COMPARATIVE ANALYSIS OF AEROBIC AND ANAEROBIC PROKARYOTES TO IDENTIFY CORRELATION BETWEEN OXYGEN REQUIREMENT AND GENE-GENE FUNCTIONAL ASSOCIATION PATTERNS

YAMING LIN HONGWEI WU
linyaming@gatech.edu hongwei.wu@gatech.edu

School of Electrical and Computer Engineering, Georgia Institute of Technology
210 Technology Circle Savannah, Georgia, 31407, USA

Activities of prokaryotes are pivotal in shaping the environment, and are also greatly influenced by the environment. With the substantial progress in genome and metagenome sequencing and the about-to-be-standardized ecological context information, environment-centric comparative genomics will complement species-centric comparative genomics, illuminating how environments have shaped and maintained prokaryotic diversities. In this paper we report our preliminary studies on the association analysis of a particular duo of genomic and ecological traits of prokaryotes – gene-gene functional association patterns vs. oxygen requirement conditions. We first establish a stochastic model to describe gene arrangements on chromosomes, based on which the functional association between genes are quantified. The gene-gene functional association measures are validated using biological process ontology and KEGG pathway annotations. Student's t-tests are then performed on the aerobic and anaerobic organisms to identify those gene pairs that exhibit different functional association patterns in the two different oxygen requirement conditions. As it is difficult to design and conduct biological experiments to validate those genome-environment association relationships that have resulted from long-term accumulative genome-environment interactions, we finally conduct computational validations to determine whether the oxygen requirement condition of an organism is predictable based on gene-gene functional association patterns. The reported study demonstrates the existence and significance of the association relationships between certain gene-gene functional association patterns and oxygen requirement conditions of prokaryotes, as well as the effectiveness of the adopted methodology for such association analysis.

Keywords: oxygen requirement conditions, gene-gene functional association, gene ontology similarity, KEGG pathway annotation

1. Introduction

There are countless ways in which prokaryotes influence our daily life. On the other hand, environments have undoubtedly left footprints on prokaryotic morphological, physiological and functional diversities [16]. It is, therefore, very important to study genomes in ecological contexts, and such importance has been recognized by the community. This has led to substantial progresses in multiple enabling fields. Environment Ontology (EnvO) and Habitat-Lite [5] have been proposed to stan-

dardize descriptions of the environment, and to be incorporated into the standard descriptions of genomes [3]. The reduced cost of DNA sequencing and world-wide sequencing efforts have made available sequences of ∼1,000 complete and ∼2,000 in-progress prokaryotic genomes [12]. With the wealth of genomic/metagenomic sequences and about-to-be-standardized ecological context information, environment-centric comparative genomics has the potential to complement species-centric comparative genomics in illuminating how environments have shaped and maintained prokaryotic diversities.

Bohlin *et al.* analyzed the genome composition of host-associated versus free-living prokaryotes, and demonstrated that the oligonucleotide usage in non-coding regions varied more for the former than the latter group of organisms [1]. Paul *et al.* [13] studied the genomic and proteomic features of halophilic and non-halophilic organisms. They discovered distinctive molecular-level signatures in halophiles that are independent of their GC-content or phylogenetic origins. These signatures include, for example, (i) over- and under-representations of certain amino acids, (ii) different propensities for the helix and coil structures, and (iii) dinucleotide abundance and synonymous codon usage preference patterns that are not species-specific but salt adaptation-specific. It was also suggested that the amount and source of horizontal gene transfer is linked to an organism's lifestyle. For instance, bacterial hyperthermophiles seem to have exchanged genes with archaea to a greater extent than other bacteria, whereas transfer of certain classes of eukaryotic genes is most common in parasitic and symbiotic bacteria [10]. Evidence has shown that the genome size and gene content in bacteria are associated with their lifestyles. For example, species with larger genome size are more metabolically versatile, able to exploit a larger number of ecological niches and exhibit larger intra-species differences; and, host-associated bacteria typically have a smaller genome size and fewer genes [8, 9]. Comparative analysis on host-associated and free-living bacteria also indicated that host-associated bacteria in general have fewer rRNA genes, more split rRNA operons and fewer transcriptional regulators. And, between mutualists and parasites, the former group has significantly more genes that enable nutrient provisioning, whereas the latter group has more genes of secretion systems [11]. Suen *et al.* used the distribution of protein domains in various Pfam families to classify prokaryotes, and compared this classification result against the 16S rRNA-based phylogenetic map. The comparison revealed that the prokaryotic organisms occupying the same ecological niche tend to possess a similar genetic repertoire due to the evolutionary pressure exerted by the ecological niche [14].

