Quantifying environmental adaptation of metabolic pathways in metagenomics

Tara A. Gianoulis*,†, Jeroen Raes‡, Prianka V. Patel§, Robert Bjornson¶, Jan O. Korbelt**, Ivica Letunic‡, Takuki Yamada‡, Alberto Paccanaro*, Lars J. Jensen‡, Michael Snyder†, Peer Bork‡, and Mark B. Gerstein*.†,‡,§

*Program in Computational Biology and Bioinformatics, Departments of †Molecular Biophysics and Biochemistry, ‡Computer Science, and §Molecular, Cellular, and Developmental Biology, Yale University, New Haven, CT 06520; †Royal Holloway, University of London, Egham, TW20 0EX United Kingdom; ‡Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, DK-2200 Copenhagen, Denmark; §Max Delbrück Center for Molecular Medicine, D-13125 Berlin-Buch, Germany; and †European Molecular Biology Laboratory, D-69117 Heidelberg, Germany

Edited by Michael Levitt, Stanford University School of Medicine, Stanford, CA, and approved November 20, 2008 (received for review August 14, 2008)

Recently, approaches have been developed to sample the genetic content of heterogeneous environments (metagenomics). However, by what means these sequences link distinct environmental conditions with specific biological processes is not well understood. Thus, a major challenge is how the usage of particular pathways and subnetworks reflects the adaptation of microbial communities across environments and habitats—i.e., how network dynamics relates to environmental features. Previous research has treated environments as discrete, somewhat simplified classes (e.g., terrestrial vs. marine), and searched for obvious metabolic differences among them (i.e., treating the analysis as a typical classification problem). However, environmental differences result from combinations of many factors, which often vary only slightly. Therefore, we introduce an approach that employs correlation and regression to relate multiple, continuously varying factors defining an environment to the extent of particular microbial pathways present in a geographic site. Moreover, rather than looking only at individual correlations (one-to-one), we adopted canonical correlation analysis and related techniques to define an ensemble of weighted pathways that maximally covaries with a combination of environmental variables (many-to-many), which we term a metabolic footprint. Applied to available aquatic datasets, we identified footprints predictive of their environment that can potentially be used as biosensors. For example, we show a strong multivariate correlation between the energy-conversion strategies of a community and multiple environmental gradients (e.g., temperature). Moreover, we identified covariation in amino acid transport and cofactor synthesis, suggesting that limiting amounts of cofactor can (partially) explain increased import of amino acids in nutrient-limited conditions.

Environmental genomics | Network dynamics | Microbiology | Canonical correlation analysis

Microbes function as highly interdependent communities. Fundamental to the maintenance of the energy balance of the ecosystem, the recycling of nutrients, and the neutralization and degradation of toxins and other detritus (1), microbial community processes are intimately intertwined with ecosystem functioning. Thus, it is critical to understand the complex interplay between the influence of the environment on microbial communities, and, in turn, the microbes’ reshaping of their environment.

Until recently, the tools to systematically study global community function and environment at the molecular level were not available, because complex microbial communities are generally not amenable to laboratory study (2). The recent advent of direct sequencing of environmental samples (i.e., metagenomics) has allowed the first large-scale insights into the function of these complex microbial communities.

Comparative metagenomics approaches have revealed significant variation in sequence composition (3), genome size (4), evolutionary rates (5), and metabolic capabilities (6–8) among qualitatively dissimilar environments (e.g., terrestrial vs. marine), providing evidence for genomic adaptations. Further, variation in specific community biological processes have been shown for different water column zones at a single geographic site (9), different climatic regions in the ocean (10), and, more recently, among 9 ecosystems (7).

The wealth of information generated from these studies emphasizes the importance of investigating relative differences in biological processes among qualitatively different environments. However, to date, none of them have directly incorporated multiple, specific measurements of the environment. By treating the environment explicitly as a set of complex, continuous features, rather than relying on an implicit subjective classification, one can build models to determine how a diverse array of biochemical activities, and particularly metabolic versatility, reflect sets of or specific environmental differences.

Providing an ideal dataset for exploring these environmental–biochemical links, the Global Ocean Survey (GOS) collected quantitative environmental features and metagenomic sequences from >40 different aquatic sites (10). Here, we used GOS data to investigate and develop multivariate approaches to systematically relate metabolic pathway usage directly to quantitative environmental differences. These approaches allowed us to address multiple relationships simultaneously, as well as to relate specific environmental features to metabolic processes at different levels of resolution, including 14 broad functional categories, 111 pathways, 141 modules (sections of pathways), 191 operons, and 15,554 orthologous groups (OGs). By identifying environmentally-dependent pathways involved in energy conversion, amino acid metabolism, and cofactor synthesis, among others, we were able to define metabolic footprints of distinct environments. Our study provides an analytical framework for uncovering ways, in which microbes adapt to (and perhaps even) how they change their environment.

Results
Quantitative Approach for Footprint Detection. We mapped 37 size-filtered GOS sites (Table S1) to their respective environmental and metabolic features at several levels of complexity.


The authors declare no conflict of interest.

This article is a PNAS Direct Submission. Freely available online through the PNAS open access option.

1T.A.G. and J.R. contributed equally to this work.

