Published in final edited form as:

Methods. 2017 October 01; 129: 8–17. doi:10.1016/j.ymeth.2017.04.018.

Functional Association Prediction by Community Profiling

Dazhi Jiao^{a,b}, Wontack Han^{a,b}, and Yuzhen Ye^{a,*}

^a150 S. Woodlawn Ave, Indiana University, Bloomington, IN 47405

Abstract

Recent years have witnessed unprecedented accumulation of DNA sequences and therefore protein sequences (predicted from DNA sequences), due to the advances of sequencing technology. One of the major sources of the hypothetical proteins is the metagenomics research. Current annotation of metagenomes (collections of short metagenomic sequences or assemblies) relies on similarity searches against known gene/protein families, based on which functional profiles of microbial communities can be built. This practice, however, leaves out the hypothetical proteins, which may outnumber the known proteins for many microbial communities. On the other hand, we may ask: what can we gain from the large number of metagenomes made available by the metagenomic studies, for the annotation of metagenomic sequences as well as functional annotation of hypothetical proteins in general? Here we propose a community profiling approach for predicting functional associations between proteins: two proteins are predicted to be associated if they share similar presence and absence profiles (called community profiles) across microbial communities. Community profiling is conceptually similar to the phylogenetic profiling approach to functional prediction, however with fundamental differences. We tested different profile construction methods, the selection of reference metagenomes, and correlation metrics, among others, to optimize the performance of this new approach. We demonstrated that the community profiling approach alone slightly outperforms the phylogenetic profiling approach for associating proteins in species that are well represented by sequenced genomes, and combining phylogenetic and community profiling further improves (though only marginally) the prediction of functional association. Further we showed that community profiling method significantly outperforms phylogenetic profiling, revealing more functional associations, when applied to a more recently sequenced bacterial genome.

Keywords

guilt-by-association; functional association prediction; phylogenetic profiling; community profiling; metagenomics

*Corresponding author: yye@indiana.edu.

^bThese authors contributed equally to this work

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

1. Introduction

The rapid advancement of the new sequencing technology has resulted in the exponential growth of available genome sequences from model and non-model organisms [1]. The massive genomic sequences, however, do not necessarily indicate the accumulation of biological knowledge, as the interpretation of these genomes is a nontrivial task. Although many approaches have been developed for functional annotation integrating different sources of information [2], genome annotation relies heavily on *homology-based inference* [3]: a gene is assigned to a function if it exhibits significantly high sequential similarity with one or more genes with known functions collected in gene (e.g., the NCBI nr) or protein (e.g., Uniprot) databases. As a result, the functional annotation of a newly sequenced genome is limited to the protein-coding genes from well-studied families, and a large fraction of proteins is likely to remain un-annotated (or denoted as hypothetical genes). This is particularly the case for those genomes that are phylogenetically distant from the well-studied model organisms, because the homology search may fail when the homologs are divergent.

Non-homology-based function prediction methods exploit information beyond the gene/protein sequences for functional inference. For example, structure-based function annotation methods look for substructures (e.g., structural motifs or surface patches) that are potentially associated with known functional characteristics of proteins (e.g., biochemical activities or their binding partners) [4, 5]. In addition, many methods, which are referred to as the guilt-by association function prediction techniques, attempt to identify functionally coupled gene pairs that share similar patterns in various contexts [6, 7], including their proximity in the genomes (e.g., in the same operon) [8], the presence/absence of their homologs in genomes across the phylogenetic tree (i.e., the *phylogenetic profiles*) [9], the fusion/fission between them (i.e., the Rosetta Stone proteins) [10], their co-expression under different physiological conditions [11], commonality between their interacting partners [12], and phenotypic effects of their knockout/knockdown mutants [13]. Once the functional coupling of a gene pair is established, the known function of one gene in the pair can be assigned to the other one with unknown function. Over the years, the guilt-by-association methods have been optimized on several aspects, including the design of distance measures, the training process as well as the parameter selections, and how the functions are transferred from gene to gene (simple transfer or network-based approach). It was also shown that the integration of context-based information can improve the sensitivity and the accuracy of function prediction [7].

The *guilt-by-association* methods are useful for microbial genome annotation, as thousands of complete bacterial genomes (many more draft genomes) have been made available, and more importantly, bacterial genomes are gene-dense with genes involved in related biological processes often found in the same genomic neighborhood or operons. Bacterial genomes provide a rich source for deriving context information for functionally coupled gene pairs, including operon structure and the phylogenetic profiles. Furthermore, recently, the NGS techniques have been applied to the direct studies (known as *the metagenomic approach*) of the complex microbial communities consisting of a majority of the microbial species that are not (yet) culturable under current laboratory conditions [14]. Massive datasets of metagenomic sequencing became available from a variety of environments, ranging from

soil [15], ocean [16] and human-associated communities [17], and under different conditions (e.g., normal vs diseased [18]). The human microbiomes are one of the most extensively characterized microbial communities because of several large initiatives including the Metahit project [19] and the NIH Human Microbiome Project (HMP) [17]. Functional analysis of the metagenomes often relies on searching the protein sequences derived from short reads or metagenome assemblies (e.g., using FragGeneScan [20]) against known protein/gene families. As a result, the biological functions of a majority of the genes encoded in these communities remain unknown. Recent metaproteomics projects have also resulted in the identification of many protein sequences, many of which remain unannotated [21, 22, 23].

