Methods 102 (2016) 12–19

Contents lists available at ScienceDirect

Methods

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

Predicting microbial interactions through computational approaches

Chenhao Li [a,b], Kun Ming Kenneth Lim [a,c], Kern Rei Chng [a], Niranjan Nagarajan [a,b,*]

[a] Genome Institute of Singapore, Singapore 138672, Singapore
[b] Department of Computer Science, National University of Singapore, Singapore 117417, Singapore
[c] Computational Biology program, National University of Singapore, Singapore, 117543, Singapore

ARTICLE INFO

Article history:
Received 15 September 2015
Received in revised form 15 January 2016
Accepted 23 February 2016
Available online 26 March 2016

Keywords:
Microbial interactions
Metagenomics
Reverse ecology
Text mining

ABSTRACT

Microorganisms play a vital role in various ecosystems and characterizing interactions between them is an essential step towards understanding the organization and function of microbial communities. Computational prediction has recently become a widely used approach to investigate microbial interactions. We provide a thorough review of emerging computational methods organized by the type of data they employ. We highlight three major challenges in inferring interactions using metagenomic survey data and discuss the underlying assumptions and mathematics of interaction inference algorithms. In addition, we review interaction prediction methods relying on metabolic pathways, which are increasingly used to reveal mechanisms of interactions. Furthermore, we also emphasize the importance of mining the scientific literature for microbial interactions – a largely overlooked data source for experimentally validated interactions.

© 2016 Elsevier Inc. All rights reserved.

1. Introduction

As an essential component in various ecosystems, microorganisms aggregate to form heterogeneous communities comprising of distinct proportions of diverse microbial entities, often referred to collectively as the *Microbiome*. Microorganisms in a microbiome do not live in isolation, but instead actively interact with other members within their community. Taken as a whole, these interactions are a description of the overall function of the microbial community. As such, the characterization of microbial interactions is a key step towards the understanding of the community organization [1–3] and the engineering of microbial communities for biomedical [4–7] and industrial applications [8–10].

The pair-wise interaction between two microbes is the fundamental unit of microbial interactions. Such interactions can be categorized by their effect on the participants, i.e. positive, negative or neutral. In combination, there exist six core categories of ecological interactions: mutualism (positive-positive), competition (negative-negative), antagonism (positive-negative), commensalism (positive-neutral), amensalism (negative-neutral) and neutralism (neutral-neutral).

Traditionally, the investigation of microbial interactions required the use of laboratory experiments such as growth and co-culture assays [11–13]. However, the laborious nature of such methods renders them infeasible for large scale application. Computational approaches offer the opportunity to alleviate this issue by predicting interaction candidates for experimental validation [4,14,15]. These predictions can be based on various types of data such as the measured species abundances from high-throughput sequencing or reconstructed metabolic models for species communities. In addition, computational methods may also assist the collation of experimentally verified interactions from large compendiums of published literature. A graphical overview of computational approaches for predicting microbial interactions can be found in Fig. 1.

Following the previous review on *in silico* microbial interaction inference from microbial abundance data [16], a number of new methods have since been proposed to address the various challenges in such a task, requiring an updated review accounting for such methods. In addition, previous reviews did not consider methods using genomic information on metabolism, a paradigm that is increasingly used to characterize metabolic interactions among microbial community members. Here, we review the available computational approaches (grouped by the different types of data that they use) and the challenges that they address, discuss their advantages and limitations, and point out directions for future work in this area.

2. Inferring interactions from metagenomic survey data

Advances in high throughput sequencing technologies have made it possible to quantify the abundances of members in a

* Corresponding author.
E-mail address: nagarajann@gis.a-star.edu.sg (N. Nagarajan).

http://dx.doi.org/10.1016/j.ymeth.2016.02.019
1046-2023/© 2016 Elsevier Inc. All rights reserved.