2To whom correspondence may be addressed. E-mail: peer.bork@embl.de or mark.gerstein@yale.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0808022106/DCSupplemental.

© 2009 by The National Academy of Sciences of the USA

Fig. 1. Schematic representation of approach. The large squares labeled B1, B2, etc. represent the geographic sites (buckets). Each bucket has sequence and environmental feature data associated with it. (A) Mapping quantitative environmental features [salinity (ppt), sample depth (position in water column from which the sample was collected), water column depth (measured from surface to floor), and chlorophyll]. (B) Metagenome-derived metabolism at different levels (see Materials and Methods). Reads are color-coded according to their corresponding pathway elements (shapes). Different pathways are represented by different shapes (square, circle, etc.). All of the instances of a particular pathway are summed and normalized to compute the pathway score. (C) Schematic representation of CCA (see details in text). (D) Schematic representation of DPM (see details in text).

Fig. 2. Predicting specific environmental parameters from subsets of metabolic pathways. Linear model for temperature built from subsets of highly correlated pathways, including N-acetylglucosamine biosynthesis, many components of amino acid metabolism, and fatty acid biosynthesis (for full list and coefficients, see Table S4). Axes are normalized actual and predicted temperature for x and y, respectively.

(pathways, modules, and operons; see Fig. 1A and B). These data can naturally be represented as matrices, where the rows are geographic sites and the columns are either environmental or metabolic features. We interrelated these matrices to examine how pathway usage across different sites is related to environmental parameters. The simplest and most direct approaches for performing such operations are correlation and regression (for comparisons with other types of methods see Fig. S1 and Table S2). Thus, we examined the first order relationships by computing the pairwise correlation between each metabolic pathway and each environmental feature (e.g., photosynthesis and temperature). Note that for clarity, we use the word pathway to refer to the usage of the pathway, as in photosynthesis as opposed to usage of photosynthesis, in the remainder of the text. This analysis revealed a number of significant correlations (environmentally-dependent pathways; see Table S3). Such pathways were used to build linear models (LM) of each environmental feature (see Table S4). Although these models performed well in predicting single environmental features (Fig. 2), there are limitations to viewing each environmental measurement in isolation, because there are hidden dependencies among the environmental features.

To discover the complex, higher order interactions between and within environmental features and metabolism, we used a second complimentary approach, regularized canonical correlation analysis (CCA). CCA has 2 primary functions: (i) to determine whether a global relationship between 2 types of features (here, environmental and metabolic) exists; and (ii) to calculate the relative contribution of each feature to the global relationship (e.g., temperature or photosynthesis), by weighting both sets of features simultaneously (11). In brief, CCA computes a linear combination for each feature set and simultaneously attempts to maximize the correlation between the 2 feature vectors (Fig. 1D). Thus, CCA is able to simultaneously assess relationships both between and among the environmental features and metabolic pathways. Because the sites are quite similar, we developed a more robust but less sensitive method called discriminative partition matching (DPM). DPM first partitions the sites into site sets on the basis of their environmental parameters, then tests which pathways give the greatest discriminatory power among the site sets (Fig. 1C).

Footprint Characteristics. The goal of DPM and CCA is to simultaneously explore the relationship between metabolism and the quantitative environmental parameters by identifying environmentally-dependent or covarying metabolic pathways (footprints). The main difference between DPM and CCA is that DPM identifies those pathways that discriminate the best between site sets, but when defining the site sets, all of the environmental variables are considered equally important. Thus, although robust to noise, DPM is more coarse-grained, and, at this resolution, the individual differences among sites and their relationships to the environment can be lost. In contrast, CCA can highlight these individual differences by weighting each environmental feature and each metabolic pathway independent of any partitioning, making it both more sensitive but also more susceptible to noise (Fig. 1D).

DPM Footprint. Applying DPM, the sites were partitioned into 2 different site sets that can loosely be classified as open ocean and...
coastal. We found that the distribution of those clusters of orthologous groups of proteins (COGs) and Kyoto Encyclopedia of Genes and Genomes (KEGG) maps annotated as having a role in metabolism was significantly different between site sets ($P < 9 \times 10^{-3}$ and $4 \times 10^{-14}$, respectively); however, no statistically significant difference was found for control matrices that were composed of translational/transcriptional machinery (see SI Materials and Methods).

Also, we find 10 KEGG maps, 24 modules, 61 operons, and 98 gene families were significantly different [false discovery rate (FDR)-corrected $q < 0.05$] between the 2 site sets. These pathways together form the DPM footprints (Table S5). By examining the broader trends of these footprint pathways, we found that secondary metabolite biosynthesis, lipid transport and metabolism, amino acid metabolism, and energy production and conversion were significantly different between site sets. Finally, we showed that the cluster similarity between the environment-based site partitioning and metabolic footprint-based site partitioning was quite high (normalized mutual information, 0.46; rand index, 0.76; $P < 0.001$), suggesting that footprints have predictive power in recapturing features of the environment based purely on pathways identified as significant in DPM.