In this paper, we propose a new guilt-by-association method, the community profiling approach that exploits the community information of the metagenomic sequence datasets for protein function prediction. Given a protein of interest, we represent the presence/absence of its homologs in many metagenomic datasets from different environments (or hosts) as a vector, denoted as the *community profile*. Then, for any pair of proteins, their functional coupling can be measured by the distance between their community profiles. The community profile can be viewed as a generalization of the phylogenetic profile, as the phylogenetic profile is based on the pattern of presence/absence of homologs in individual genomes, while a community can be viewed as a *bag* of many microbial genomes. Likewise, similar community profiles between two genes indicate these genes tend to co-occur in the same communities, and thus they may be functionally coupled. We note that community profiles also provide some complementary information to phylogenetic profiles: due to common co-occurrences of microbial species [24], and the frequent horizontal gene transfer events among microbial species in the same environment [25], the functional association between genes may not be indicated by gene co-occurrences in individual genomes, but by gene co-occurrences in bacterial communities.

We note that our community profiling is fundamentally different from the co-occurrence approaches for inference of metagenomic clusters of genes [26] and for binning of contigs/scaffolds in metagenome assembly [27, 28]. For our approach, it is important to use a collection of diverse metagenomes, whereas for the latter, it is important to use similar metagenomes, for example derived from longitudinal studies or gut metagenomes from different hosts.

We tested our approach using two well-studied species (*E. coli* and *Bacillus subtilis*) and a newly sequenced genome. We demonstrate that the community profiling method alone achieved even slightly better performance than the phylogenetic profiling method on well represented genomes (i.e., *E. coli* and *B. subtilis*), and combining phylogenetic and community profiling methods further improved the prediction. More importantly, we showed that community profiling method outperforms phylogenetic profiling and reveals more functional associations, when it is applied to a more recently sequenced microbial genome (*Prevotella copri*). Considering metagenomic datasets will eventually outnumber the sequenced microbial genomes, we believe our method provides a venue to lift the species boundary for predicting potential functional associations between genes/proteins. The main focus of this paper is to demonstrate the effectiveness of the community profiling method for

predicting functional associations, and to optimize the fundamental design of the method (e.g., the profile construction, the distance measures and selection of metagenomes).

2. Materials and Methods

2.1. Community profiling

Similar to the phylogenetic profiling, the community profiling method is based on the profiles of genes or proteins that represent their presence or absence in a set of metagenomes, each from a microbial community in a specific environment or host (see Figure 1). Strictly speaking, a community profile of a gene is defined as a vector, in which each dimension represents the possibility of the gene to be present in the corresponding metagenome. In this work, we focus on the prediction of functional association between genes that encode for proteins and therefore their protein products. But in principle, the same approach can be applied to other genes as well.

To build the community profile of a given gene/protein sequence (herein referred to as the *query sequence*), RAPSearch2 [29] (other fast similarity search tools including Diamond [30] and MICA [31] can also be utilized) is applied to search the sequence against a set of metagenomes. As described in details below, the set of metagenomes (herein referred to as the *reference metagenomes*) were downloaded from metagenomic data repositories. The *E*-value of the top match of the query sequence in each metagenome is used for the corresponding dimension in the vector. When different distance measures are used, this *E*-value vector can be transformed into different numerical representations through different transformation functions. The transformed vectors or the original *E*-value vectors are then used for distance calculation between pairs of community profiles. This pairwise distance score is finally used to predict the association between the corresponding proteins.

2.2. Data

The reference metagenomes for the community profiling method contain metagenomes of microbial community samples from two widely used repositories: JGI Integrated Microbial Genomes with Microbiome (IMG/M) [1] and the Human Microbiome Project (HMP) [32]. A total of 5671 metagenomes from different environments and hosts, such as soil, water, air or animal guts were downloaded from the JGI Genome Portal [33]. A total of 741 metagenomes were downloaded from the HMP [17]. We only consider the metagenomes with genes assigned to at least 100 functions by myRAST [34], leaving 6287 metagenomes to be further considered.

In order to compare the community profiling with the phylogenetic profiling method, we also assembled a set of reference genomes for building the phylogenetic profiles. A total of 5807 complete bacterial genomes were downloaded from the NCBI ftp site as of October 2016.

Functional links between pairs of proteins in *E. coli* from the publication [35] were used as the gold standard for assessing the performance of both profiling methods. These links were originally obtained from the STRING database of predicted functional associations between proteins [36] and were selected to meet the following requirements: the links have to be

supported by experimental data, curated data or both, and at least one of these evidence channels has a confidence score ≥ 0.9 , the highest confidence score in STRING. The non-interacting pairs were generated by randomly selected pairs of proteins, excluding the pairs found in STRING database with the evidence of interaction at any confidence level. The ratio of interacting and non-interacting pairs was kept as 1:2, but testing datasets with higher ratios for non-interacting pairs were shown to have little impacts on the performance assessment of phylogenetic profiling methods [35]. Using the same criteria, we created a dataset of interacting and non-interacting pairs of proteins for *B. subtilis*, by using interaction data from STRING and sequences from UniProt [37].