As described above, the association between prokaryotic genomic and ecological traits lies in multiple aspects. The genomic traits can include basic genomic and proteomic features (e.g., genome size, GC-content, preference for various synonymous codons, and various protein structures), as well as features indicative of evolutionary traces (e.g., horizontal gene transfer), functional potentialities (e.g., distribution of genes in various functional categories), and regulation efficiency (e.g., operon structures). And, the ecological traits can include habitat, temperature, pH, salinity,

pressure, light intensity, oxygen, nutrient sources, etc. We here focus on a particular duo of genomic and ecological traits – gene-gene functional association patterns vs. oxygen requirement conditions, to determine whether there exist statistically as well as biologically significant association relationships between these two traits. We first establish a mathematical model to quantify gene-gene functional association, and validate this model using biological process ontology [15] and KEGG pathway annotations [7] (Section 3). We then perform student's *t*-tests [6] to identify those gene pairs that exhibit different functional association patterns for aerobic and anaerobic organisms (Section 4). Finally, to further validate the significance of the identified association relationships, we examine whether and to what extent the oxygen requirement property of an organism can be predicted based on its gene-gene functional association properties (Section 5). Conclusions are drawn in Section 6.

2. Aerobic and Anaerobic Prokaryotes

As of March 2009, the NCBI Microbial Genome Project Database [12] contains 841 complete prokaryotic genomes, including 260 aerobes and 147 anaerobes. By *aerobic* we mean that the organism can grow in the presence of oxygen and probably uses oxygen as an electron acceptor; and, by *anaerobic* we mean that the organism grows in the absence of oxygen and utilizes alternative electron acceptors.

The ideal strategy to investigate the association between genomes and a particular ecological factor (e.g., oxygen requirement) is to select those organisms whose diversities in other ecological aspects (e.g., salinity, habitat, and temperature) are comparable for the different ecological conditions being focused on (e.g., aerobic and anaerobic conditions). However, as revealed by Table 1 in supplementary materials, the distributions of the aerobic and anaerobic organisms in various salinity, habitat and temperature conditions are slightly different. If we selected the aerobic and anaerobic prokaryotes with their other ecological conditions being identical (e.g., non-halophilic, host-associated and mesophilic temperature), we would end up with a much smaller number of organisms, which will hinder us from drawing statistically significant conclusions. Therefore, we have decided to use all these aerobic and anaerobic prokaryotes for the comparative and association analysis.

In terms of the phylogenetic diversity, as revealed by Table 2 in supplementary materials, the majority ($\sim$80%) of the aerobic prokaryotes belong to *Proteobacteria*, *Actinobacteria* and *Fimicutes* with *Proteobacteria* being the dominant phylum; whereas the majority ($\sim$70%) of the anaerobic prokaryotes belong to *Firmicutes*, *Proteobacteria* and *Euryarchaeota* with these three phyla being approximately equally populated. To investigate the association between particular genomic features and ecolgocial contexts, we should ideally exclude the impact of factors other than ecological contexts (such as phylogenetic origins) on the genomic features. This could be achieved by focusing on strains belonging to the same taxonomic lineage but with different ecological contexts, such as strains in the *Geobacter* genus that

have different oxygen requirements and/or habitat types. However, for comparative genomics that involves a wide range of genomes, it is difficult to distinguish whether phylogenetic origins and ecological contexts are in the same or different roles in shaping the genomic features. Therefore, we have decided to consider all taxonomic lineages together for the environment-centric comparative and association analysis.

The number of genes in each genomes ranges from 536 to 9,703, and on average ∼42.8% of the genes are annotated with KEGG orthologous groups [7]. We here focus on these KEGG orthology-annotated genes, and consider genes with the same KEGG orthology annotations as the "same" (orthologous). There are totally 5,436 different KEGG orthologous groups, which lead to over 14 million gene pairs to be investigated.