CCA Footprint. Next, we applied regularized CCA to measure the strength of the metabolic pathway’s covariation with environmental features. We identified 22 KEGG maps, 53 modules, and 35 operons as being environmentally dependent (absolute value of structural correlations >0.3; see Fig. 3). These
pathways form footprints that can be investigated for subtler environment-based changes in metabolic capabilities (Table S6). In this manner, we identified diverse functional processes that covaried significantly with the environment, including xenobiotic degradation, energy conversion, lipid metabolism, and amino acid metabolism.

**Adaptation of Energy Conversion Strategies to Specific Environmental Challenges.** Many of the environmentally-dependent pathways were associated with energy conversion. The diversification in energy-conversion strategies is reasonable given that a primary challenge to all microbial communities is how to maintain adequate energy reserves despite challenging conditions in their specific environment.

Our results demonstrate ample diversification in energy-conversion strategies linked to such quantitative environmental differences. In particular, we show that proteins involved in (photo)autotrophic processes, such as photosynthesis, oxidative phosphorylation, and carbon and nitrogen fixation, are strongly influenced by variation in environmental parameters (Fig. 3). This link is seen at all functional levels and reinforced by multiple method-ologies (Table S7). The module-level analysis showed that only photosynthetic modules involved in light capture and electron transport (photosystem I, II, and the cytochrome b6/f complexes) correlated with the environment. In contrast, the abundance of the module for the ATP synthase complex, whose function is independent of the particular energy conversion strategy, does not change significantly (Fig. 3A). A similar trend can be seen for oxidative phosphorylation, although not as strongly (Fig. 3B). The seeming lack of environmental constraint on the ATP synthases probably reflects their role in coupling energy to a proton gradient (e.g., oxidative phosphorylation, etc.) that are required regardless of which specific energy-conversion strategy is used. Also, in some cases, our approach allows the 3-way linking of functional, phylo-genetic, and environmental patterns. For example, in respiratory complex I, the module covering the cyanobacterial NADH dehy-drogenases (i.e., most likely those from Prochlorococcus-like spec-ies) covaries positively with temperature and other photosynthesis modules (Tables S3 and S4). However, the module covering the proteobacterial NADH dehydrogenase (i.e., most likely from SAR11-like species) varies inversely with the temperature gradient. Such observations can be associated with their respective geo-graphic distributions. Photosynthetic Prochlorococcus are mostly absent in the northern, temperate sites but dominate in tropical waters (10, 12); whereas SAR11-like proteobacteria, which do not rely on the classical photosynthetic machinery to collect energy, dominate the northern, temperate regions (13). Thus, although variation in the reliance on autotrophic processes is not unexpected, these observations illustrate the potential of the proposed meth-odology to detect biologically relevant covariation.

**Balancing Amino Acid Synthesis vs. Import: Adapting to Nutrient-Limited Conditions.** We observed that metabolic pathways associated with amino acid and cofactor transport and metabolism varied significantly with the environmental features. Given the oligotrophic nature of the oceans (14), this observation may reflect the variability in amino acid uptake and recycling pathways as an alternative nutrient source in the various environments sampled, a strategy used by many of the dominating species in ocean surface waters (15). Lending further support to this hypothesis, operons with significant structural correlations consisted of both amino acid metabolism pathways and transporters necessary for exogenous uptake (Table S8). Amino acid uptake is sensitive to light availability (15), which, given the north to south sample collection gradient, could be an additional factor in their variation. The strength of this covariation is further reflected by the positioning of many of the amino acid metab-olism maps along the same principal axis as temperature and chlorophyll in the positive direction (Fig. S2).

**Environmentally Variant/Invariant Amino Acid Pathways Differ by Cofactor Cost.** One of the most striking aspects of our findings is that amino acid biosynthetic pathways could be divided into those that vary with the environment (high structural correlation coefficient) and those that do not. Interestingly, covariation of amino acid biosynthesis with the environment was unrelated to the energetic cost of synthesizing a particular amino acid (e.g., metabolic optimization). This simple result is seemingly counterintuitive, as one would expect that those pathways that used the most energy might vary the most with the energetic potential of their environment. However, we observe a significant positive correlation ($P < 0.05$) between the structural correlation of the amino acid pathways (strength of environmental covariation) and their dependence on potentially limiting cofactors (e.g., thiamin, tetrahydrofolate, cobalamin; see Materials and Methods and Table S9), corroborated by concordant variation in the ABC transporters of the cofactors.

This result suggests that the “cost” of obtaining trace metals for use in cofactors could be more expensive than the energetic cost of synthesizing transport machinery and degradative components that would allow for import of exogenous amino acids reducing the need for cofactor. The relationship among an environmental covariation of an amino acid, cofactor dependency, and transporters suggests the idea of “synthesis vs. import ” as an adaptive strategy in aquatic environments; i.e., the import of exogenous amino acids may be more favorable than direct synthesis in environments where the manufacture of the cofactors required for their synthesis is limiting.