The draft genome of *P. copri* DSM 18205 (BACT 893.fa) was downloaded from the DACC website at <http://hmpdacc.org/>.

2.3. Prediction of functional associations

In the scope of this manuscript, the goal of our method is to predict the functional associations between pairs of proteins using the community profiling method. The method was first applied to the prediction of the functional links between the proteins in *E. coli* and *B. subtilis*. To evaluate the performance of the profiling methods, we used the area under curve (AUC) as a measure of prediction accuracy. Using the 10-fold cross validation, we compared the AUCs for different combinations of *E*-value transformations, distance measures and optimization methods to select the one with the best performance.

2.4. E-value transformation functions and distance measures

For community profiling (and phylogenetic profiling), profiles need to be first constructed for the proteins, and then a proper distance measure needs to be used to calculate the distance between proteins according to their profiles. We have tested different ways of constructing profiles (binary or non-binary) based on similarity search results and distance measures for measuring the correlation between the profiles, as shown below.

1. Hamming distance between binary profiles. The Hamming distance was used in the first application of phylogenetic profiling [9] as the distance measure. The Hamming distance is the number of reference genomes that do not have the same absence/presence value (0 or 1). The binary values can be transformed from the *E*-values output by RAPSearch2 by simply testing whether the *E*-value is less than a cutoff and can be considered as an indicator of whether the gene is present in the reference genome. We used the cutoff value $E = 10^{-5}$ to transform the *E*-values to binary values for all the binary profile-based methods discussed in this section.
2. Hypergeometric distribution based approach. Wu and colleagues [38] calculated the distance measure for predicting interactions between pairs of proteins using binary profiles and the hypergeometric distribution, which is the probability distribution of a given number of co-occurrence of a pair of proteins across the set of reference genomes. The distance is then computed as the *P*-value for the number of matches between two profiles being as large as the number of matches under the hypergeometric distribution.

3. Jaccard coefficient between binary profiles. Contrary to the Hamming distance, the Jaccard coefficient only accounts for the similarity introduced by co-occurrences of a pair of proteins and ignores the number of reference genomes that do not contain any protein in the pair [39, 40].
4. Pearson correlation coefficient between profiles of E -value. The Pearson correlation coefficient quantifies the degree of linearity of two E -value vectors and is only zero if there is no correlation [39, 38].
5. Spearman's rank correlation between profiles of E -value. The Spearman's rank correlation coefficient was previously used in ranked phylogenetic profiling methods [41].
6. Mutual information. Mutual information is a widely used statistical correlation measure in phylogenetic profiling [42, 38, 35]. We calculated mutual information based on two types of transformation of the E -values. First, we followed the approach in [42], and transformed E -values with the negative inverse of the logarithm of E -value: $p = -\frac{1}{\log E}$, with values of $p > 1$ set to 1 to avoid logarithm-induced problems. Our second transformation function is the negative of logarithm of E -values: $p = -\log E$, and similarly p is set to 1 when $p > 1$. We have found that the second transformation function outperforms the first. The mutual information score is then converted to a distance measure as following [43]:

$$d = \sqrt{1 - e^{-2\delta}},$$

where δ is the mutual information.

2.5. Selection of reference metagenomes

In phylogenetic profiling, various approaches have been proposed to reduce the bias introduced by the un-even phylogenetic distances among the reference genomes [36, 44, 45, 35]. One can also reduce the bias automatically by using machine learning (ML) based dimension reduction algorithms such as the genetic algorithm. It was shown that in phylogenetic profiling, the tree-guided selection methods are better in terms of improving the prediction accuracy [35].

We built a hierarchical tree for the metagenomes by clustering them based on their functional similarities, since the goal is to develop a tool for predicting functional associations. Putative proteins were predicted for each metagenome using FragGeneScan [20]. We then applied myRAST [34] (which provides extremely fast functional annotation based on sequence similarity) to assign putative functions to all predicted proteins from each of the metagenomes. A table was produced such that each row is a metagenome and each column is a function (a FigFam ID; we only kept functions each of which were annotated in at least 10 metagenomes), and the number in each cell shows the number of proteins in the corresponding metagenome that are assigned to the corresponding function. We then computed the similarity of the functional contents between any two metagenomes using Spearman's rank correlation, and the resulted similarity table was used to compute the tree

of metagenomes using a hierarchical clustering algorithm (average linkage). To test the size impacts of metagenome collections on the performance of community profiling, from the hierarchical clustering result, we prepared three collections of non-redundant metagenomes: *mg-300* (with 300 metagenomes), *mg-500* and *mg-1000*. We denoted the entire collection of metagenomes as *mg-all*.

In order to compare the performance of community profiling with the performance of phylogenetic profiling, we used the list of all available complete bacterial and archaeal genomes (denoted as *ref-all*), and a non-redundant subset containing 1000 reference genomes using the same approach based on the functional contents of the genomes (we denoted this non-redundant set as *ref-1000*).

2.6. Implementation

The distance measures, transformation functions, and the optimization methods are implemented mainly in Java. All programs are available for download at <https://sourceforge.net/projects/communityprofiling/>. Other files including the lists of reference metagenomes/genomes and their information are available on our website at <http://omics.informatics.indiana.edu/CP/>.

