Mapping genomic features to functional traits through microbial whole genome sequences

Wei Zhang

Department of Computer Science and Engineering,
University of Notre Dame,
Notre Dame, IN 46556, USA
Email: wzhang7@nd.edu

Erliang Zeng*

Department of Computer Science and Engineering,
University of Notre Dame,
Notre Dame, IN 46556, USA
and
Eck Institute for Global Health,
University of Notre Dame,
Notre Dame, IN 46556, USA
Email: ezeng@nd.edu
*Corresponding author

Dan Liu and Stuart E. Jones

Department of Biological Science,
University of Notre Dame,
Notre Dame, IN 46556, USA
Email: dliu1@nd.edu
Email: sjones20@nd.edu

Scott Emrich

Department of Computer Science and Engineering,
University of Notre Dame,
Notre Dame, IN 46556, USA
and
Eck Institute for Global Health,
University of Notre Dame,
Notre Dame, IN 46556, USA
Email: semrich@nd.edu

Abstract: Recently, the utility of trait-based approaches for microbial communities has been identified. Increasing availability of whole genome sequences provide the opportunity to explore the genetic foundations of a variety of functional traits. We proposed a machine learning framework to quantitatively link the genomic features with functional traits. Genes from

bacteria genomes belonging to different functional traits were grouped to Cluster of Orthologs (COGs), and were used as features. Then, TF-IDF technique from the text mining domain was applied to transform the data to accommodate the abundance and importance of each COG. After TF-IDF processing, COGs were ranked using feature selection methods to identify their relevance to the functional trait of interest. Extensive experimental results demonstrated that functional trait related genes can be detected using our method. Further, the method has the potential to provide novel biological insights.

Keywords: microbial; functional traits; genomic signatures; sporulation; feature selection; machine learning; phenotype-genotype association; microbial diversity; functional genomics.

Reference to this paper should be made as follows: Zhang, W., Zeng, E., Liu, D., Jones, S.E. and Emrich, S. (2014) 'Mapping genomic features to functional traits through microbial whole genome sequences', *Int. J. Bioinformatics Research and Applications*, Vol. 10, Nos. 4/5, pp.461–478.

Biographical notes: Wei Zhang is currently a research engineer at Adchemy Inc. He received his PhD in 2013 from the Department of Computer Science and Engineering at the University of Notre Dame. His research interests include bioinformatics, data mining and machine learning.

Erliang Zeng is currently a research assistant professor in the Department of Computer Science and Engineering at the University of Notre Dame (UND). Prior to joining UND, he worked as a postdoctoral associate at the University of Miami. He received a PhD degree in Computer Science from the Florida International University in 2008, and a MS degree in Biochemistry and Molecular Biology from the Shanghai Jiao Tong University in 2001. His research interests are in the areas of bioinformatics, computational biology, and biological big data mining.

Dan Liu is currently a research assistant in the Taub Institute at the Columbia University. Before then she was a graduate student in the Department of Biological Sciences at the University of Notre Dame from 2011 to 2013.

Stuart E. Jones is an ecologist that seeks to link microbial genetics, physiology, and ecology to ecosystem processes. He received his PhD in Limnology and Marine Sciences from the University of Wisconsin-Madison. He is currently an assistant professor in the Department of Biological Sciences at the University of Notre Dame.

Scott Emrich received the BS degree in Biology and Computer Science from the Loyola College in Maryland and the PhD degree in Bioinformatics and Computational Biology from the Iowa State University. His research interests include computational biology, bioinformatics and parallel computing with a focus on arthropod genome analysis with applications to global health and ecology. As of 2007 he has been a member of the Computer Science and Engineering department at the University of Notre Dame.

This paper is a revised and expanded version of a paper entitled 'A machine learning framework for trait based genomics' presented at the '2nd IEEE International Conference on Computational Advances in Bio and Medical Sciences (ICCABS)', Las Vegas, NV, USA, 23–25 February 2012.

1 Background

Microbes are the most abundant and diverse biota on Earth. Universal trees of life demonstrate that microbial archaea, bacteria and eukaryotes constitute the vast majority of life's diversity (Ciccarelli et al., 2006). These diverse organisms perform many important ecological functions across a wide range of natural and man-made environments, such as: they regulate global biochemical cycles, influence human health and are responsible for the fate of contaminants in the environment. Despite their importance, we do not understand what maintains microbial diversity, how microbial communities are assembled or how microbial species will adapt to survive and thrive in the face of global change.

In this paper, we present novel bioinformatics approaches to begin forging quantitative links between genetic, taxonomic and functional aspects of microbial diversity and therefore help understand implications of microbial diversity and activity for important ecosystems. We leverage functional traits in our work (McGill et al., 2006), which are definable and measurable properties of an organism (or a group of organisms). Functional traits can strongly influence an organism's fitness, and examples include basal metabolic rate, seed or egg size, adult body mass and potential photosynthetic rate (see Green et al., 2008; McGill et al., 2006). Microbes are ideal candidates for a trait-based approach because emerging culturing approaches expand the diversity of cultivable microorganisms, and short generation times and small sizes facilitate rapid functional trait characterisation.

Here, we provide tools to address the growing interest in understanding relationships between functional traits and underlying genetic blueprint of a genome. Studying genotype-phenotype relationships is a fundamental problem in evolutionary biology and has vast societal importance. This information then can be used to help predict microbial response to climate change, and other human impacts. One significant challenge, however, is the absence of any sort of microbial trait database that collates links between genome features and ecological functions.

Any attempt to develop a microbial trait-based ecology, therefore, must begin with an initial characterisation of the linkage between genomic features and a functional trait of interest. Trying to accomplish this, we applied machine learning techniques to derive associations between bacterial genome content and observed functional traits directly from whole genome sequencing data. Specifically, given a data set containing bacteria genomes with a known functional trait of interest, we aim to identify genes whose presence/absence correlate with the known functional trait. The development of quantitative links between genome content and observed traits derived from reference strains would enable prediction of functional potential of yet-to-be cultured bacteria through culture-independent genome sequencing (i.e. metagenomics or single amplified genome sequencing). Further, we could use the identified gene(s) to characterise individuals, populations and communities, which is increasingly important given concomitant advances in high-throughput sequencing.