**Environment-Driven Variation in Methionine-Dependent Pathways.** Methionine, a central amino acid in oceanic microorganisms, presents a particularly interesting example of this phenomenon and, also, illustrates the importance of a complex network of metabolic adaptations to limiting factors. Reduction in the use of methionine in nutrient limited environments has been noted previously (16). Our results suggest this reduction may stem from cofactor (and perhaps more specifically metal) cost optimization rather than (or in addition to) energetic constraints. We find environmentally-linked variation throughout methionine metabol-ism, including methionine synthesis, salvage, and degradation reinforced at multiple levels of pathway resolution. More specifically, we note that synthesis of both methionine and its cofactor cobalamin (contains cobalt) both decrease as methionine degradation and amino acid transporters (e.g., spermidine and putrescine) increase. Oceanic microorganisms have been shown to take extreme measures to conserve limited metals (e.g., iron) (17); these observations suggest an analogous adaptive response to cobalt limitation.

If such a limitation exists, one would expect to find equally wide-spread changes throughout methionine- (and, thus, cobal-amin-) dependent pathways; in particular, in those that depend on the cofactor $S$-adenosylmethionine (SAM), such as methyl-ation and secondary metabolites biosynthesis. Indeed, we do find evidence for environmental dependence for a whole suite of methionine processes, including cobalamin biosynthesis, as well as variation in many of the SAM-dependent processes (e.g., polyamines, ubiquinones, chlorophyll, and heme), hinting that methionine has a significant role in shaping downstream environ-mental adaptations. These observations provide evidence in support of a synthesis vs. import theory.

**Modulating Lipid and Glycan Metabolism As an Adaptation to Physicochemical Conditions.** Lipids and glycans are important compo-nents of the microbial outer membrane; thus, it would be expected to be particularly responsive to environmental condi-
tions. We do find strong environment-linked variation in a plethora of lipid and glycan metabolism-related processes (see Fig. 3 and Table S10). Indeed, modification of the cell wall is a known adaptive mechanism (e.g., for membrane fluidity) (9, 18), and the variation of pathways involved in extracellular polysaccharide synthesis, lipopolysaccharide synthesis, cell wall maintenance, and glycerophospholipid synthesis (Table S4) along the salinity, sample depth, and temperature gradients sampled in the GOS sites could be a reflection of this adaptation. Also, significant contributions of lipid metabolism modules in the construction of a LM for sample depth (Table S4) may illustrate an adaptation strategy to maintain buoyancy for optimal growth conditions (e.g., to optimally profit from light scavenging machinery adaptations for certain wavelengths, see ref. 12). Alternatively, it could reflect an adaptation of heterotrophic prokaryotes to the varying composition of phytoplankton-produced dissolved organic matter with depth. Due to the diversity of these roles of pathways without further experimentation, one can only speculate on the validity of these particular interpretations. However, undoubtedly, the extreme variation and flexibility of these pathways indicate their central importance in metabolic adaptation to the environment.

**Discussion**

As different evolutionary strategies are required to cope with the unique set of challenges specific to each geographic site sampled, our results suggest how environmental pressures shaped these pathway differences. The detailed analysis of 3 case studies revealed particular pathway adaptations that provide numerous testable hypotheses for linking metabolic versatility to the environment.

Recently, Dinsdale et al. (7) demonstrated that functional differences can be used to discriminate among 9 qualitatively categorized, discrete ecosystems. However, as in genome-wide association studies where methods using binned data have been supplemented by more sensitive methods that make use of continuous measurements (19), we have demonstrated the utility of a similar transition in microbial ecology by using comparative metagenomics. Our methods associate microbial community functions with quantitative, continuous features of the environment, allowing for an objective, data-driven framework to classify sites both on the basis of their metabolism and environmental parameters. We show evidence for widespread environmentally-dependent metabolic versatility even in seemingly similar sites (sharing same habitat classification).

The methods implemented here also provide a valuable and sensitive assay for simultaneously assessing a number of environmental parameters, allowing us to predict both individual and groups of environmental features (see Fig. 2 and Table S4). In reverse, we also predict the usage of a particular metabolic pathway given a set of environmental conditions (Table S11). Thus, our results suggest that metabolic footprints can be used as the basis of biosensors in situations where no clear measurable environmental factors are available (e.g., monitoring water quality and predicting health state from clinical samples). Indeed, such biosensors would provide more information than the current practice based on species composition (20), which measure downstream effects (e.g., marker species in pollution), instead of focusing on the molecular processes of the ecosystem as a whole, objective, data-driven framework to classify sites both on the basis of their metabolism and environmental parameters. We show evidence for widespread environmentally-dependent metabolic versatility even in seemingly similar sites (sharing same habitat classification).

The methods implemented here also provide a valuable and sensitive assay for simultaneously assessing a number of environmental parameters, allowing us to predict both individual and groups of environmental features (see Fig. 2 and Table S4). In reverse, we also predict the usage of a particular metabolic pathway given a set of environmental conditions (Table S11). Thus, our results suggest that metabolic footprints can be used as the basis of biosensors in situations where no clear measurable environmental factors are available (e.g., monitoring water quality and predicting health state from clinical samples). Indeed, such biosensors would provide more information than the current practice based on species composition (20), which measure downstream effects (e.g., marker species in pollution), instead of focusing on the molecular processes of the ecosystem as a whole, objective, data-driven framework to classify sites both on the basis of their metabolism and environmental parameters. We show evidence for widespread environmentally-dependent metabolic versatility even in seemingly similar sites (sharing same habitat classification).