3. Quantification of Gene-Gene Functional Association

For prokaryotic organisms, the arrangement of genes on chromosomes affects how efficiently genes are transcribed, and consequently reflects how different parts of the cellular machinery are coordinated in response to environmental changes. The information regarding whether genes are adjacent or how distantly genes are separated has been demonstrated to be effective for operon prediction [2], functional module prediction [17], etc. Here we also use the same information to described gene-gene functional association.

3.1. *Stochastic Model for Gene Arrangement*

The hypothesis of this stochastic model for gene arrangement is that there exists no functional association between two genes so that one gene's presence or position on the chromosome is independent of that of the other gene. Given any two genes g_i and g_j, the evidence in each genome to support this hypothesis can be described as:

$$\begin{aligned} L(g_i, g_j) = {} & I(g_i, g_j, |g_i - g_j| \le d_{ij}) \left[p_i p_j \sum_{m=1}^{M} \frac{N_m + (2N_m - 1)d_{ij} - d_{ij}^2}{N^2} \right] \\ & + I(g_i, g_j, NA) \left[p_i p_j \left(1 - \sum_{m=1}^{M} \left(\frac{N_m}{N} \right)^2 \right) \right] \\ & + I(g_i, \bar{g}_j) p_i (1 - p_j) + I(\bar{g}_i, g_j)(1 - p_i) p_j + I(\bar{g}_i, \bar{g}_j)(1 - p_i)(1 - p_j) \end{aligned} \tag{1}$$

where the five terms on the right hand side account for the scenario that (*i*) both g_i and g_j belong to the same directon (a directon is a continuous stretch of genes transcribed along the same direction), (*ii*) g_i and g_j are present in the genome but do not belong to the same directon, (*iii*) g_i is present in the genome but g_j is not, (*iv*) g_i is not but g_j is present in the genome, and (*v*) neither g_i nor g_j is present in the genome, respectively. p_i denotes the probability of g_i being present, which is estimated as the ratio of the number of organisms containing g_i to the

total number of organisms. d_{ij} denotes the number of genes present between g_i and g_j. N is the total number of genes. All these genes are grouped into M directons according to their position on the chromosome and transcription direction, with N_m denoting the number of genes in the m-th directon.And, $I()$ is the indicator function. The so-calculated $L(g_i, g_j)$ reflects the level of validity of the hypothesis associated with the stochastic model. The smaller $L(g_i, g_j)$ is, the more strongly is the alternative hypothesis that the two genes are functionally associated being supported. We therefore use $A(g_i, g_j) \equiv -\log L(g_i, g_j)$ to quantify the functional association between g_i and g_j.

3.2. *Validation of Gene-Gene Functional Association Measures*

We here examine whether the model in Eq. (1) can faithfully capture the functional association between genes, using biological process ontology and KEGG pathway annotations.

3.2.1. *Validation of the $A(g_i, g_j)$ Measures based on Biological Process Ontology Annotations*

The Gene Ontology(GO) describes properties of gene products from three independent perspectives – cellular component, molecular function, and biological process. In particular, the biological process ontology describes to what biological objectives gene products contribute and in what series of biological events gene products are involved. The GO is structured as a directed acyclic graph, wherein (*i*) each term is a child of one or more terms, (*ii*) child terms are instances or components of the parent terms, and (*iii*) child terms can provide more specific descriptions than the parent terms [15]. Genes with the same biological process GO annotations can be considered as functionally associated to a certain extent. Therefore, our validation is based on biological process GO annotations.

For each biological process GO term, we define its depth as the number of terms in the longest path from it to the root term (GO:0008150, biological process). For a pair of biological process GO terms, go_1 and go_2, we define their similarity as the maximum depth of all their common ancestor terms, i.e.,

$$S^{GO}(go_1, go_2) = \max\{depth(go), go \in C_1 \cap C_2\}, \tag{2}$$

where C_i denotes the ancestor terms of go_i $(i = 1, 2)$, and $C_1 \cap C_2$ denotes the common ancestors of go_1 and go_2. And, for a pair of genes, g_1 and g_2, among all the pairs formed by their biological process GO terms, we take the maximum similarity as the GO similarity of these two genes, i.e.,

$$S^{GO}(g_1, g_2) = \max\left\{S^{GO}(go_1, go_2), go_1 \in S_1, go_2 \in S_2\right\} \tag{3}$$

where S_i denotes the biological process GO terms applicable to g_i $(i = 1, 2)$.