Several approaches have analysed genotype-phenotype association by linking genes to a particular phenotype, e.g. endospore formation, gram stain, motility or oxygen requirement (Jim et al., 2004; Kastenmüller et al., 2009; Liu et al., 2006; Slonim et al., 2006; Tamura and D'haeseleer, 2008). Lingner et al. (2010) proposed an approach to

predict microbial phenotypes based on discriminative learning from protein domain frequencies. The genomic features extracted from genome sequences and used by these methods are from the mapping between genomic content to existing databases such as NCBI COG database (Tatusov et al., 1997), metabolic pathway (Kanehisa and Goto, 2000) and protein domains (Finn et al., 2005). While providing successful hints for linking genes to a particular functional trait, currently available methods have several limitations arising from the incomplete and sometimes biased knowledge presents in these databases. Our approach, on the other hand, provides *de novo* predictions of a species' functional trait from whole genome features and only uses annotation as the last step to help interpret results. Thus, all detectable genomic features from emerging genomes can be linked to functional traits, which will be increasingly important as sequences are rapidly generated from a growing number of microbial organisms of interest. We have also included weighted network analysis in our framework to reveal intrinsic interactions among the genetic factors that affect the functional trait under-investigation, which, to the best of our knowledge, it is the first time that weighted network techniques have been used in a genomic-trait association study.

2 Results

Microorganisms have different strategies to resist environment stress, and dormancy is a common and important functional trait in soil bacterial communities. Upwards of 90% of bacterial cells and 50% of bacterial populations can be dormant in a sample of soil (Lennon and Jones, 2011). The formation of an environmentally resistant spore is one, and perhaps the most famous, means of a bacteria being dormant. As most of the endospore-forming bacteria are soil organisms (Moir and Smith, 1990), we acquired the complete genomic sequences of 100 soil bacteria from NCBI ftp (ftp://ftp.ncbi.nih. gov/genbank/genomes/Bacteria/). Based upon literature information, 46 of them were classified as capable of spore formation while 54 not. Our approach was tested on this data set to computationally identify sporulation-related genes. This was achieved by applying feature selection to the organisms' Cluster of Orthologous Genes (COG). Here, COGs refer to orthologous gene groups obtained directly from results of the computational tool OrthoMCL (Li et al., 2003) on ORFs from the genomes of 100 soil bacteria. It should be noted that these COGs are different from those as present in the NCBI COG database (Tatusov et al., 1997), which contains clusters from 66 unicellular organisms.

There were almost half a million genes in total from the data set ($n = 431{,}168$). OrthoMCL (Li et al., 2003) was applied to cluster genes with similar functions (see methods in Section 4) into 42,029 COGs. In this scenario, each genome is represented as a data vector of size 42,029, with each element being the counts of genes from that genome belonging to a corresponding COG. So the entire data set can be formalised as a data matrix with size $n \times m$, where n represents the number of total COGs and m indicates the number of genomes.[1] Next, the TF-IDF technique was used to weight each element in the matrix and feature selection was used to rank COGs (see methods in Section 4). Top ranked COGs were expected to be more correlated with the sporulation functional trait.

To evaluate performance, two well-studied classification methods: SVM and Naive Bayes were used to classify the data using selected COGs. Classification precision was

recorded according to a tenfold cross validation. The larger the precision, the better the prediction power of the selected COGs is. To see how feature selection could influence classification performance, all of the COGs were first used to build a model and classify the data (i.e. using no feature selection at all). The average precision was 73.78% using the SVM classifier and 87.40% using the Naive Bayes classifier, respectively. A small subset of ranked COGs obtained from feature selection methods were then used to train the classifier and evaluate prediction performance. The number of selected COGs ranged from 2 to 200.

Figures 1 and 2 show the plots of the average classification precision of two classifiers (SVM and Naive Bayes) versus variant number of COGs selected using three feature selection methods: mRMR, InfoGain and RFE (for details see the methods Section 4), and versus all COGs without any feature selection (black solid lines). As expected, using a small subset of top-ranked features achieves better prediction power. This is more apparent for the SVM classifier because it tends to perform poorly when a lot of noisy features are included. In both figures, precision increases gradually as more features are used and then begins to saturate. Among the three feature selection methods, RFE performs the best and InfoGain is the worst in this analysis. The performance of mRMR is comparable to that of RFE when the feature size is small (less than 50), which is consistent with observations from our previous work (Zhang et al., 2010). Considering the computing intensity of RFE is substantial when the number of features is large, we believe mRMR can be a good initial choice given its decent performance and much less computing time.

Figure 1 Precision curves using SVM classifier. Average classification precision of tenfold cross validation using an SVM with different numbers of ranked COGs. The blue, green and red plots are results using variant number of COGs ranked by feature selection methods mRMR, InfoGain and RFE, respectively. The black solid line is the average precision of SVM classifier without feature selection

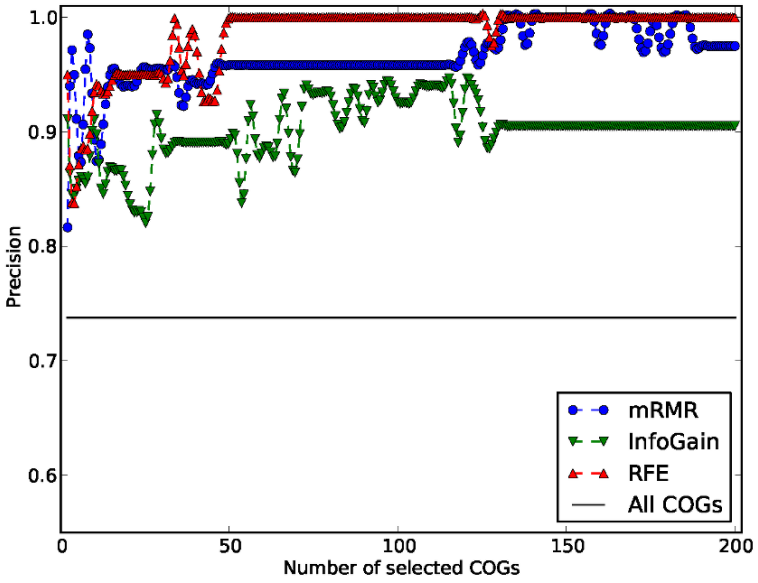


Figure 2 Precision curves using Naive Bayes classifier. Average classification precision of tenfold cross validation using Naive Bayes with different numbers of ranked COGs. The blue, green and red plots are results using variant number of COGs ranked by feature selection methods mRMR, InfoGain and RFE feature selection, respectively. The black solid line is the average precision of Naive Bayes classifier without feature selection