Like all current metagenomics datasets, the GOS dataset provides only a snapshot of the total genomic content of a site. However, by quantifying the difference in pathway usage along different environmental gradients, one can observe the SPATIAL dynamics of pathways—analogous to the temporal dynamics in usage of pathways between different cellular states (27).

Although we have taken precautions to ensure the coverage across sites is the same (SI Materials and Methods), the potential remains for important but rare components of metabolic adaptation to be overlooked. Similarly, although we were able to map 74% of proteins to STRING OGs, there is still a fraction of hypothetical proteins that may harbor unknown and, thus, “unmapped” metabolic components. Indeed, environmental correlation may provide counterexamples for the annotation of proteins of unknown function. Novel techniques to functionally characterize this fraction represent a significant challenge and an avenue of active research (21, 22). Also, the 5 features reported do not fully encapsulate environmental complexity, and the integration of more environmental measurements will likely reveal many new and exciting discoveries. Despite the inherent limitations of the data, they do not compromise the ideas or the conceptual framework presented. Indeed, although the available datasets have constrained us to an analysis of aquatic habitats, the same methodology could readily be applied to investigate the specific metabolic capabilities for any ecosystem in which (physical) environmental parameters are collected including (e.g., different anatomical locations, which form “microbial habitats” in humans).

The potential contribution of large viruses as a reservoir for microbial diversity has recently been shown (23, 24). However, <0.3% of proteins in our set can be characterized as viral, suggesting a negligible impact on our reported findings (see Materials and Methods). Repeating this analysis on just the viral sample represents an interesting avenue for future research.

It is clear that microbial communities have a critical role in shaping our world from aiding in global climate regulation (25) and geochemical cycles to degrading hazardous byproducts; however, the complicated, intertwined nature between microbial communities and the environments they inhabit and influence remains poorly understood. We have presented a methodological framework that provides a roadmap to explore these questions in a systematic and statistically rigorous fashion.

**Materials and Methods**

**General Overview of Methodology and Data Used.** For full data and code dump, see http://networks.gersteinlab.org/metagenomics. For full details and extensive discussion of the validation of chosen procedures, see SI Materials and Methods.

**Preprocessing GOS Data.** Sequence and metadata (Fig. S3) from 37 sites (Table S1) from the first phase of the GOS expedition (0.1- to 0.8-μm filter size i.e., mostly prokaryotes), except for Sargasso Sea station 11; see ref. 35) was downloaded from CAMERA (36).

**Mapping.** Peptides were mapped to sites (Figs. S4 and S5) on the read-to-scaffold and orf-to-scaffold mappings available at CAMERA (36). Then, 111 KEGG maps, 141 modules, and 191 operons were assigned as indicated in ref. 14. Module definitions were downloaded from KEGG (37), and operons were constructed as indicated in ref. 38. In brief, protein sequences were searched against the extended database of proteins assigned to OGs in STRING 7.0 (38) by using BLASTP (39). A pathway was called present when a hit matched any of its components (bitscore >60; 80% consistency among top 5 hits; see SI Materials and Methods). Cofactors were mapped to each module by means of EC number by using the Brenda database (40). All results described were manually scrutinized to reduce artefactual assignments. The pathway frequency for each site was calculated by summing the total number of instances of that pathway in a particular site, then normalized by total number of assignments for all pathways in that site to compensate for sample coverage differences. Further normalizations were performed when necessary (see SI Materials and Methods). For all analyses, pathways for which the summed count over all sites constituted equal to or <0.01% of the total count were removed to avoid artifacts.

**Correlation and Regression.** We computed pairwise Spearman correlations between each pathway frequency vector and each environmental metadata vector. Linear regression models were constructed in 2 directions: (i) the environmental factor was treated as the response variable and predicted from a subset of pathway frequencies; and (ii) the inverse model where pathway
frequency was treated as the response variable and predicted from environmental factors (see SI Materials and Methods).

**DPM.** DPM was used to analyze whether groupings of sites based on similar environmental features also shared functional similarities. Sites were clustered based on their quantitative environmental metadata into “site-sets.” Next, we partitioned the sites in the metabolism matrices (Fig. 1A) into the same site sets and tested whether the means of each individual map, module, operon, and COG between the 2 site sets differed significantly (Benjamini–Hochberg corrected $P < 0.05$). Significant pathways were combined to form the DPM footprint (Table S5).

**CCA.** We used a regularized version of CCA to identify the set of projections that maximally correlate pathways and environmental variables (11). Those pathways, which had a structural correlation coefficient $>0.3$, formed the CCA footprint (Table S5).

**ACKNOWLEDGMENTS.** We thank Andrea Sboner, Roger Alexander, Albert Colman, and Nick Ornston for many thoughtful discussions, and Minoru Kanehisa’s team for early access to the KEGG module database. The instrumentation was supported by the Yale University Biomedical High Performance Computing Center and National Institutes of Health Grant R019895. J.O.K. was supported by a Marie Curie fellowship and J.R. and P.B., by the European Union FP7 Program Contract no. HEALTH-F4–2007–201052.