The above defined GO similarities have been used to derive information for predicting protein-protein interaction [19] and for validating the functional association measures derived from other resources [17]. We here use them to validate

the $A(g_i, g_j)$ measures. Specifically, we will verify whether gene pairs that are indicated as strongly/moderately/weakly functionally associated by the $A(g_i, g_j)$ measure are also indicated so by the $S^{GO}(g_i, g_j)$ measure. We first calculate the average of the $A(g_i, g_j)$ measure across all organisms for each gene pair. Then based on the average $A(g_i, g_j)$ measures, we form three groups, consisting of strongly, moderately, and weakly functionally associated gene pairs, respectively. By *strongly/moderately/weakly*, we mean the 93,118 gene pairs whose average $A(g_i, g_j)$ measures are among the top/middle/bottom one percent. We then compare the distributions of the $S^{GO}(g_i, g_j)$ measure for these three groups of gene pairs, as shown in Fig. 1.

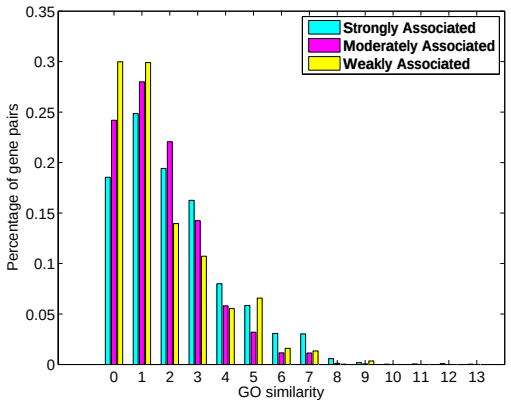

Fig. 1. Distributions of $S^{GO}(g_i, g_j)$ for gene pairs that are considered as strongly (cyan), moderately (magenta) and weakly (yellow) functionally associated based on the $A(g_i, g_j)$ measures. The horizontal axis corresponds to the $S^{GO}(g_i, g_j)$ measure, ranging from 0 to 13. The bars correspond to the percentages of gene pairs whose $S^{GO}(g_i, g_j)$ measures fall within particular ranges.

Kolmogorov-Smirnov tests [6] for pairwise comparisons between these three distributions have indicated that the groups of the $S^{GO}(g_i, g_j)$ measure are different (with p-value $\leq 1e - 103$) for strongly-, moderately-, and weakly- associated gene pairs. As shown in Fig. 1, gene pairs with stronger functional association are more likely than the gene pairs with weaker functional association to be associated with larger $S^{GO}(g_i, g_j)$ measures. Therefore, the $A(g_i, g_j)$ and $S^{GO}(g_i, g_j)$ measures are positively correlated; and, the $A(g_i, g_j)$ measure is as qualified as the $S^{GO}(g_i, g_j)$ measure to quantify the functional association between genes.

3.2.2. *Validation of the $A(g_i, g_j)$ Measures based on KEGG Pathway Annotations*

The KEGG Pathway database is a collection of manually drawn pathway maps that represent the knowledge about the molecular interactions and reaction networks [7].

Because genes involved in the same pathway can be considered as functionally associated to a certain extent, we here validate the $A(g_i, g_j)$ measures based on the KEGG Pathway annotations. Specifically, we will examine whether those gene pairs with different levels of the $A(g_i, g_j)$ measure are involved in the same pathway with different levels of likelihood.

Given a group of gene pairs, the likelihood that the two genes of a randomly drawn pair are involved in the same pathway can be estimated as the ratio of the number of gene pairs being involved in the same pathway to the total number of gene pairs in the group. For the three groups, which contains the strongly, moderately and weakly functionally associated gene pairs, respectively, the ratios are 3.897%, 0.772%, and 0.607%. That is, those gene pairs that are indicated as strongly functionally associated by the $A(g_i, g_j)$ measure are (5∼6 times) more likely to be involved in the same pathway than those indicated as moderately/weakly associated. Therefore, based on the KEGG pathway annotations, the $A(g_i, g_j)$ measures can capture the functional association between genes.

