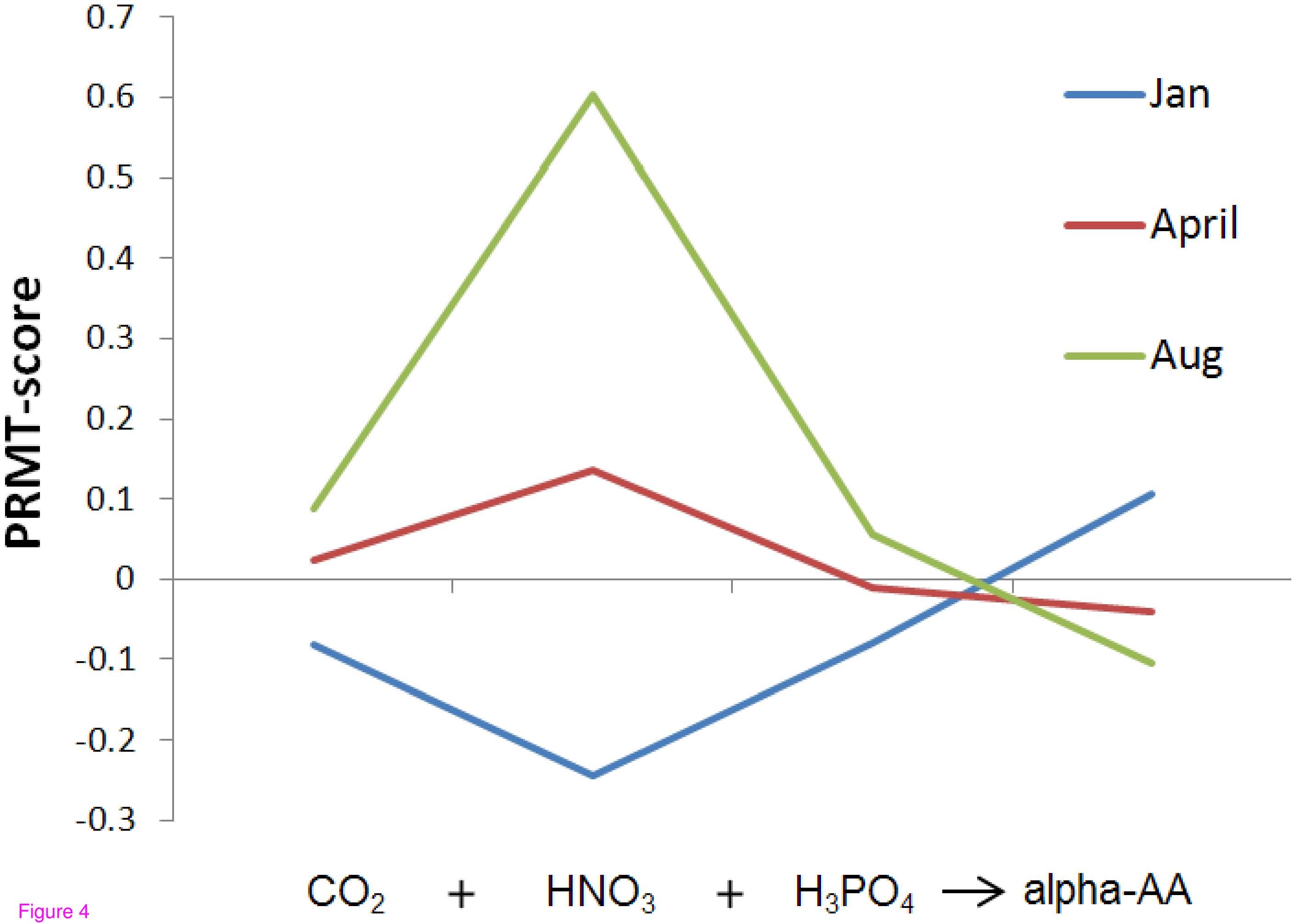

PRMT-score
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
Jan
April
Aug
CO_2 + HNO_3 + H_3PO_4 → alpha-AA