Supporting Information

Gianoulis et al. 10.1073/pnas.0808022106

SI Materials and Methods

GOS Data Collection and Preprocessing. For this study, we filtered the data from the first phase of the GOS expedition to keep only those sites that used a 0.1- to 0.8-μm filter size (with the exception of the Sargasso Sea station 11, which was excluded because it is suspected of contamination; see ref. 1); thus, only prokaryotes are part of this analysis. For the remaining 37 sites (Table S1), the site metadata were downloaded from the CAMERA database (2). For this study, measurements for temperature, sample depth, water depth, salinity, and monthly average chlorophyll level were used. Because 10 salinity measurements were missing, we averaged the salinity for all nonzero (excluded freshwater site) salinity measurements. In some cases, we were able to corroborate the missing measurements’ validity through extrapolating from the World Ocean Database (3). For the protein sequence data, the 6.1 million predicted proteins (4) were downloaded from CAMERA.

Mapping Peptides to Sites. Peptides were mapped to sites based on the read-to-scaffold and orf-to-scaffold mappings available at CAMERA (2). Thus, to assign these peptides to a particular site, we used a mapping algorithm that cross-referenced between reads, scaffolds, and peptides based on predicted gene coordinates (Fig. S5). Therefore, there were instances in which reads that formed part of a single peptide originated from 2 different sites; because this allowed peptides to be “present” in multiple sites, we term these “multisite” peptides (for additional details, see below).

Mapping Cofactors for Modules. Cofactors were mapped to each module via EC number by using the Brenda database (5). To normalize the effects of module size, the fraction of chemical reactions requiring certain cofactors per module is regarded as the cofactor-dependence of module (Table S8). We then used a goodness of fit test (K-S test) to compare the distribution of canonical correlation analysis (CCA) structural correlation coefficient between the amino acids that have no cofactors (score = 0) and those with cofactors (score > 0) (P < 0.05).

Assignment and Pathway Score. The 111 Kyoto Encyclopedia of Genes and Genomes (KEGG) maps, 141 modules, and 191 operons were assigned as in ref. 6. For clarity, in the remainder of the text, we use the term pathway to refer to all of these levels. Module definitions were downloaded from KEGG (7), and operons were constructed as in ref. 8. In brief, protein sequences were searched against the extended database of proteins assigned to orthologous groups (OGs) in STRING 7.0 (8), by using BLASTP (9), and a pathway was called present when a hit matching 1 of its proteins occurred (with a BLAST score of at least 60 bits). All results described were also manually scrutinized to reduce artefactual assignments.

The pathway frequency for each site was assigned by summing the total number of instances of that pathway for a particular site and normalizing by total number of assignments for that site to compensate for sample coverage differences. For all analyses, pathways for which the summed count over all sites constituted equal to or <0.01% of the total count were removed to avoid artifacts.

In addition, we calculated a mismatch rate where we looked to see how many times the top 5 BLAST hits for each peptide mapped to the same pathway. We find that 80% of the top-5 hits will map to the same pathway with a corresponding drop at less stringent bit scores, suggesting our results are threshold-independent. A second source of miscalling could be cross hitting of pathways by more “generalist” enzymes. Therefore, we have manually checked the assignments and sought confirmation at multiple levels of resolution (map-module-operon-OG) for all of the case stories reported in this study.

Pairwise Correlations and Linear Regression. We computed pairwise Spearman correlations between each pathway frequency vector and each environmental metadata vector for the same sample set, (P values corrected for multiple testing by using the Benjamini–Hochberg false discovery rate; see ref. 10). Linear models were constructed in 2 directions: (i) the environmental factor was treated as the response variable and predicted from a subset of pathway frequencies; and (ii) the inverse model where pathway frequency was treated as the response variable and predicted from environmental factors. To identify the subset of predictive variables, we used a stepwise regression analysis based on Akaike’s information criterion (implementation in R stats package). To avoid overfitting in (ii), we used only the top 20 pathways that showed the highest pairwise correlation (as measured by uncorrected P value) with the environmental feature modeled. As in many feature selection methods, one is not guaranteed the “best” subset, and we acknowledge that there can be multiple suboptimal solutions. Linear models were considered significant at P < 0.05 for both the total model and the estimate of the variable coefficients. For regressions in both directions, the pathway frequencies were standardized to a mean of 0 and a SD of 1. For (ii), we used the centered, quantile-normalized environmental data transformed into percentiles to ensure a truly normal distribution and, thus, accurate P values.

Clustering. The environmental data matrix was first standardized to mean of 0 and SD of 1. We evaluated distances by using 1-correlation and used average linkage hierarchical clustering. The clustering procedure was repeated by using spectral k-means without significance differences (data not shown).

Discriminative Partition Matching (DPM). To analyze whether groupings of sites based on similar environmental features also shared functional similarities, we clustered the sites based on their quantitative environmental metadata, resulting in 2 distinct clusters or site sets. Next, we partitioned the sites in the metabolism matrices (see Fig. 1A) into the same 2 site sets and calculated the mean normalized frequency for each pathway in each site set (see below for generalized approach). If the means of the pathway frequency between the 2 site sets were not significantly different, this would suggest that the environment-based partitioning does not reflect functional differences. If the distributions do differ significantly, it would imply that the environmental features are related to the specific aspect of metabolism. Also, we computed the 2-sample t test for each individual map, module, operons, and cluster of orthologous groups of proteins (COG). Those pathways that were significantly different (Benajamini–Hochberg corrected P < 0.05) were combined to form the DPM footprint (Table S4).