4. Identification of Gene Pairs with Different Functional Association Patterns under the Two Different Oxygen Requirement Conditions

The goal of our study is to identify those gene pairs whose functional association patterns are statistically as well as biologically linked with the oxygen requirement conditions of prokaryotes. Each pair of genes are associated with two collections of $A(g_i, g_j)$ measures that corresponds to the 260 aerobic and 147 anaerobic organisms, respectively. By conducting an unpaired two-sample student's t-test [6] on the two collections (See supplementary materials for details of student's t-test), we can determine to what extent (measured by the p-value derived from the test) the functional association between g_i and g_j exhibit different distribution profiles for the two different oxygen requirement conditions.

4.1. *Student's t-Test Results*

Based on the p-values derived from the t-tests (See supplementary Fig. 1 for p-value distributions), we have formed two gene pair groups. The first group consists of 93,118 gene pairs whose p-values range from 0.9576 to 1 and are among the largest one percent of all the p-values. The second group consists of 93,118 gene pairs whose p-values range from 0 to $1e-18$ and are among the smallest one percent of all the p-values. The first group can be interpreted as to contain those gene pairs whose functional association are invariant to the oxygen requirement condition; and the second group contains those gene pairs whose functional association measures exhibit different distribution patterns for the two different oxygen requirement conditions. Fig. 2 shows the average $A(g_i, g_j)$ measure across all aerobic organisms (x axis) vs. the average $A(g_i, g_j)$ measure across all anaerobic organisms (y axis) for these two group of gene pairs. Note that the first group(with large p-values) are

mainly distributed along the $y = x$ line, which is consistent with our interpretation that gene pairs with large p-values have their functional association relationships invariant to the oxygen requirement condition. Also note that the second group of gene pairs (with small p-values) are generally off the $y = x$ line, with 50,293 pairs above and 42,879 pairs below the line. Those gene pairs above the $y = x$ line can be interpreted as to be more strongly functionally associated in the anaerobic condition. Whereas, those gene pairs below the $y = x$ line can be interpreted as to be more strongly functionally associated in the aerobic condition.

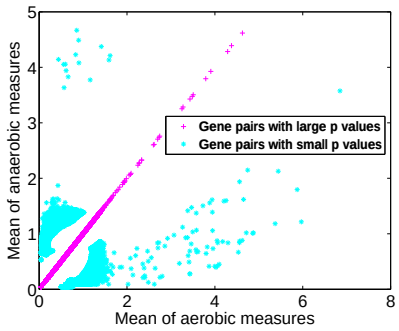

Fig. 2. The average $A(g_i, g_j)$ measure across aerobic organisms (x-axis) vs. the average $A(g_i, g_j)$ measure across anaerobic organisms (y-axis).

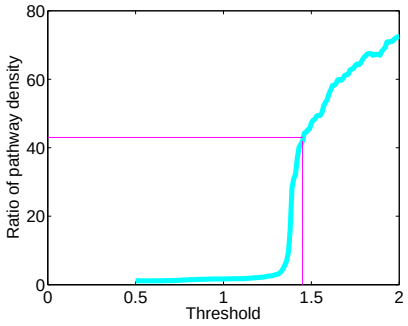

Fig. 3. Likelihood ratio of being involved in the same KEGG pathway as a function of the threshold value. The threshold distinguishes those strongly associated gene pairs in some oxygen requirement condition from the other gene pairs.

4.2. *Biological Implications of the Gene Pairs with Large/Small p-Values*

Within the group of gene pairs whose functional association relationships are invariant to the oxygen requirement condition (the first group aforementioned), we are more interested in those that are always strongly associated in both aerobic and anaerobic conditions. Similarly, within the group of gene pairs with stronger functional association relationships in one of the oxygen requirement conditions (the second group aforementioned), we are more interested in those that are strongly associated in either aerobic or anaerobic condition. This is because these strongly associated gene pairs, whether uniformly or conditionally, may reflect which part(s) of the cellular machinery are invariantly or conditionally "wired" with respect to different oxygen requirement conditions.