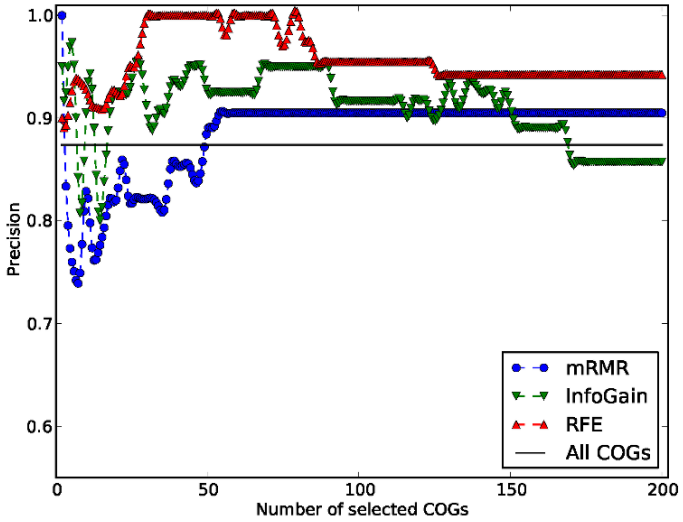


Figure 3 shows the associated Receiver Operator Characteristics (ROC) curves and area under ROC curve (AUC) of the three feature selection methods. The results were obtained from a tenfold cross validation process on the data using an SVM classifier and the top 50 ranked COGs from each of the three-feature selection methods.

2.1 Investigating the biological relevance of discriminative COGs

From the bacterial sporulation literature, we identified 135 important sporulation-related genes (from here on referred to as 'sporulation genes'). This list includes genes involved in sporulation initiation, such as *spo0A*, *sigmaE*, *sigmaK*, *sigmaF*, and *sigmaG*; some that control or regulate the germination of spores, such as *gerAA*, *gerAB*, *gerAc*, *gerBA*, *gerBB*, and *gerK*; and others that help spores resist heat and UV photochemistry, such as *sspA*, *sspB* and *spoVB* (Stragier and Losick, 1996). To assess the ability of our method to identify biologically relevant features, we compared the list of 'sporulation genes' to the list of ranked COGs. HMMER (Finn et al., 2011) was used to locate homology between 'sporulation genes' and genes in ranked COGs, while the Pfam scan tool (Finn et al., 2005) was used to annotate ranked COGs. Table 1 presents the top ten COGs ranked by the feature selection method mRMR, the homologous 'sporulation genes' each of them mapped to, and their Pfam annotations. Among the ten COGs, five can be directly mapped to at least one of the 'sporulation genes' identified in the literature.

Figure 3 ROC curves and AUC using SVM classifier. Receiver Operator Characteristics (ROC) curves and associated area under ROC curve (AUC) representing the classification performance of tenfold cross validation using the top 50 ranked COGs selected from the three feature selection methods. Axes are limited to minimum 50% true positive ratio and maximum 50% false positive ratio

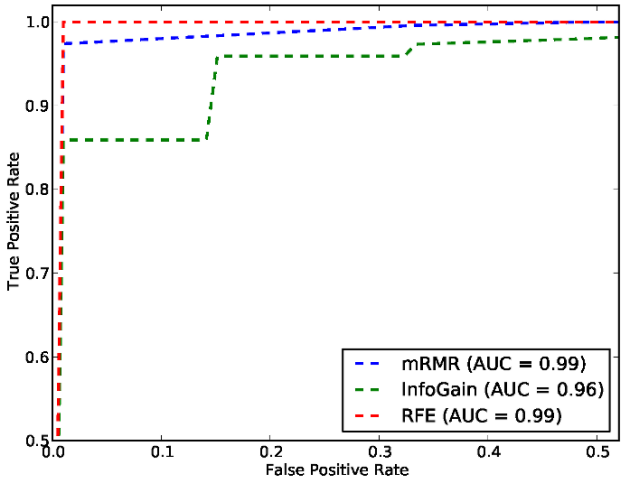


Table 1 Top ranked COGs mapped to sporulation genes. Top ten ranked COGs mapped to sporulation genes and their Pfam annotation. The information in the parenthesis of Pfam annotation column is Pfam clan ID. A clan contains two or more Pfams that have originated from a single evolutionary origin (Finn et al., 2005). COGs that have known sporulation gene mapping are subject to enrichment analysis. The *p*-value tells if the mapped sporulation genes are enriched in the corresponding COG or not

COG Rank	*Mapped to Known Sporulation Gene*	*Pfam Annotation*	*p-value*
1	sigA, spoIIIC, sigmaE, sigK, sigmaF, spoIVCB, sigmaG, sigmaH	Helix-turn-helix (CL0123)	1.63e-17
2	N/A	NADP_Rossmann (CL0063)	N/A
3	spo0KE, spo0KD	P-loop_NTPase (CL0023)	2.95e-04
4	N/A	SGNH_hydrolase (CL0264)	N/A
5	spoIIAA	STAS (CL0502)	1.78e-02
6	N/A	Helix-turn-helix (CL0123)	N/A
7	spo0KE, spo0KD	P-loop_NTPase (CL0023)	5.24e-05
8	N/A	Carbonic anhydrase	N/A
9	spo0KE, spo0KD	P-loop_NTPase (CL0023)	5.24e-05
10	N/A	Zn_Beta_Ribbon (CL0167)	N/A