CCA. The goal of CCA is to identify the set of projections that maximally correlate 2 sets of variables (11). For a more detailed description of the relations of CCA to other common techniques, including principal components analysis and least squares regression, see ref. 12.
Due to the large number of dimensions and small number of data points, the solution can be unstable; thus, we applied a variant of CCA, regularized CCA (see ref. 13; implementation in ref. 14). We estimated regularization parameters $\lambda_1$ and $\lambda_2$ (penalty to covariance matrices) via a leave-one-out cross-validation procedure (implementation in ref. 14; see Table S11). Because of the interdependencies between metabolic pathways, canonical weights must be interpreted with caution. For this reason, we also calculated the structural correlation coefficient, which is the correlation between the original variable and the canonical variate. This allows one to specifically answer the question how important is this one variable (metabolic pathway) relative to all of the other variables (metabolic pathways) (see below for additional evaluation metrics). Those pathways, which had a structural correlation coefficient $>0.3$, formed the CCA footprint (Table S5). Also, we investigated the effect of changing this threshold (see Table S11). Principal components analysis and the resultant biplot on the environmental features show these features to be basically orthogonal (Fig. S4).

Evaluation Metrics and Controls. Construction and results from control matrices. To control for relative differences in metabolic pathways among the geographic locations simply reflecting sampling bias, we constructed 2 control matrices, composed of proteins that would not be expected to change among sites, such as those involved in basal transcription or translational machinery. The first is composed of those COGs categorized as information processing, and the second, those involved in cellular processes.

We used Student’s $t$ test and found that, although the distributions of the means for the control matrices (composed of those COGs annotated as belonging to either information or cellular processing) are not significantly different between the 2 environmental site sets ($P = 0.07$ and $0.08$, respectively), there are significant differences in metabolism ($P = 9 \times 10^{-3}$ and $4 \times 10^{-14}$ COG and KEGG annotated metabolism definitions; see refs. 7 and 15). However, we do see the same asymmetry as originally noted in the GOS paper for DNA polymerase, topoisomerase, and gyrase (4), by aggregating across the basal machinery this effect is minimized. Thus, the greatest strength of DPM is as a means of evaluating the functional significance of a particular partitioning and in controlling for potential sampling bias through the testing of control matrices (expected to be environmentally invariant) alongside matrices that are suspected of being environmentally variant.

More detailed CCA evaluation metrics. As in PCA, there are a number of metrics that can be used to determine the number of dimensions, in this case canonical variates, that should be included in the analysis (12). The overall canonical correlations for both dimension 1 and dimension 2 are high for KEGG maps, module, and operon; however, there is a significant drop in average redundancy between dimension 1 and dimension 2 and further dimension 2 and dimension 3, making it appropriate to use only these 2 dimensions in the overall analysis (Table S11). We can also measure the amount of information the environment is able to “cover” from the environment and vice versa by calculating the average variance of the dimensions and the redundancy (11). These measurements are both high for the environment but lower for the metabolic pathways. This suggests that there are many weaker signals coming from the metabolic matrices as opposed to a few strong ones.

Generalization of DPM. We provided a specific use of DPM in the text; however, DPM can be generalized. There are 3 basic steps to DPM. (i) The sites from the first matrix are partitioned to create site sets. (ii) The second matrix is partitioned in accordance with these site sets. (iii) A test (or ANOVA for more than 2 site sets) is performed to test whether the site sets are statistically different in the attributes of the second matrix.

Distributions of multisite and single-site peptides. In cases where a single peptide came from reads from different sites (multisite peptides) (16), we calculated the overlap between the reads and the peptide in 2 different ways. In the first, we considered the percentage of the read that was within a peptide (Eq. 1), and in the second case, we assessed the amount that the read contributed to the peptide (Eq. 2).

\[
\text{Fraction of read within peptide} = \frac{\text{Read overlap with peptide}}{\text{Read length}} 
\]

[1]

\[
\text{Fraction of peptide within read} = \frac{\text{Read overlap with peptide}}{\text{Peptide length}} 
\]

[2]

As illustrated in Fig. S3, the distributions of both peptide and read overlap are identical. This suggests that there are no major differences in assembly quality between the multisite and single site differences. However, it does not mean anything about the assembly quality itself. Only that if one assumes the assembly to be correct, there is no discernable differences between single and multisite peptides.

Comparison with Variance-Maximization Approaches. Compare and contrast with other methods. An entirely different approach to the one presented in the text assumes that the inherent variability of the environments could be directly observed by examining the global variance in the metabolic dataset; i.e., one identifies the pathways with greatest variance without directly measuring whether they covary specifically with the environment. First, we used standard deviation to find pathways that changed the most across the sites. We also used a PCA to identify the pathways that encapsulate the greatest proportion of variance. We then assessed the performance of these methods to identify metabolic adaptation to environmental parameters, based on their ability to recapture the environmental-based partitioning by using only the metabolic pathways identified as significant for each method by measuring cluster similarity (see below). Simply identifying the metabolic pathways with the greatest variance did not always reflect changes in the environmental parameters (see Fig. S2). Indeed, both methods that simultaneously incorporate environmental and metabolic data significantly outperform the variance-based, independent methods, and, perhaps, unsurprisingly, the linear models, which are more appropriate for investigating single relationships than looking at global context. These results were consistent, despite varying the number of pathways by using a variety of different thresholds for all methods SD, PCA, LM, DPM, and CCA (see Table S11).