To distinguish those gene pairs that are strongly associated in either aerobic or anaerobic or both conditions, we have utilized the following approach involving KEGG pathway annotations. Assume t is a threshold value, and those gene pairs

whose average $A(g_i, g_j)$ measure across aerobic organisms, or average $A(g_i, g_j)$ measure across anaerobic organisms, or both averages is greater than or equal to t are considered as strongly associated in some oxygen requirement condition. Intuitively, the strongly associated gene pairs are more likely than the other pairs to be involved in the same pathway; hence, we have compared the likelihood of strongly associated gene pairs being involved in the same pathway against the likelihood of the other gene pairs being involved in the same pathway. Fig. 3 shows this likelihood ratio as a function of t. Note that this likelihood ratio increases abruptly when t is in the range of [1.35, 1.45]. We have therefore selected $t = 1.45$ to screen out those gene pairs that are strongly associated in either aerobic or anaerobic or both oxygen requirement conditions.

By applying this threshold value, we have identified 99 gene pairs with strong association in the aerobic condition, 28 gene pairs with strong functional association in the anaerobic condition, and 51 gene pairs with strong association in both aerobic and anaerobic associations. Complete lists are provided in the supplementary material. It is expected that genes involved in energy metabolism (e.g., pyruvate/oxoglutarate oxidoreductases, oxidative phosphorylation, nitrogen metabolism) are arranged and functionally associated in different ways in different oxygen requirement conditions. It is also expected that genes involved in the genetic information processing (e.g., transcription and translation machinery) are conserved for organisms of various ecological contexts and genes related to cell motility (e.g., flagellar and motor genes) are subject to impacts of other environmental factors and are therefore invariant to oxygen requirement conditions. However, it is interesting to observe that genes relevant to virulence (e.g., urease, secretion system) also exhibit different functional association patterns for different oxygen requirement conditions.

5. Prediction of Oxygen Requirement Conditions Based on certain Gene-Gene Functional Association Patterns

The significance of the association between the gene-gene functional association patterns and the oxygen requirement condition can be reflected statistically by the p-values obtained through the student's t-tests. However, because the association between genomic and ecological traits are the result of long-term accumulative interactions between prokaryotes and the environment, it is difficult to design wet-lab experiments that can generate relatively long-term observations for further validations. We therefore focus on computational validations, and examine whether and to what extent the ecological trait (e.g., oxygen requirement condition) of an organism can be predicted based on the genomic traits (e.g., gene-gene functional association patterns). The rationale for this computational validation approach lies in that the more predictable the ecological traits are based on the genomic traits, the more significantly the ecological and genomic traits are associated. The procedure for this predication-based computational validation is as follows.

First, all of the 407 prokaryotic organisms, either aerobic or anaerobic, are randomly partitioned into two sets, where the set consisting of 333 (∼80%) organisms is used for training, and the other set consisting of 74 (∼20%) organisms is used for testing. Secondly, using the training set, we estimate for each gene the probability of being present in a genome (p_i in Eq. (1)), quantify for each gene pair the functional association, and conduct the student's t-test to identify gene pairs whose functional association relationships are either invariant or stronger in one of the oxygen requirement conditions. We are particularly interested in those gene pairs whose functional association measures follow different distribution profiles for different oxygen requirement conditions, and are strong in either aerobic or anaerobic conditions, since the functional association measures of these gene pairs will serve as the features for the predictions. We use $t = 1.45$ (see Section 4.2) to screen out those gene pairs whose average functional association measure across all training aerobic (or anaerobic) organisms is greater than or equal to $t = 1.45$ and use them as the genomic features of the aerobic (or anaerobic) condition. And finally, the performance of predicting the ecological trait based on genomic traits will be evaluated using organisms in the testing set. For each testing organism, two genomic feature vectors will be generated that correspond to the aerobic and anaerobic conditions, respectively. The aerobic feature vector consists of the functional association measures of those signature gene pairs for the aerobic condition, and is compared with the genomic features of the aerobic condition to determine the Euclidian distance between the testing organism and training aerobic organisms. Similarly, we can determine the Euclidian distance between the testing organism and training anaerobic organisms. The oxygen requirement trait of the testing organisms is predicted as aerobic (or anaerobic) if its distance to the aerobic (or anaerobic) condition is smaller.