3. Results

We evaluated our new community profiling approach and compared its performance with phylogenetic profiling approach in two well-studied bacterial species, *E. coli* and *B. subtilis*. We further showed that integrating the two profiling methods can reach slightly more accurate predictions than individual approaches. Since our community profiling approach only uses the community information, it opens up opportunity for predicting functional links for sequenced genomes that are poorly represented by the reference genomes (so the profiling approach will be of limited usage in such cases), as demonstrated in application of our method to *P. copri*.

3.1. Predicting functional associations in *E. coli* and *B. subtilis* using community profiles

We first evaluated our approach for functional association prediction in *E. coli* and *B. subtilis* because these two species have been extensively studied, providing more complete data for validation. Note that even for these two organisms, we only have partial information about the protein interactions, and new interactions are yet to be discovered. Therefore, our “gold standard” can be inaccurate because some non-interacting (negative) pairs in the dataset can be mis-classified.

We applied the 10-fold cross validation and the results are shown in Figure 2a and 3a. For the prediction in both genomes, mutual information with negative logarithm transformation of *E*-values and the Spearman’s rank correlation method resulted in the highest AUCs. In general, the prediction performance is better for *E. coli* than for *B. subtilis*, perhaps due to the fact that the known functional association data are more complete for *E. coli* than for *B. subtilis*.

We evaluated if using a non-redundant collection of the reference metagenomes can improve the prediction accuracy. Performance on both the *E. coli* and *B. subtilis* genomes (Figure 2a and 3a) showed that for some distance metrics (e.g., Pellegrini/Hamming distance), using non-redundant collections of metagenomes improved the prediction of functional association. However, for the methods that achieved the highest accuracy, including the mutual information with negative logarithm transformation of *E*-values and the Spearman's rank correlation method, using all metagenomes helped improve the performance. By contrast, using the non-redundant reference genomes (ref-1000) for phylogenetic profiling consistently resulted in better performance than using the entire set of reference genomes (ref-all) across the different metrics (see Figure 2b and 3b). This contrast indicates that mutual information and Spearman's rank correlation are less sensitive to the distribution of reference metagenomes (and genomes) and there is a greater genomic diversity in the collection of metagenomes than the collection of genomes. Based on these results, we recommend that the whole set of reference metagenomes to be used for prediction of functional association in new species. However, considering that computing the community profiles using the whole set of reference metagenomes will be slower, it is reasonable to use a non-redundant set of metagenomes (such as mg-1000) for community profiling, which as we show in Figure 2a and 3a, resulted in only slightly worse performance than using the whole set of metagenomes.

3.2. Comparing the performance of community profiling and phylogenetic profiling methods

Table 1 summarizes the performances of the phylogenetic profiling and community profiling approaches. Overall, their performances using the two best distance metrics—mutual information and Spearman's rank correlation—are comparable, with the community profiling method performing slightly better.

Notably, not only prediction accuracies are comparable between the community profiling and phylogenetic profiling methods, the distance measures from these two profiling method are highly correlated with each other when the same distance measure is applied (Figure 4). Still there are a small proportion of protein pairs having different scores between the two profiling methods, indicating that more accurate predictions can be achieved by combining these two profiling approaches. A simple integration approach is to choose the smaller distance score reported by these two methods for each pair of proteins for the final prediction. When applying this integrated score in the mutual information method or the Spearman's rank correlation-based method, we obtained a better prediction performance than that of individual methods (see Table 1 and Figure 5).

3.3. Predicting functional links in a newly sequenced human-associated bacterial genome

P. copri was one of the reference genomes sequenced by the HMP, and it is one of the common bacterial species associated with human guts [17]. Protein sequences encoded by this genome, and their putative functions were predicted using myRAST. Out of the 3,080 protein sequences, there are 1,789 proteins that have functional annotations (assigned to FigFams; see details in Supplementary File 1), among which 647 can be assigned to metabolic pathways (see Supplementary File 2 for details). We considered that proteins

assigned to the same metabolic pathway are functionally linked. In total, there are 2252 such pairs of proteins. We also prepared a list of 2252 pairs of proteins that are unlikely to be associated, by randomly selecting a pair from the list of proteins in this genome (but excluding the pairs that were assigned to the same pathways). We note that other collections of pathways such as KEGG pathways (<http://www.kegg.jp/kegg/>) and MetaCyc (<http://metacyc.org>) may also be used for the validation purposes.

We applied the mutual information based community profiling and phylogenetic profiling to predict the functional associations between the proteins in this genome. We tried different cutoffs to see how that impacted on the performance in terms of recall and false positive rate (FPR). The results are summarized in Table 2. With small cutoffs (e.g., 0.1 or 0.2), the community profiling method identified a reasonable fraction of functionally coupled proteins with high accuracy (low FPR). As the distance cutoff increases, the community profiling method predicts more associated pairs, but the accuracy drops quickly (higher FPRs). By contrast, phylogenetic profiling only predicted very few proteins with functional associations, even when relatively high distance cutoffs (e.g., 0.5) were applied. Overall, at the same recall rate, the community profiling approach achieved lower FPR than the phylogenetic profiling approach.