Compare and contrast PCA and CCA. PCA and CCA actually have a deep relationship through formulation of the eigen problem, nicely illustrated in Borga et al. (12). Although they are related, there are completely different underlying assumptions motivating the use of one type and not the other. Although PCA attempts to capture the variance in a single dataset, CCA captures the within and between covariance (cross-variance) between 2 datasets. Thus, PCA can be used to extract components with the highest global variance, and has been used extensively in comparative metagenomics under the assumption that any variance observed could be attributed to environmental changes. Such reasoning makes sense when comparing qualitatively dissimilar environments. As an example, the difference between soil and water cannot really be quantified in a meaningful way, because all of the variables are changing simultaneously. Thus, more precise measurements to see how say metabolism varies as a function of the environment would not be...
informative or even necessarily feasible. However, where environments change only slightly along a continuum of features, CCA provides the strength of covariation between both environmental and metabolic differences.

**Metrics to compare cluster similarity.** Cluster similarity was measured by comparing both a normalized mutual information (NMI) score, which measures the amount of information lost if one applies the classification “clustering” from the first partition to the second (17), and the rand index (18), which computes the number of “correct” pairwise interactions between the 2 sets of clusters. The closer the NMI score is to 1.0 the better the metabolic footprint generated from the method performed in recapitulating the structure of the environmental data. For each set, a significance value for the rand index was computed by randomly shuffling the clustering assignment, recalculating the index after each iteration, and counting the number of times the index computed from the random data exceeded the index computed from the real data after 10,000 iterations.

Fig. S1. Comparison of different classes of methods. We evaluated the efficacy of 3 different classes of methods based on their explicit use of the quantitative environmental data, which we term independent, isolated, and simultaneous. Independent methods include no environmental description (green), isolated only one environmental feature at a time (purple), and simultaneous methods incorporate all environmental features simultaneously (blue). For clarity, we refer to the highly-weighted set of pathways generated for each method as a footprint. Each of the 5 methods was used to generate a metabolic footprint, and each bar represents the NMI score for that footprint of method. No statistically significant difference was observed between scores within each particular category (P > 0.05).}

Gianoulis et al. www.pnas.org/cgi/content/short/0808022106
Color Legend for Pathway Functional Category
- Carbohydrate Metabolism
- Energy Metabolism
- Lipid Metabolism
- Amino Acid Metabolism
- Nucleotide Metabolism
- Glycan Biosynthesis and Metabolism
- Cofactor and Vitamin Metabolism
- Biosynthesis of Secondary Metabolites
- Xenobiotic Degradation and Metabolism

Fig. S2. Bullseye plot of CCA-derived structural correlations. Shown are results from CCA for KEGG (A) and module (B). The x and y axes represent the structural correlation coefficients (normalized weights) in the first and second dimension, respectively. The closer either environmental features (red triangles) or metabolic pathways (color coded by functional category) are to the perimeter of the outer ellipse, the better they fit the model. Also, the closer an environmental feature is to a metabolic pathway the stronger the covariation between them. The inner ellipse (radius 0.3) represents those features that did not fit the model (for further explanation, see ref. 14). Those pathways in the inner ellipse can be thought of as environmentally invariant, and those outside this ellipse as environmentally variant.
Fig. S3. Biplot and boxplot of standardized environmental variables. To examine possible dependencies between the variables, we performed principal-component analysis. (A) We next plotted component 1 and 2. One can see that the variables with the exception of temperature and salinity are basically orthogonal to one another. (B) Boxplot of standardized variables.
Fig. S4. Distribution of multisite and all-peptide and read overlaps. We term peptides whose reads derived from multiple geographic locations “multisites.” (A and B) Fraction of each peptide that is within read (A) and the fraction the read “contributes” to the peptide (B). The same plots are shown for all peptides: peptide overlap (C) and read overlap (D). While B and D are in part assessing if differences in assembly between multi- and single-site peptides exist, A and C illustrate more functionally that there is no observable difference between 2 types.
Given a scaffold where
R - set of reads
BS - set of buckets to which reads from R belong
P - set of peptides

Pseudocode:
foreach p in P:
s = p's scaffold
find R(s)
foreach r in R(s):
    if r overlaps with p
        put p in r's bucket

Fig. S5. Mapping peptides to geographic locations (sites). Schematic and pseudocode for mapping of peptides to a particular site are shown. The goal of this algorithm is, given this set of reads (color coded blue or green, depending on which bucket (site) they were recovered from), this set of peptides (boxes), and the coordinates from the scaffold (long black line), to determine to which buckets the peptides (boxes) belong.

Other Supporting Information Files

Table S1
Table S2
Table S3
Table S4
Table S5
Table S6
Table S7
Table S8
Table S9
Table S10
Table S11