The above procedure is repeated for 10 times, so that the effectiveness of the prediction can be evaluated using different testing sets. The prediction accuracy rates for these 10 repeats are summarized in Table 1. Observe that the prediction accuracy for all the 10 repeats of the random-partitioning-training-and-testing procedure is higher than or equal to 80%, and can even reach ∼90% or higher for some repeats. Considering that a random guess of an organism's oxygen requirement condition can only yield a 65% accuracy at most, the prediction results based on the organism's gene-gene functional association patterns, which can reach an 80% or even higher accuracy rate, indicate that it is feasible to predict the ecological trait based on the genomic trait. Consequently, the association between certain gene-gene functional association patterns and the oxygen requirement condition is far from being random or trivial; rather, such association reflects the interactions between prokaryotic genomes and the environment.

Table 1. Accuracy rate of predicting the oxygen requirement condition of an prokaryotic organism based on its gene-gene functional association patterns, obtained for 10 repeats of the random-partitioning-training-and-testing procedure.

Experiment Number	**Prediction Accuracy**	**Experiment Number**	**Prediction Accuracy**
Experiment 1	91.89%	Experiment 2	81.08%
Experiment 3	87.84%	Experiment 4	85.13%
Experiment 5	87.84%	Experiment 6	86.49%
Experiment 7	90.54%	Experiment 8	86.49%
Experiment 9	79.73%	Experiment 10	85.13%

6. Conclusion

In this paper, we have reported our preliminary investigations on the correlation between gene-gene functional association patterns and oxygen requirement conditions of prokaryotes. We have first established a stochastic model to quantify functional association between genes. The gene-gene functional association measures have then been validated using biological process GO and KEGG pathway annotations. Such validations have revealed that gene pairs indicated as strongly associated by the proposed measure are also more likely to be associated with high GO similarities and be involved in the same pathway, which means that the stochastic model-based quantification can well capture the functional association between genes. We have then performed the student's t-tests on the gene-gene functional association measures of the aerobic and anaerobic organisms, and have then identified those gene pairs that exhibit different functional association patterns under different oxygen requirement conditions. We then have conducted computational validations to examine whether the oxygen requirement trait of an organism can be predicted based on the gene-gene functional association patterns. The 10 repeats of the random-partitioning-training-and-testing experiment have shown that an 80% or even higher ($\sim$90%) accuracy rate can be achieved for such predictions. Compared to the maximum 65% accuracy rate that can be achieved by random guessing, our computational validation results may indicate the existence and significance of the association between certain gene-gene functional association patterns and oxygen requirement conditions of prokaryotes, as well as the effectiveness of the methodology we have adopted in exploring the genome-environment association relationships.

Though our preliminary studies have rendered very promising results, there are several questions yet to be answered in our future studies. First, how well can the association relationships between certain gene-gene functional association patterns and oxygen requirement conditions that have been identified through the reported preliminary study be generalized to other prokaryotic genomes? Our preliminary studies have been based on the complete prokaryotic genomes, which only account for a tiny percentage ($< 1\%$) of the entire microbial world. We will take advantage of the in-progress prokaryotic genomes and metagenomes in our future study to generalize our comparative and association analysis. Particularly, we will predict

the oxygen requirement traits of the in-progress genomes and metagenomes using the predictor model built on the complete genomes to determine how accurate the prediction can be and consequently how well the association relationships can be generalized. Secondly, how well can the methodology adopted in the reported preliminary study, including quantification of the gene-gene functional association, statistical test, and predication-based computational validations, be generalized to other environmental factors? We will mainly focus on the phenotypic and ecological traits that are specified in the NCBI Microbial Genome Project database, including motility, salinity, habitat, temperature, and pathogenicity. We will investigate whether certain gene-gene functional association patterns are statistically and biologically associated with these phenotypic and ecological traits of prokaryotes. Finally, is it possible to decouple the impact of phylogenetic diversities and the impact of environmental diversities on the genomic features of prokaryotes? As we have observed from Table 2 in supplementary materials, the aerobic prokaryotes are mainly distributed in the *Proteobacteria* phylum, wherease the anaerobic prokaryotes are mainly distributed in the *Firmicutes*, *Proteobacteria* and *Euryarchaeota* phyla. We may argue whether the difference in the phylogenetic distributions also contributes to the difference in the gene-gene functional association patterns between the aerobic and anaerobic organisms. Because the environment has undoubtedly been involved in the prokaryotic evolution, it is quite difficult to determine whether the environmental factors are directly linked with the genomic traits, or indirectly through the phylogenetic traits. Though we do not have answers for this question yet, we will compare those genomic traits that are signatures of certain phylogenetic origins against those that are signatures of certain ecological contexts. If the two sets of the signature genomic traits are significantly different, we may claim that the phylogenies and environments are in different roles in shaping the the genomic traits of prokaryotes.