Figure 4

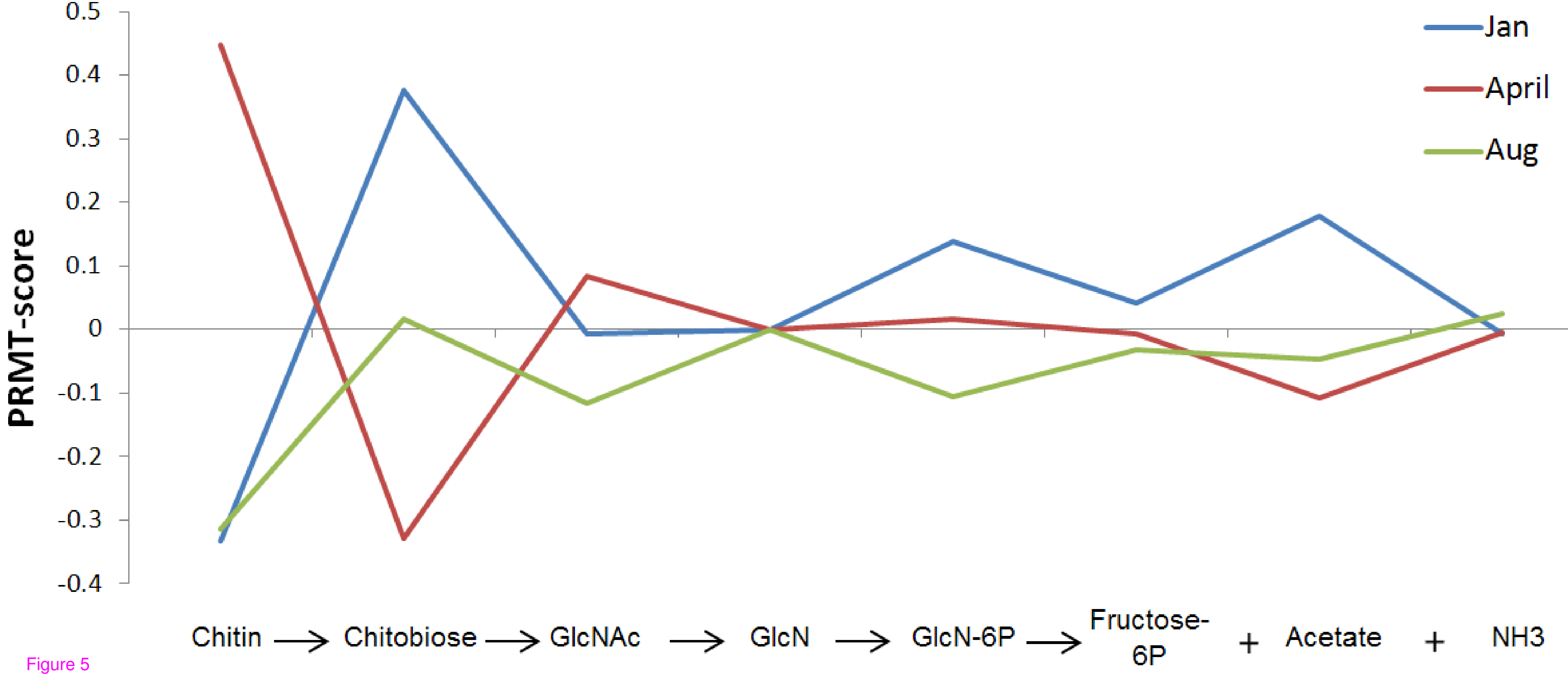

0.5
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
PRMT-score
Jan
April
Aug
Chitin → Chitobiose → GlcNAc → GlcN → GlcN-6P → Fructose-6P + Acetate + NH3

Figure 5

Additional files provided with this submission:

Additional file 1: AdditionalFile1_Table_S1_ListOfKEGGPatwhays.txt, 5K
http://www.microbialinformaticsj.com/imedia/7227940825562799/supp1.txt
Additional file 2: AdditionalFile2_Table_S2_AllPRMT-scores.txt, 783K
http://www.microbialinformaticsj.com/imedia/1673417203556279/supp2.txt
Additional file 3: AdditionalFile3_L4-EnvironmentalMetabolome.sif, 221K
http://www.microbialinformaticsj.com/imedia/2290518556280007/supp3.sif