The different performances of the two approaches can be explained by the different tendency of finding homologs for query proteins in metagenomes and genomes. Figure 6 shows that there are more proteins in *P. copri* with homologs found in the reference metagenomes (in community profiling) than proteins with homologs found in genomes (in phylogenetic profiling). This indicates that the community profiling has its advantage of predicting functional associations when applied to less-well studied bacterial genomes (or metagenomes).

We built a protein functional network from all predicted pairs of associated proteins (using 0.1 as the distance cutoff so we can show the network) in *P. copri* (see Figure 7a). In this network, there are a total of 670 pairs (see Supplementary File 3 for pairs of genes and their correlation scores). We show a few modules in Figure 7 (b)–(d). Figure 7b shows a module consisting of a few proteins that are likely to form a large protein complex, including prot 02552, which was predicted to be Na(+)-translocating NADH-quinone reductase subunit B (EC 1.6.5.-) by myRAST. Figure 7c shows an example where proteins all have predicted functions, but some were not assigned to the same pathway (prot 00204 and prot 00365 were predicted to participate in Histidine Biosynthesis pathway, but the other two proteins prot 02834 and prot 00202 were not). We note that we identified a tightly connected module with six proteins (Figure 7d), all of which have no functional assignments by myRAST based on sequence similarity. Although we cannot assign a specific function to this group of proteins (since all proteins had no functional annotation), a future direction is to infer hints for functional annotation for this group of proteins by studying their associated microbial communities.

4. Discussion

We have developed a community profiling approach for functional association prediction. We showed that its performance is dependent on the metrics to be used. It was shown that a reasonable and tree-guided selection of reference genome can improve the prediction accuracy by reducing the phylogenetic bias in phylogenetic profiling [35]. Our results however showed the community profiling achieved its best performance when it uses all metagenomes for profiling. We expect that this will change as metagenomes keep accumulating. We will also try different approaches of selecting reference metagenomes, aiming to further improve the performance of the community profiling approach.

In community profiling method, the cutoff value for distance score may affect the accuracy of the functional association prediction significantly. The prediction of associated pairs of proteins is a highly unbalanced learning problem, where there are many more negatives (pairs without any association) than the positives (pairs with association). Therefore, a small change in the cutoff value can greatly affect the accuracy of the prediction. Still we show that when the cutoff value is set to a small value, the prediction can be very accurate. We can further build a network of genes based on the distance data, and such a network may be useful for improving the prediction of functional associations between genes. We note that a refined annotation of genes (e.g., with binding specificity) however can only be achieved by integrating multiple lines of evidence.

In this paper, we only reported results of applying community profiling approach to predicting functional associate between genes/proteins in individual genome. However, community profiling method can also be applied to predict functional links between genes in a metagenome, or functional links between gene clusters predicted from metagenomes such as the gut microbial gene catalogue [46] (in this case, representative sequences of the gene clusters can be used for similarity search and profiling). We released our tool as open source codes on sourceforge; users can get our tool and run it as a standalone program for predicting functional associations, or integrate it with other approaches for functional prediction. One of the future development directions for us is to pre-compute functional associations for important bacterial genomes and metagenomes and make the predictions available through a web server.

Critical Assessment of protein Function Annotation (CAFA) experiment [2] is a worldwide effort, aiming to analyze and evaluate protein function prediction methods. According to this assessment, methods based on massive integration of evolutionary analyses and multiple data sources outperform other approaches. The main focus of this paper, however, is to demonstrate the usefulness of community profiling for functional annotation, so a comparison of our method with all the existing methods, esp. those integrating multiple sources of informations (or predictions) will not be necessary. We believe that the community information will be of increasing importance considering the on-going research activities in metagenomics, and it can be integrated with other approaches (similarity-based and context-based approaches) for functional predictions.

5. Conclusions

In this paper we proposed a new method to use community sequencing data in a profiling method to predict associations between proteins. We have shown that the community profiling method slightly outperformed the phylogenetic profiling method in predicting functional associations in well-studied genomes. A method combining the two profiling methods can further improve the prediction accuracy. At the same time, the community profiling is more suitable for predicting functional associations in genomes (or metagenomes) that have not been well studied. With the rapid increase of (meta)genomic data, an immediate application of our method is to predict associations for the new sequences that are hardly matched to any sequences in current genome databases.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by the NIH grant 1R01AI108888 and NSF grant DBI-0845685 to Ye. The authors thank Dr. Haixu Tang for helpful discussions.

