Methods 129 (2017) 8–17

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

Functional association prediction by community profiling

Dazhi Jiao [1], Wontack Han [1], Yuzhen Ye *

Indiana University, 150 S. Woodlawn Ave, Bloomington, IN 47405, United States

ARTICLE INFO

Article history:
Received 18 January 2017
Received in revised form 31 March 2017
Accepted 20 April 2017
Available online 26 April 2017

Keywords:
Guilt-by-association
Functional association prediction
Phylogenetic profiling
Community profiling
Metagenomics

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.

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

* Corresponding author.
E-mail address: yye@indiana.edu (Y. Ye).
[1] These authors contributed equally to this work.

http://dx.doi.org/10.1016/j.ymeth.2017.04.018
1046-2023/

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 [21–23].

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-occurence 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 Fig. 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 assiciation 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