Acknowledgements

This research is supported in part by National Science Foundation (NSF/OCE-0928196).

References

[1] Bohlin, J., Skjerve, E., and Ussery, D.W., Investigations of oligonucleotide usage variance within and between prokaryotes, *PLoS Computational Biology*, 4(4):e1000057, 2008.

[2] Ermolaeva, M.D., White, O., and Salzberg, S.L., Prediction of operons in microbial genomes, *Nucleic Acids Res.*, 29(5):1216-1221, 2001.

[3] Field, D., Garrity, G., Gray, T., Morrison, N., The minimum information about a genome sequence (MIGS) specification, *Nature Biotechnology*, 26(5):541–547, 2008.

[4] Fukuchi, S., Yoshimune, K., Wakayama, M., Moriguchi, M., and Nishikawa, K., Unique Amino Acid Composition of Protein in Halophilic Bacteria, *J. Mol. Bio.*, 327(2) 347–357, 2003.

[5] Hirschman L., Clark C., Cohen K.B., Mardis, S., *et al.*, Habitat-Lite: a GSC case study based on free text terms for environmental metadata, *OMICS*, 12(2):129–136, 2008.
[6] Kanji, G.K., 100 Statistical Tests, *Sage Publications Ltd*, 2006.
[7] Kanehisa, M., Araki, M., Goto, S., Hattori, M, Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T., and Yamanishi, Y., KEGG for linking genomes to life and the environment, *Nucleic Acids Res.*, 36(Database issue):D480-4, 2008.
[8] Konstantinidis, K.T. and Tiedje, J.M., Trends between gene content and genome size in prokaryotic species with larger genomes, *Proc Natl Acad Sci U S A*, 101(9):3160–3165, 2004.
[9] Konstantinidis, K.T., Ramette, A., and Tiedje, J.M., The bacterial species definition in the genomic era, *Philos Trans R Soc Lond B Biol Sci*, 361(1475):1929–1940, 2006.
[10] Koonin, E.V., Makarova, K.S., and Aravind, L., Horizontal gene transfer in prokaryotes: quantification and classification, *Annual Review Microbiology*, 55:709–742, 2001.
[11] Merhej, V., Royer-Carenzi, M., Pontarotti, P., and Raoult, D., Massive comparative genomic analysis reveals convergent evolution of specialized bacteria, *Biology Direct*, 4:13, 2009.
[12] NCBI Microbial Genome Project, Available from: `http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi`.
[13] Paul, S., Bag, S., Das, S, Harvill, E., and Dutta, C., Molecular signature of hyper saline adaptation: insight from genome and proteome composition of halophilic prokaryotes, *Genome Biology*, 9(4):R70, 2008
[14] Suen, G., Goldman, B., Welch, R., Predicting prokaryotic ecological niches using genome sequence analysis, *PLoS ONE*, 2(1):e473, 2007.
[15] The Gene Ontology Consortium, Gene ontology: tool for the unification of biology, *Nature Genet.*, 25(1):25–29, 2000.
[16] Todar, K., *Todar's Online Textbook of Bacteriology*, Available from: `http://textbookofbacteriology.net/`, 2008.
[17] Wu, H., Mao, F., Su, Z., Olman, V., Xu, Y., Prediction of functional modules based on gene distributions in microbial genomes, *Genome Informatics*, 16(2):247–259, 2005.
[18] Wu, H., Su, Z., Mao, F., Olman, V., Xu, Y., Prediction of functional modules based on comparative genome analysis and Gene Ontology application, *Nucleic Acids Res.*, 33(9):2822–2837, 2005.
[19] Wu, X., Zhu, L., Guo, J., Zhang, D.-Y., Lin, K., Prediction of yeast protein-protein interaction network: insights from the Gene Ontology and annotations, *Nucleic Acids Res.*, 34(7):2137–2150, 2006.