References

1. Markowitz VM, Chen IMA, Chu K, Szeto E, Palaniappan K, Grechkin Y, Ratner A, Jacob B, Pati A, Huntemann M, Liolios K, Pagani I, Anderson I, Mavromatis K, Ivanova NN, Kyrpides NC. IMG/M: the integrated metagenome data management and comparative analysis system. *Nucleic Acids Res.* 2012; 40(Database issue):D123–D129. DOI: 10.1093/nar/gkr975 [PubMed: 22086953]
2. Radivojac P, Clark WT, Oron TR, Schnoes AM, Wittkop T, Sokolov A, Graim K, Funk C, Verspoor K, Ben-Hur A, et al. A large-scale evaluation of computational protein function prediction. *Nat Methods.* 2013; 10(3):221–227. [PubMed: 23353650]
3. Loewenstein Y, Raimondo D, Redfern OC, Watson J, Frishman D, Linial M, Orengo C, Thornton J, Tramontano A, et al. Protein function annotation by homology-based inference. *Genome Biol.* 2009; 10(2):207. [PubMed: 19226439]
4. Lee D, Redfern O, Orengo C. Predicting protein function from sequence and structure. *Nature Reviews Molecular Cell Biology.* 2007; 8(12):995–1005. [PubMed: 18037900]
5. Gherardini PF, Helmer-Citterich M. Structure-based function prediction: approaches and applications. *Briefings in functional genomics & proteomics.* 2008; 7(4):291–302. [PubMed: 18599513]
6. Korbel JO, Jensen LJ, Von Mering C, Bork P. Analysis of genomic context: prediction of functional associations from conserved bidirectionally transcribed gene pairs. *Nature biotechnology.* 2004; 22(7):911–917.
7. Sharan R, Ulitsky I, Shamir R. Network-based prediction of protein function. *Mol Syst Biol.* 2007; 3:88. [PubMed: 17353930]
8. Overbeek R, Fonstein M, Dsouza M, Pusch GD, Maltsev N. The use of gene clusters to infer functional coupling. *Proceedings of the National Academy of Sciences.* 1999; 96(6):2896–2901.
9. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci USA.* 1999; 96(8):4285–4288. [PubMed: 10200254]
10. Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D. Detecting protein function and protein-protein interactions from genome sequences. *Science.* 1999; 285(5428):751–753. [PubMed: 10427000]

11. Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, Bennett HA, Coffey E, Dai H, He YD, et al. Functional discovery via a compendium of expression profiles. *Cell*. 2000; 102(1):109–126. [PubMed: 10929718]
12. Brun C, Chevenet F, Martin D, Wojcik J, Guénoche A, Jacq B. Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network. *Genome biology*. 2003; 5(1):R6. [PubMed: 14709178]
13. Carpenter AE, Sabatini DM. Systematic genome-wide screens of gene function. *Nature Reviews Genetics*. 2004; 5(1):11–22.
14. Stulberg E, Fravel D, Proctor LM, Murray DM, LoTempio J, Chrisey L, Garland J, Goodwin K, Graber J, Harris MC, Jackson S, Mishkind M, Porterfield DM, Records A. An assessment of US microbiome research. *Nat Microbiol*. 2016; 1:15015. [PubMed: 27571759]
15. Fierer N, Leff JW, Adams BJ, Nielsen UN, Bates ST, Lauber CL, Owens S, Gilbert JA, Wall DH, Caporaso JG. Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. *Proc Natl Acad Sci USA*. 2012; 109(52):21390–21395. [PubMed: 23236140]
16. Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S, Wu D, Eisen JA, Hoffman JM, Remington K, et al. The sorcerer ii global ocean sampling expedition: northwest atlantic through eastern tropical pacific. *PLoS biology*. 2007; 5(3):e77. [PubMed: 17355176]
17. Huttenhower C, Gevers D, Knight R, Abubucker S, Badger JH, Chinwalla AT, Creasy HH, Earl AM, FitzGerald MG, Fulton RS, et al. Structure, function and diversity of the healthy human microbiome. *Nature*. 2012; 486(7402):207–214. [PubMed: 22699609]
18. Zmora N, Zeevi D, Korem T, Segal E, Elinav E. Taking it Personally: Personalized Utilization of the Human Microbiome in Health and Disease. *Cell Host Microbe*. 2016; 19(1):12–20. [PubMed: 26764593]
19. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T, et al. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*. 2010; 464(7285):59–65. [PubMed: 20203603]
20. Rho M, Tang H, Ye Y. FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Res*. 2010; 38(20):e191. [PubMed: 20805240]
21. Tang H, Li S, Ye Y. A Graph-Centric Approach for Metagenome-Guided Peptide and Protein Identification in Metaproteomics. *PLoS Comput Biol*. 2016; 12(12):e1005224. [PubMed: 27918579]
22. Lü F, Bize A, Guillot A, Monnet V, Madigou C, Chapleur O, Mazéas L, He P, Bouchez T. Metaproteomics of cellulose methanisation under thermophilic conditions reveals a surprisingly high proteolytic activity. *The ISME journal*. 2014; 8(1):88–102. [PubMed: 23949661]
23. Erickson AR, Cantarel BL, Lamendella R, Darzi Y, Mongodin EF, Pan C, Shah M, Halfvarson J, Tysk C, Henrissat B, Raes J, Verberkmoes NC, Fraser CM, Hettich RL, Jansson JK. Integrated metagenomics/metaproteomics reveals human host-microbiota signatures of crohn's disease. *PloS one*. 2012; 7(11):e49138. [PubMed: 23209564]
24. Faust K, Sathirapongsasuti JF, Izard J, Segata N, Gevers D, Raes J, Huttenhower C. Microbial co-occurrence relationships in the human microbiome. *PLoS computational biology*. 2012; 8(7):e1002606. [PubMed: 22807668]
25. Lorenz MG, Wackernagel W. Bacterial gene transfer by natural genetic transformation in the environment. *Microbiological reviews*. 1994; 58(3):563. [PubMed: 7968924]
26. Nielsen HB, Almeida M, Juncker AS, Rasmussen S, Li J, Sunagawa S, Plichta DR, Gautier L, Pedersen AG, Le Chatelier E, et al. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat Biotechnol*. 2014; 32(8):822–828. [PubMed: 24997787]
27. Albertsen M, Hugenholtz P, Skarshewski A, Nielsen KL, Tyson GW, Nielsen PH. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat Biotechnol*. 2013; 31(6):533–538. [PubMed: 23707974]
28. Alneberg J, Bjarnason BS, de Bruijn I, Schirmer M, Quick J, Ijaz UZ, Lahti L, Loman NJ, Andersson AF, Quince C. Binning metagenomic contigs by coverage and composition. *Nat Methods*. 2014; 11(11):1144–1146. [PubMed: 25218180]