Given that biological processes can be complex and involve multiple genes, the unmapped five COGs may indirectly participate in the sporulation process or may not yet be identified as sporulation genes. Many sporulation proteins belong to a DNA-binding family that is essential for gene expression of the mother-cell compartment during sporulation. For example, the Helix-Turn-Helix (HTH), which is a major structural motif capable of binding DNA to regulate gene expression, has been found in ‘sporulation

genes' that correspond to our top ranked COGs. Interestingly, the sixth unknown COG in Table 1 also has HTH domain according its Pfam annotation suggesting this COG could contain novel sporulation genes. Another example is the tenth COG that shares a zinc-binding ribbon motif with spoVE, the stage V sporulation protein E (http://projects. biotec.tu-dresden.de/memotif/en/L-[IMV]-L-x-[LV]-[ILV]-x-[GSV]-[AIV]-[AGL]). Finally, the second COG shares the same Pfam clan group (CL0063) with the human MT-A70 protein; MT-A70-like proteins such as the yeast IME4 protein are important for inducing sporulation in other species. See supplementary information[1] for the Pfam annotations for top ranked 200 COGs. We conclude that top-ranked COGs from our method contain both known 'sporulation genes' and new sporulation gene candidates for experimental validation.

We observed additional COGs with good discriminative power, which can also be used by biologists to generate hypotheses, to validate or to compare to previously conducted biological studies. Because such efforts can likely increase biological understanding, we broadened our analysis by comparing the top n (n ranges from 1 to 200) COGs to 135 known 'sporulation genes'. The evaluation was performed using two metrics: (a) the percentage of top n COGs that are homologous to any of the 135 sporulation genes (referred to as 'recall of COGs') and (b) the percentage of 135 'sporulation genes' that can be mapped to genes contained in the top n selected COGs (referred to as 'recall of sporulation genes'). Figure 4 plots the changes of these two metrics when increasing the number of COGs selected from an mRMR-based ranked list. We observed high 'recall of COGs' when the number of selected COGs n were relatively small, indicating our method was able to find gene markers that were highly related to sporulation. We also saw a relative flat plot of 'recall of sporulation genes', which may result from a lack of a comprehensive microbial trait database; the 135 sporulation genes we curated from literature are not a complete set. As new functional traits revealed by our method are confirmed by wet-lab experiments, these new trait genes can then be deployed and result in a more comprehensive database for future studies.

2.2 Network view of discriminative COGs

COGs identified by feature selection methods are predicted to be the most discriminative among the different functional trait groups. Optimal feature selection, however, requires an exhaustive search of all possible subsets of features of the chosen cardinality, which is computationally hard and impractical. As a result, most feature selection methods rely on certain heuristic strategies, and there remains a need to study the combined effect of multiple features and their correlations. Network analysis has been widely used in microarray data analysis to reveal genes' co-expression patterns, and we used this technique here to investigate the intrinsic interplay of sporulation-related COGs. The weighted network analysis of the top 200 mRMR-based COGs generates four modules of cooccurring COGs (Figure 5). Among these four modules, three contain enriched sets of sporulation genes that are known to be involved in sporulation (Table 2). Another view of the four modules shown in Figure 6 reveals their intrinsic network structure. The turquoise module, which is the most enriched for sporulation-related COGs, has the most dense intra-module edges; while the yellow module, which contains only two sporulation-related COGs, has sparse intra-module relationships. The three modules (turquoise, brown and blue) that have enriched sporulation-related COGs show more inter-module interactions than their interplays with the yellow module, indicating that sporulation genes have more similar cooccurring profiles across genomes. Interestingly,

the yellow module gets involved in a sporulation network only by interacting with the brown module, while the brown module connects the turquoise module and blue module. For the detailed module information for this analysis see supplementary information.[1] Although it is out of the scope of this paper, we believe further analysis of the role the COGs play in the network will further help decipher the sporulation functional trait.

Figure 4 Comparison of top selected COGs to known 'sporulation genes'. Plots of 'recall of COGs' and 'recall of sporulation genes' versus variant number of COGs selected by feature selection method mRMR. The 'recall of COGs' is the percentage of top *n* COGs that are homologous to any of the 135 sporulation genes, and the 'recall of sporulation genes' is the percentage of 135 'sporulation genes' that can be mapped to genes contained the top *n*

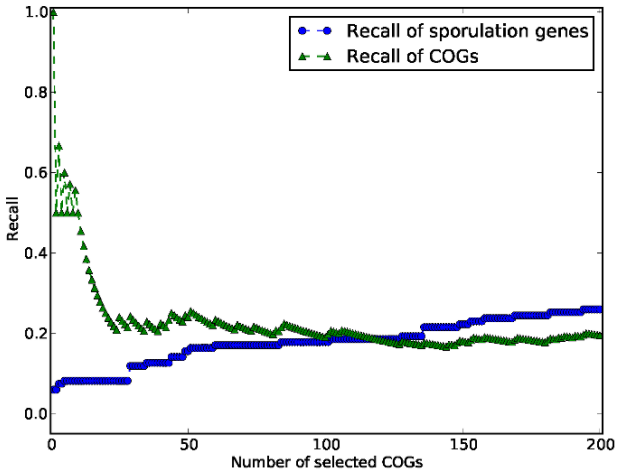


Figure 5 Heatmap of the top 200 selected COGs by method mRMR. The weighted network analysis of the top 200 COGs suggested by the feature selection method mRMR results in four modules of co-occurring COGs

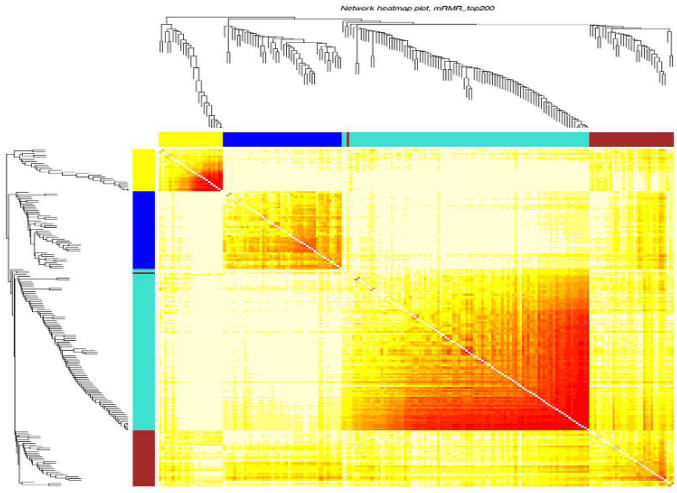


Figure 6 Modules in COG correlation network. Four modules of co-occurring COGs (represented by different colours) in the COG correlation network. Nodes marked with a triangle represent COGs containing 'sporulation genes'. Node labels are COG IDs obtained from method OrthoMCL