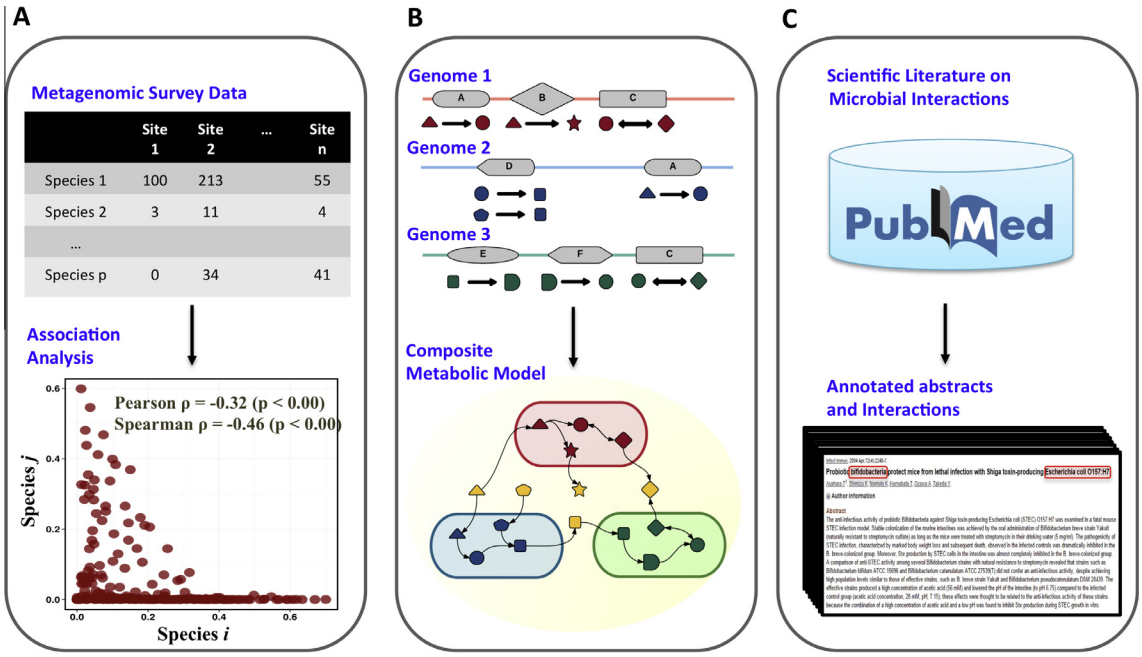

Fig. 1. A graphical overview of computational approaches to predict microbial interactions. (A) Microbial interactions are frequently inferred by observing correlations in species abundances in metagenomic survey datasets as depicted here for a pair of species. (B) Interactions can also be predicted by reconstructing pathways from annotated genomes for each species and then jointly modeling community metabolism to identify metabolites that serve as interaction interfaces (show in yellow). Genes are depicted here as gray shapes while associated metabolites are shown as colored shapes. (C) Text mining of scientific literature databases (e.g. NCBI PubMed) is another approach for cataloguing microbial interactions that are experimentally validated and can serve as a gold-standard for the field.

microbial community in a relatively unbiased manner by sequencing marker genes or whole metagenomes (i.e. total DNA from a microbial community). The abundance of each species (or higher taxa) is then estimated from the counts of reads assigned to the respective taxa (see Peabody et al. [17] for a comprehensive comparison of available methods for taxonomic classification of metagenomic sequencing reads). The data collected can then be further tabulated into a data matrix where each row represents read counts of a species across all the samples. To account for differences in sequencing depth (i.e. total number of reads generated for a sample), read counts are often normalized into proportions (relative abundances) by dividing by the column sums. Alternatively, the data matrix can be simplified to record only presence (1) or absence (0) information by setting a minimum threshold on read counts or relative abundances. Metagenomic survey datasets can be collected across different sites or across different time points within the same site, with the techniques used to infer microbial interactions from them being somewhat distinct. In the next section we review methods that use survey data without a temporal component, referred to here as "spatial" metagenomic survey data.

2.1. Microbial interaction inference with spatial metagenomic survey data

Spatial metagenomic survey data provide a static view of the composition of microbiota across different sites. For human-associated microbiota, several recent studies have generated a significant amount of data across different patients and different body sites [18,19]. The size of such datasets varies across different studies. For example, data is currently available for over 1200 samples across different body sites (22–207 samples per site) from the Human Microbiome Project (HMP), providing relative abundances for more than 40,000 taxonomic groups in total. While many studies have focused on investigating the composition of microbial communities or identifying species associated with certain phenotypes, these datasets can also be used to infer interactions between species. Although this can be a coarse-grained approach, inferring microbial interactions from available spatial metagenomic survey data can serve as the basis for understanding community structure and to generate useful hypotheses for further investigation [16]. The underlying rationale of the inference is that the observed community structure is driven by the ecological interactions between species, and therefore the non-random pattern of species distribution can be used to infer these interactions (Fig. 1A). Such patterns include simple associations such as co-occurrence or co-exclusion and correlation, as well as more complex associations such as limited cycles in predator-prey systems [20].

2.1.1. Using co-occurrence or co-exclusion patterns

The simplest and yet interesting pattern that serves to inform about species interactions is the co-occurrence or co-exclusion of two species, providing evidence that there is strong dependency (i.e. positive interactions including mutualism and commensalism) or competition (i.e. negative interactions including competition, antagonism and amensalism) between them. The detection of such patterns can be formulated into a statistical test of whether the species pair co-occurs or co-excludes each other more than random using the *Fisher's exact test*, which compares the co-occurrence pattern with the hypergeometric distribution to assign statistical significance [21].

In addition to the Fisher's exact test, it is possible to quantify the similarity between the distributions of two species across sites, with ecological distance (or similarity) scores. These scores, originally designed for comparing the overall species composition of two sites (e.g. the *Jaccard distance* [22] or *Bray-Curtis dissimilarity* [1]), have since been applied to compare the composition of two species across sites (using species abundances normalized across sites). Such distance scores obtain maximum (e.g. 1) when species are mutually exclusive and minimum (e.g. 0) when species have

Download English Version:

https://daneshyari.com/en/article/1993177

Download Persian Version:

https://daneshyari.com/article/1993177

Daneshyari.com