29. Zhao Y, Tang H, Ye Y. RAPSearch2: a fast and memory-efficient protein similarity search tool for next-generation sequencing data. *Bioinformatics*. 2012; 28(1):125–126. [PubMed: 22039206]
30. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nat Methods*. 2015; 12(1):59–60. [PubMed: 25402007]
31. Yu YW, Daniels NM, Danko DC, Berger B. Entropy-scaling search of massive biological data. *Cell Syst*. 2015; 1(2):130–140. [PubMed: 26436140]
32. Peterson J, Garges S, Giovanni M, McInnes P, Wang L, Schloss JA, Bonazzi V, McEwen JE, Wetterstrand KA, Deal C, et al. The NIH Human Microbiome Project. *Genome Res*. 2009; 19(12):2317–2323. [PubMed: 19819907]
33. Grigoriev IV, Nordberg H, Shabalov I, Aerts A, Cantor M, Goodstein D, Kuo A, Minovitsky S, Nikitin R, Ohm RA, Otillar R, Poliakov A, Ratnere I, Riley R, Smirnova T, Rokhsar D, Dubchak I. The genome portal of the department of energy joint genome institute. *Nucleic Acids Res*. 2012; 40(Database issue):D26–D32. DOI: 10.1093/nar/gkr947 [PubMed: 22110030]
34. Overbeek R, Olson R, Pusch GD, Olsen GJ, Davis JJ, Disz T, Edwards RA, Gerdes S, Parrello B, Shukla M, Vonstein V, Wattam AR, Xia F, Stevens R. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Res*. 2014; 42(Database issue):D206–214. [PubMed: 24293654]
35. Simonsen M, Maetschke S, Ragan M. Automatic selection of reference taxa for protein–protein interaction prediction with phylogenetic profiling. *Bioinformatics*. 2012; 28(6):851–857. [PubMed: 22219205]
36. Von Mering C, Huynen M, Jaeggi D, Schmidt S, Bork P, Snel B. String: a database of predicted functional associations between proteins. *Nucleic acids research*. 2003; 31(1):258–261. [PubMed: 12519996]
37. Leinonen R, Diez FG, Binns D, Fleischmann W, Lopez R, Apweiler R. Uniprot archive. *Bioinformatics*. 2004; 20(17):3236–3237. DOI: 10.1093/bioinformatics/bth191 [PubMed: 15044231]
38. Wu J, Kasif S, DeLisi C. Identification of functional links between genes using phylogenetic profiles. *Bioinformatics*. 2003; 19(12):1524–1530. [PubMed: 12912833]
39. Glazko GV, Mushegian AR. Detection of evolutionarily stable fragments of cellular pathways by hierarchical clustering of phyletic patterns. *Genome Biol*. 2004; 5(5):R32.doi: 10.1186/gb-2004-5-5-r32 [PubMed: 15128446]
40. Yamada T, Goto S, Kanehisa M. Extraction of phylogenetic network modules from prokaryote metabolic pathways. *Genome Inform*. 2004; 15(1):249–258. [PubMed: 15712127]
41. Freilich S, Goldovsky L, Gottlieb A, Blanc E, Tsoka S, Ouzounis CA. Stratification of co-evolving genomic groups using ranked phylogenetic profiles. *BMC Bioinformatics*. 2009; 10:355.doi: 10.1186/1471-2105-10-355 [PubMed: 19860884]
42. Date SV, Marcotte EM. Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages. *Nat Biotechnol*. 2003; 21(9):1055–1062. DOI: 10.1038/nbt861 [PubMed: 12923548]
43. Joe H. Relative entropy measures of multivariate dependence. *Journal of the American Statistical Association*. 1989; 84(405):157–164.
44. Sun J, Xu J, Liu Z, Liu Q, Zhao A, Shi T, Li Y. Refined phylogenetic profiles method for predicting protein-protein interactions. *Bioinformatics*. 2005; 21(16):3409–3415. DOI: 10.1093/bioinformatics/bti532 [PubMed: 15947018]
45. Sun J, Li Y, Zhao Z. Phylogenetic profiles for the prediction of protein-protein interactions: how to select reference organisms? *Biochem Biophys Res Commun*. 2007; 353(4):985–991. DOI: 10.1016/j.bbrc.2006.12.146 [PubMed: 17207465]
46. Li J, Jia H, Cai X, Zhong H, Feng Q, Sunagawa S, Arumugam M, Kultima JR, Prifti E, Nielsen T, et al. An integrated catalog of reference genes in the human gut microbiome. *Nat Biotechnol*. 2014; 32(8):834–841. [PubMed: 24997786]

Highlights

- A new guilt-by-association approach for functional association prediction is proposed.
- The approach utilizes gene co-occurrence across microbial communities.
- The selection of reference metagenomes and correlation metric matters.