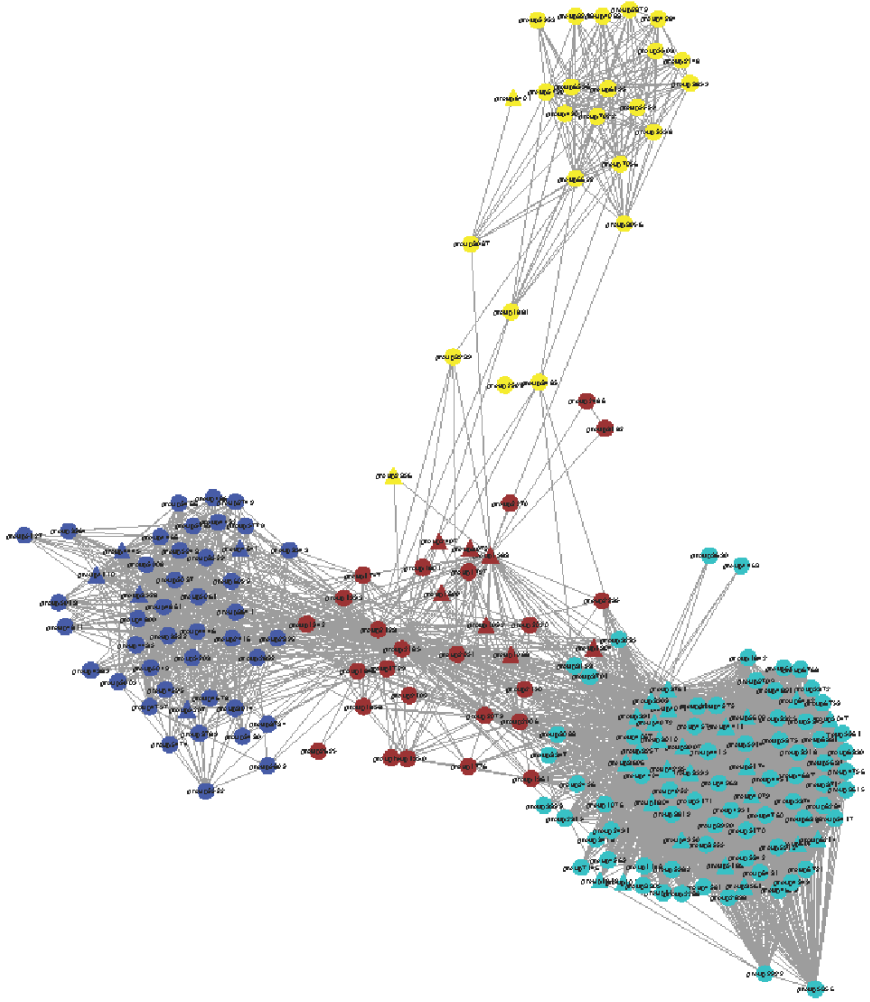


Table 2 Enrichment analysis of sporulation genes in modules. The *p*-values of sporulation genes enriched in four modules were calculated by hypergeometric distribution. The COGs containing 'sporulation genes' is referred to as 'Sporulation COGs'

Module	*# of COGs*	*# of Sporulation COGs*	*p-value*
turquoise	95	25	8.88e-14
brown	34	7	4.32e-4
blue	46	5	4.18e-2
yellow	25	2	0.28

2.3 Evaluation of our approach on other data sets

In our genomic-sporulation association study, orthologous groups obtained from OrthoMCL (Li et al., 2003) were used as genomic features. To further test the effectiveness of our method, we processed data obtained from the work of Lingner et al. (2010) that included 1475 bacterial organisms associated with four functional traits: endospores, gram stain, motility and oxygen requirement. According to the annotation on the NCBI prokaryotic genome project website (ftp://ftp.ncbi.nlm.nih.gov/genomes/genomeprj/lproks_0.txt), Lingner et al. (2010) labelled organisms that were annotated as 'yes' (endospores), '+' (gram stain), 'motile' (motility) or 'aerobic' (oxygen requirement) as positive examples and organisms that are annotated 'no'/'–'/'non-motile'/'anaerobic' as negative examples. In addition, Lingner et al. (2010) created a domain profile for each organism based on domain knowledge from the public Pfam database (Finn et al., 2005). A total of 1475 organisms were divided into two groups: training (1032 organisms) and testing (443 organisms). For each functional trait, we applied our approach on training data to select the top ranked Pfam IDs and evaluated classification performance using SVMs on the test data. Table 3 shows the performance in terms of different metrics and Figure 7 shows the associated ROC curves and AUC. The results reveal that the selected Pfam domains have good predictive power in classifying the test data. While the classification performance using our method is comparable to those as shown in the work of Lingner et al. (2010), the top ten most discriminative Pfam for phenotype category 'endospores' make our method standout (Table 4). Of the top ten domains, all but two with unknown functions are directly involved in sporulation and four of them have function description containing the terms 'spore' or 'sporulation' (Table 4).