Community profiling

	soil	ocean	river	gut	skin
g1	1	1	1	0	0
g2	1	1	1	0	0
g3	0	1	0	1	1
g4	0	1	1	1	1

Phylogenetic profiling

	E.coli	A.urinae	Y. pestis	B. fragilis	C. efficiens
g1	1	1	1	0	0
g2	1	1	1	0	0
g3	0	1	0	1	1
g4	0	1	1	1	1

Figure 1. A schematic illustration of the community profiling approach and its comparison with the phylogenetic profiling

Each row in the tables represents a community profile of a gene across microbial communities associated with different environments (e.g., soil microbiome) and hosts (e.g., skin microbiome) or a phylogenetic profile across genomes. In this illustration, the values in the profiles are binary with 1 indicating presence and 0 indicating absence of the gene or its homolog in the corresponding metagenome/genome. However, in actual calculation, E-values or scores transformed from E-values are used. In this figure, g1 and g2 share similar community/phylogenetic profiles and are likely to be functionally linked; likewise g3 and g4 share similar profiles.

(a) Community profiling

(b) Phylogenetic profiling

Figure 2. Comparison of the performance in *E. coli*

The distance metrics evaluated include (from left to right in each figure) Pellegrini (Pellegrini/Hamming distance), HG (the hypergeometric distribution method), Jaccard (the Jaccard index), MI (the mutual information based on negative logarithm of *E*-values), Date (the mutual information based on negative inverse of logarithm of *E*-values), Pearson (the Pearson's correlation), and Spearman (the Spearman's rank correlation). For each distance metric, we compared the performance using all metagenomes (mg-all) and three subsets of reference metagenomes (mg-300, mg-500 and mg-1000).

(a) Community profiling

(b) Phylogenetic profiling

Figure 3. Comparison of the performance in *B. subtilis*
Refer to Figure 2 legend for explanations of the metrics and reference metagenomes (metagenomes).

Figure 4. Correlation of the distance scores between protein pairs computed by the community profiling and phylogenetic profiling methods

(a) The pairwise distance scores based on mutual information with negative logarithm of E -values; (b) The pairwise distance scores based on Spearman's rank correlation. In both figures, the red dots are the non-interacting pairs of proteins and the blue dots are the interacting pairs. The top and the right panels show the distributions of the distances for interacting and non-interacting pairs.

Figure 5. Integration of the two profiling methods can improve the prediction of functional associations

(a) Receiver operating characteristic (ROC) curve based on mutual information with negative logarithm of E -values. (b) ROC curve based on Spearman's correlation. In both cases, the AUC increases when the minimum of the distances computed using the two profiling approaches is used for prediction.

Figure 6. More *P. copri* proteins have homologs in reference metagenomes as compared to reference genomes

A reference (meta)genome is considered to have a homolog for a query protein, if the *E*-value of the top hit in the (meta)genome is $\leq 10^{-5}$.

Figure 7. A network of proteins with functional associations in *P. copri*

Each node represents a protein and an edge between two nodes represents a predicted functional association between the proteins. Proteins that are assigned to the same metabolic pathway are labeled with the same colors and other proteins (including those with functional annotations but not assigned to a metabolic pathway, and those without functional annotations) are shown in white. (a) shows the global network with all predicted associations (distance cutoff=0.1). A few modules in (a) (highlighted in red circles) are shown in (b)–(d), showing different situations: (b) all proteins have functional annotations and are assigned to the same metabolic pathway by myRAST; (c) some proteins have functional annotations, but are not assigned to any pathway; and (d) all proteins have no functional annotations.

Table 1

A summary of the performance of the community profiling (CP) and phylogenetic profiling (PP) methods, and their integration as measured by AUC.

	CP	PP	Minimum
Spearman's Rank Correlation	0.840	0.838	0.851
Mutual Information	0.830	0.821	0.840

The Minimum column lists the performance when the smaller value of the CP and PP distances was used for prediction.

Table 2

The prediction accuracy of the profiling methods for *Prevotella copri* at different distance cutoffs.

	Community Profiling		Phylogenetic Profiling	
Cutoff	Recall	FPR	Recall	FPR
0.1	0.027	0.000	0.000	0.000
0.2	0.342	0.038	0.016	0.000
0.3	0.575	0.119	0.063	0.001
0.4	0.773	0.240	0.143	0.009
0.5	0.935	0.369	0.254	0.027
0.6	0.980	0.488	0.427	0.067
0.7	0.998	0.591	0.570	0.130
0.8	1.000	0.685	0.786	0.253
0.900	-	-	0.963	0.491
0.925	-	-	0.980	0.587
0.950	-	-	0.990	0.702
0.975	-	-	0.998	0.803
1.0	-	-	1.0	1.0