Figure 7 ROC curves and AUC using SVM classifier on Pfam domain profiles. ROC curves and associated AUC representing the prediction performance using SVM classifier on Pfam domain profiles

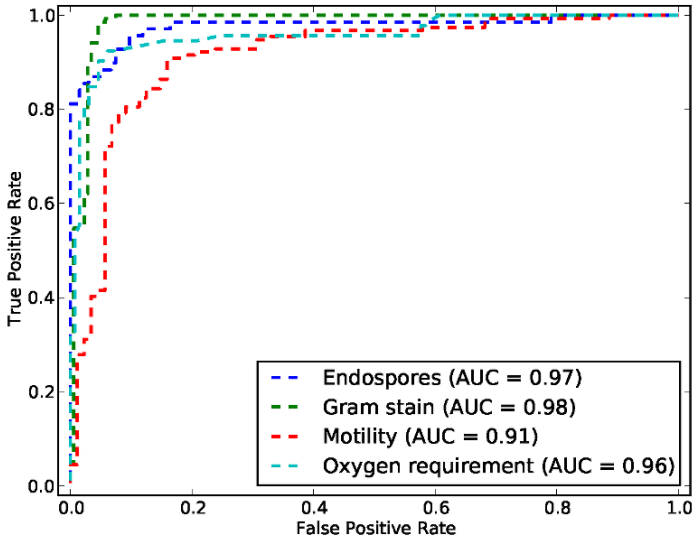


Table 3 SVM classification performance on datasets of pfam domain profiles. SVM classification performance on four functional trait datasets of Pfam domain profiles using top 50 ranked Pfams. The first column indicates the functional trait category, the remaining columns represent the classification performance in terms of different measures

Functional Trait	*Precision*	*Recall*	*F1 score*	*AUC*
Endospores	0.947	0.884	0.906	0.974
Gram stain	0.947	0.943	0.944	0.984
Motility	0.874	0.877	0.876	0.916
Oxygen Requirement	0.925	0.894	0.904	0.942

Table 4 Top ten discriminative Pfam domain families for trait 'Endospores'. List of top ten discriminative Pfam domain families obtained from feature selection method mRMR for the trait "Endospores". The first column indicates the rank. The second column denotes the Pfam family ID. The third column corresponds to the Pfam family description.

Rank	*Pfam Family ID*	*Pfam Description*
1	PF00269	Small, acid-soluble spore proteins, alpha/beta type
2	PF04456	Domain of unknown function (DUF503)
3	PF04026	Stage V sporulation protein G (SpoVG)
4	PF05582	YabG peptidase U57
5	PF00704	Glycosyl hydrolases family 18
6	PF08486	Stage II sporulation protein (SpoIID)
7	PF04070	Domain of unknown function (DUF378)
8	PF05103	DivIVA protein
9	PF03419	Sporulation factor SpoIIGA
10	PF01522	Polysaccharide deacetylase

3 Conclusions

The use of whole genome sequence data allows us to link functional traits with genomic context; however, many challenges remain. In this study, we investigated using machine learning methods to determine genomic features that are correlated to a binary functional trait from a pool of classified bacterial genomes. Any genomic features such as those curated in existing databases (e.g. Pfam domain) could be used in our method. In this paper, we contribute the use of orthologous clusters, which are gene groups with similar functions from multiple genomes and are presumed to have evolved from a common ancestor. One advantage of using orthologous clusters as features is that it makes the comparison among hundreds and thousands of genomes possible and substantially reduces the feature dimensionality. Another contribution is applying a text mining technique to transform the genome-feature matrix, which helps to control the fact that some genomic features are generally more common than others. We tested our framework on several data sets and demonstrated that a selected subset of genomic features has more predictive power than using all the features and known trait-related

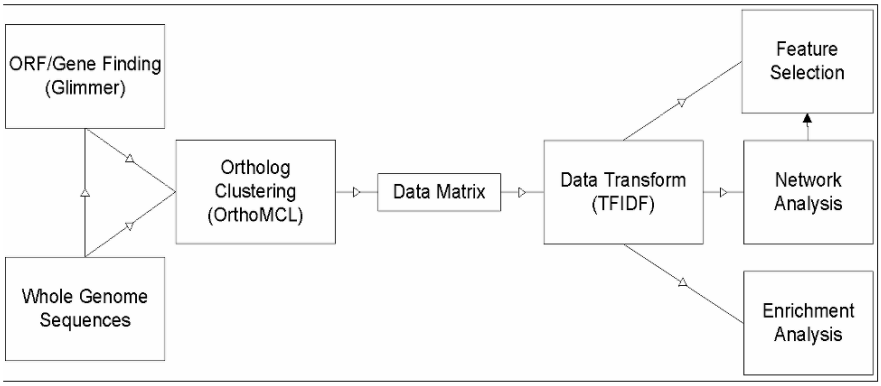
genes were detected. Although we focused on a discrete functional trait, this method can be extended to continuous functional traits. Our method is able to identify informative genomic features that were previously not linked to the functional trait of interest, and we investigated their roles in genotype-phenotype associations through network analysis. These techniques leverage whole genome data with known characteristics to help to better understand the genetic foundation of many important ecological functional traits.

4 Methods

Our goal is to build features from well-characterised reference genomes to help better classify environmental samples from a functional perspective. The same method can be applied to any set of sequenced genomes, the availability of which has improved significantly with the advent of high-throughput sequencing and improved microbial culturing techniques.

To begin developing quantitative links between genome content and functional traits, reference genomes with known functional trait group labels were used as training data. Supervised machine learning techniques were then used to develop predictive models from these training data, or more simply which genes (features) could distinguish one group from another. Figure 8 illustrates the flowchart of our proposed framework. The details are explained as follows.

Figure 8 Flowchart. Framework flowchart for trait based microbial genomics method



4.1 Generate features

Each training datapoint is represented with a set of input features and a qualitative-dependent trait variable giving the correct class label of that datapoint. It is important to identify what to use as features to train the model. Intuitively, one could use genes alone as features. Different bacterial species, however, contain up to thousands of genes, which could lead to millions of features that could make the learning process intractable. In addition, gene-based methods can be deteriorated by simple genomic rearrangements

such as gene fusion/fission events or domain shuffling (Gabaldón et al., 2009). Further, different genes may have similar functions. To address these concerns we chose to use predicted orthologs, which are genes presumed to have evolved from a common ancestor by speciation and therefore likely retain similar functions, as our basic feature. Specifically, we used OrthoMCL (Li et al., 2003), which is a BLAST-driven method, to identify putative Clusters of Orthologous Groups (COG) based on protein sequences, and used COGs as features for training data. OrthoMCL constructs sets of orthologous genes based on sequence similarity search results for all versus all genome searches (Li et al., 2003). If the gene annotation data was not available, Glimmer (Delcher et al., 2007) was used to determine putative Open Reading Frames (ORFs) that were then translated to protein sequences for inclusion in OrthoMCL. Glimmer is a system for finding genes in microbial sequences and it uses interpolated Markov models to identify the coding regions and distinguish them from non-coding DNA (Delcher et al., 2007).

4.2 Data weighting and transformation

Instead of simply assigning a binary value to a genome by virtue of its presence/absence in a COG, we counted the number of genes from a genome that belong to each COG. The data is represented using a matrix, as shown below:

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}$$

where X is the count matrix with n COGs (features) and m species, and x_{ij} denotes the total number of genes assigned to COG i in species genome j. The goal of our proposed approach is identifying COGs with the most discriminative power among analysed genomes with different functional traits.

We applied the Term Frequency-Inverse Document Frequency (TF-IDF) approach (Wu et al., 2008) in text mining to weight each element in matrix X. TF-IDF is a widely used information retrieval technique to measure the amount of information of a term weighted by its occurrence probability. TF-IDF works by determining the relative frequency of terms in a specific document compared to the inverse proportion of that term over the entire document corpus. Intuitively, this calculation determines how relevant a given term is in a particular document. We viewed each species genome in this paper as a document and each COG as a term.

The TF-IDF measure consists of two components. First, we calculate the term frequency (TF), which is usually defined as the word frequency divided by the total number of words in the document. In our case, TF is the number of genes in one COG divided by the total number of genes in all COGs in the genome j. We also used a modified version of the TF calculation method as below:

$$TF_{ij} = 1 + \log x_{ij} \tag{1}$$

The intuition behind the second part, IDF, is to down-weight the terms that appear in many documents since they tend to be less discriminative. Here, to adjust for the fact that some COGs might appear in many or all genomes in our training data, we computed IDF as:

$$IDF_i = \log \frac{N}{\sum_{j=1}^{m} n_{ij}} \tag{2}$$

$$n_{ij} = \begin{cases} 1 & \text{if } x_{ij} > 0 \\ 0 & \text{if } x_{ij} = 0 \end{cases} \tag{3}$$

where N is the total number of genomes.

Combining the above two measures, each element x_{ij} in initial matrix M is weighted and transformed to a continuous value as follows:

$$TF - IDF_{ij} = TF_{ij} * IDF_i \tag{4}$$

4.3 Feature selection

A common challenge in biological data mining is small sample size combined with high-dimensional input features. Even if we have already reduced the dimension of available features by grouping genes with similar functions into COGs, the number of features (on the order of thousands) is still very large compared to the number of species (on the order of tens or hundreds in our preliminary work). Examples of this type of problem can be seen in the various applications of supervised classification approaches to microarray data (Lee et al., 2005), in which the goal is to identify a small, but highly predictive subset of genes for further investigation from up to tens of thousands of candidates in the case of higher eukaryotic organisms. Feature selection can be employed along with classifier construction to avoid over-fitting, to generate more reliable classifier and to provide more insights into the underlying causal relationships (Lee et al., 2005; Zhang et al., 2010; Zhang et al., 2012).

There are typically three categories of feature selection solutions: filter methods, wrapper methods and embedded methods. Filter methods perform a univariate test or multivariate test to rank each feature independent of the model construction step; wrapper methods evaluate various possible feature subsets by training and testing on a specific classification model, while embedded methods build the process of searching an optimal subset of features into classifier construction. We focus on filter and wrapper methods in our work. Wrapper methods often achieve better performance but are more computationally intensive. Here, we applied three-feature selection methods: Information Gain (InfoGain; Elomaa and Rousu, 1999), minimum Redundancy Maximum Relevance (mRMR; Ding and Peng, 2005) and Recursive Feature Elimination (RFE; Guyon et al., 2002). The first two are filter methods: InfoGain ranks features by measuring the decrease in entropy when the feature is given versus absent, and mRMR selects features through the calculation of mutual information. RFE is a wrapper method that uses a Support Vector Machine (SVM) classifier as the evaluator.

4.4 Network analysis

Network analysis was performed using R package WGCNA (Langfelder and Horvath, 2008) on the 200 presence/absence COG profiles selected by method mRMR for the sporulation data. The package WGCNA was originally designed for weighted gene co-expression network analysis of microarray data. The package uses a topological overlap measure (Ravasz et al., 2002) to construct weighted correlation networks. We applied it in this study to analyse COG profiles and to detect COG cooccurring modules, which are defined as group of highly correlated COGs. The network whose topological overlap is above the threshold of 0.10 was exported to Cytoscape (Shannon et al., 2003), a widely used network visualisation software platform.

4.5 Functional enrichment analysis

The hypergeometric distribution is used for functional enrichment analysis such as *p*-values (as shown in Tables 1 and 2) calculation. The hypergeometric distribution is given by

$$h(k|N,M,n)=\frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}. \tag{5}$$

For COG enrichment analysis (Table 1), N represents the total number of genes in all COGs, n is the number of genes in a COG of interest, k denotes the number of trait relevant genes in the COG of interest, and M corresponds to total number of trait relevant genes in all COGs. For module enrichment analysis (Table 2), N represents the total number of COGs, n is the number of COGs in a module of interest, k denotes the number of COGs containing trait relevant genes in the module of interest, and M indicates the total number of COGs containing trait relevant genes in the entire data set. Given hypergeometric distribution as shown in equation (5), the p-value for COG enrichment analysis (Table 1) and module enrichment analysis (Table 2) can be calculated as follows:

$$p(x\geq k)=\sum_{x=k}^{n}h(x|N,M,n). \tag{6}$$

Acknowledgements

This work has been supported by a National Research Initiative Grants (2011-67019-30225) from the USDA National Institute of Food and Agriculture (to SJ, SE collaborator) and in part by the strategic University of Notre Dame investments in Global Health and Environmental Change research (to SE). Publication of this paper was supported by strategic University of Notre Dame investments in Global Health and Environmental Change research.

References

Ciccarelli, F.D., Doerks, T., von Mering, C., Creevey, C.J., Snel, B. and Bork, P. (2006) 'Toward automatic reconstruction of a highly resolved tree of life', *Science*, Vol. 311, pp.1283–1287.

Delcher, A.L., Bratke, K.A., Powers, E.C. and Salzberg, S.L. (2007) 'Identifying bacterial genes and endosymbiont DNA with Glimmer', *Bioinformatics*, Vol. 23, pp.673–679.

Ding, C. and Peng, H. (2005) 'Minimum redundancy feature selection from microarray gene expression data', *Journal of Bioinformatics and Computational Biology*, Vol. 3, No. 2, pp.185–205.

Elomaa, T. and Rousu, J. (1999) 'General and efficient multisplitting of numerical attributes', *Machine Learning*, Vol. 36, No. 3, pp.1–49.

Finn, R.D., Clements, J. and Eddy, S.R. (2011) 'HMMER web server: interactive sequence similarity searching', *Nucleic Acids Research*, Vol. 39, pp.29–37.

Finn, R.D., Mistry, J., Schuster-Böckler, B., Griffths-Jones, S., Hollich, V., Lassmann, T., Moxon, S., Marshall, M., Khanna, A., Durbin, R., Eddy, S.R., Sonnhammer, E.L.L. and Bateman, A. (2005) 'Pfam: clans, web tools and services', *Nucleic Acids Research*, Vol. 34, pp.D247–D251.

Gabaldón1, T., Dessimoz, C., Huxley-Jones, J., Vilella, A.J., Sonnhammer, E.L. and Lewis, S. (2009) 'Joining forces in the quest for orthologs', *Genome Biology*, Vol. 10, No. 9, 403p.

Green, J.L., Bohannan, B.J.M. and Whitaker, R.J. (2008) 'Microbial biogeography: from taxonomy to traits', *Science*, Vol. 320, No. 5879, pp.1039–1043.

Guyon, J.W., Barnhill, S. and Vapnik, V. (2002) 'Gene selection for cancer classification using support vector machines', *Machine Learning*, Vol. 46, pp.389–422.

Jim, K., Parmar, K., Singh, M. and Tavazoie, S. (2004) 'A cross-genomic approach for systematic mapping of phenotypic traits to genes', *Genome Research*, Vol. 14, pp.109–115.

Kanehisa, M. and Goto, S. (2000) 'KEGG: Kyoto Encyclopedia of genes and genomes', *Nucleic Acids Research*, Vol. 28, No. 1, pp.27–30.

Kastenmüller, G., Schenk, M.E., Gasteiger, J. and Mewes, H.W. (2009) 'Uncovering metabolic pathways relevant to phenotypic traits of microbial genomes', *Genome Biology*, Vol. 10, pp.R28.

Langfelder, P. and Horvath, S. (2008) 'WGCNA: an R package for weighted correlation network analysis', *BMC Bioinformatics*, Vol. 9, No. 1, pp.559+.

Lee, J., Park, M. and Song, S. (2005) 'An extensive comparison of recent classification tools applied to microarray data', *Computational Statistics & Data Analysis*, Vol. 48, No. 4, pp.869–885.

Lennon, J.T. and Jones, S.E. (2011) 'Microbial seed banks: the ecological and evolutionary implications of dormancy', *Nature Reviews Microbiology*, Vol. 9, pp.119–130.

Li, L., Stoeckert Jr., C.J. and Roos, D.S. (2003) 'OrthoMCL: identification of ortholog groups for eukaryotic genomes', *Genome Research*, Vol. 13, pp.2178–2189.

Lingner, T., Mühlhausen, S., Gabaldón, T., Notredame, C. and Meinicke, P. (2010) 'Predicting phenotypic traits of prokaryotes from protein domain frequencies', *BMC Bioinformatics*, Vol. 11, pp.481.

Liu, Y., Li, J., Sam, L., Goh, C.S., Gerstein, M. and Lussier, Y.A. (2006) 'An integrative genomic approach to uncover molecular mechanisms of prokaryotic traits', *PLOS Computational Biology*, Vol. 2, e159p.

McGill, B., Enquist, B., Weiher, E. and Westoby, M. (2006) 'Rebuilding community ecology from functional traits', *Trends in Ecology & Evolution*, Vol. 21, No. 4, pp.178–185.

Moir, A. and Smith, D.A. (1990) 'The genetics of bacterial spore germination', *Annual Review of Microbiology*, Vol. 44, No. 1, pp.531–548.

Ravasz, E., Somera, A.L., Mongru, D.A., Oltvai, Z.N. and Barabási, A.L. (2002) 'Hierarchical organization of modularity in metabolic networks', *Science*, Vol. 297, No. 5586, pp.1551–1555.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B. and Ideker, T. (2003) 'Cytoscape: a software environment for integrated models of biomolecular interaction networks', *Genome Research*, Vol. 13, No. 11, pp.2498–2504.

Slonim, N., Elemento, O. and Tavazoie, S. (2006) 'Ab initio genotype-phenotype association reveals intrinsic modularity in genetic networks', *Molecular Systems Biology*, Vol. 2, No. 1.

Stragier, P. and Losick, R. (1996) 'Molecular genetics of sporulation in Bacillus subtilis', *Annual Review of Genetics*, Vol. 30, pp.297–341.

Tamura, M. and D'haeseleer, P. (2008) 'Microbial genotype–phenotype mapping by class association rule mining', *Bioinformatics*, Vol. 24, pp.1523–1529.

Tatusov, R.L., Koonin, E.V. and Lipman, D.J. (1997) 'A genomic perspective on protein families', *Science*, Vol. 278, No. 5338, pp.631–637.

Wu, H.C., Luk, R.W.P., Wong, K.F. and Kwok, K.L. (2008) 'Interpreting TF-IDF term weights as making relevance decisions', *ACM Transactions on Information Systems*, Vol. 26, No. 3, pp.1–37.

Zhang, W., Emrich, S.J. and Zeng, E. (2010) 'A two-stage machine learning approach for pathway analysis', *Proceedings of IEEE International Conference on Bioinformatics and Biomedicine*, pp.274–279.

Zhang, W., Zeng, E., Liu, D., Jones, S. and Emrich, S.J. (2012) 'A machine learning framework for trait based genomics', *Proceedings of IEEE 2nd International Conference on Computational Advances in Bio and Medical Sciences*, pp.1–6.

Note

1 This information is available by contacting authors.