Functional gene networks based on the gene neighborhood in metagenomes

Chan Yeong Kim and Insuk Lee

Department of Biotechnology, College of Life Science and Biotechnology, Yonsei University, Seoul, Korea

ABSTRACT

The gene neighborhood in prokaryotic genomes has been effectively utilized in inferring co-functional networks in various organisms. Previously, such genomic context information has been sought among completely assembled prokaryotic genomes. Here, we present a method to infer functional gene networks according to the gene neighborhood in metagenome contigs, which are incompletely assembled genomic fragments. Given that the amount of metagenome sequence data has now surpassed that of completely assembled prokaryotic genomes in the public domain, we expect benefits of inferring networks by the metagenome-based gene neighborhood. We generated co-functional networks for diverse taxonomical species using metagenomics contigs derived from the human microbiome and the ocean microbiome. We found that the networks based on the metagenome gene neighborhood outperformed those based on 1748 completely assembled prokaryotic genomes. We also demonstrated that the metagenome-based gene neighborhood could predict genes related to virulence-associated phenotypes in a bacterial pathogen, indicating that metagenome-based functional links could be sufficiently predictive for some phenotypes of medical importance. Owing to the exponential growth of metagenome sequence data in public repositories, metagenome-based inference of co-functional networks will facilitate understanding of gene functions and pathways in diverse species.

Introduction

The functional and regulatory constraints acting on genes have contributed to shaping pathways and functional modules. As a result, consideration of the genomic context often provides useful features for inferring functional links between genes (Galperin and Koonin 2000). The prokaryotic operon, in which all encoded genes are co-transcribed and participate in the same biological process, is a well-known example of such a genomic context. Large-scale identification of genes contributing to the operon structure generally relies on measurement of the chromosomal distance between genes (Dandekar et al. 1998, Overbeek et al. 1999) or determining the probability that a given pair of conserved genes is in close proximity to each other across prokaryotic genomes (Bowers et al. 2004). In fact, these two approaches for the inference of functional links based on gene neighborhoods are complementary (Shin et al. 2014). Identifying the gene neighborhood can help to infer co-functional links in not only prokaryotic species but also in higher eukaryotes using orthologous gene neighborhoods in prokaryotic genomes.

Previous studies adopting a gene neighborhood approach have employed only completely assembled prokaryotic genomes. However, a distance-based gene neighborhood method can also be theoretically applied to incompletely assembled contigs. Recently, next-generation sequencing technology has dumped massive amounts of metagenome sequence data from host organisms and environments. For example, the Human Microbiome Project (HMP) provided approximately 700 shotgun sequencing samples across several body sites (Human Microbiome Project 2012), and the TARA Oceans project provided 7.2 terabases of metagenomic data from 243 samples collected from 68 locations worldwide (Sunagawa et al. 2015). Since the metagenome shotgun sequencing projects have released abundant metagenomic contigs, we now have the opportunity to identify functional associations between genes using the distance-based gene neighborhood. The metagenome-based gene neighborhood has been applied to the functional prediction of prokaryotic genes (Harrington et al. 2007). However, no study has yet assessed the ability of the metagenome-based gene neighborhood approach in retrieving functional links between genes and biological processes in eukaryotic organisms. In this study, we developed a method for inferring functional links from metagenome contigs and
assessed the inferred networks for various organisms, including a bacterium, a fungus, an animal, and a plant species. We found that the functional networks generated by the metagenome-based gene neighborhood outperformed those generated by the gene neighborhood across 1748 completely assembled prokaryotic genomes (122 archaea and 1626 bacteria). We also demonstrated that the metagenome-based gene neighborhood could retrieve genes related to virulence-associated phenotypes in a bacterial pathogen.

**Materials and methods**

**Human and ocean microbiome contig data**

We downloaded the human microbiome contig data from the HMP Data Portal (Human Microbiome Project 2012). The microbiome was derived from 16 different human body sites of ∼300 healthy individuals, comprising 65 billion base pairs of DNA of 47 million contigs from 754 samples. We compiled the ocean microbiome contig data from the TARA Ocean Project (Sunagawa et al. 2015). The project database provides 242 ocean microbiome samples derived from six oceans and seas: the Pacific Ocean, Atlantic Ocean, Mediterranean Sea, Red Sea, Indian Ocean, and Southern Ocean. The sequence data contain approximately 67 billion base pairs from ∼62 million contigs.

**Sequence alignment using DIAMOND**

Because of the massiveness of the microbiome data, we used DIAMOND (Buchfink et al. 2015), a fast and reliable sequence alignment tool, to align genes of organisms for network construction against metagenome contig sequences. In this work, we used the --sensitive option to obtain more reliable alignments. In addition, we set the --max-target-seqs to maximize the number of aligned sequences and set the --evalue to 1e-10 to ignore the weakly aligned sequences.

**Network inferring algorithm**

For a given organism, we aligned all genes to the human and ocean microbiome contigs using DIAMOND. If two genes were aligned in the same contig and the distance between them was less than 1000 base pairs, we considered these genes to co-exist. For all co-existing gene pairs, the significance of the gene neighborhood, $S$, was calculated as follows:

$$S = \log\left(\frac{d + 1}{n}\right),$$

where $n$ is the total number of contigs in which two genes co-exist and $d$ is the median distance (base pairs) between them across all co-existing contigs. We added a pseudo-count (1) to the median distance to avoid having to calculate the logarithm of zero, and took logarithm to reduce dispersion of $S$ scores.

**Gene neighborhood networks based on prokaryotic genomes and gold-standard functional links for network assessment**

We downloaded the gene neighborhood network based on 1748 completely assembled prokaryotic genomes and gold-standard functional links for *Arabidopsis thaliana*, *Danio rerio*, and *Saccharomyces cerevisiae* from AraNet v2 (Lee et al. 2015), DanioNet (Shim et al. 2016), and YeastNet v3 (Kim et al. 2014), respectively. For *Staphylococcus aureus*, the network was constructed based on the method described in our previous work (Shin et al. 2014). We compiled the gold-standard functional links of *S. aureus* for benchmarking networks from the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et al. 2012) and MetaCyc (Caspi et al. 2016) pathway databases.

**Virulence-associated phenotype genes of *S. aureus***

We compiled 253 genes for 10 virulence-associated phenotypes from three distinct data sources. First, we obtained the *S. aureus* genes reported to be associated with hemolysin production, protease activity, pigment formation, and mannitol fermentation from a previous large-scale mutant screening study using *S. aureus* strain USA300 (Fey et al. 2013). Second, we compiled genes for adherence, capsule, exoenzyme, toxin, and secretion systems from the Virulence Factor Database (VFDB) (Chen et al. 2005). Finally, we compiled genes that correspond to the genome-wide association study loci attributed to increased toxicity (Laabei et al. 2014). We excluded intergenic loci and mapped each locus to the gene containing the locus.

**Network-based gene prioritization and receiver operation characteristic (ROC) analysis**

We prioritized the genes for each virulence-associated phenotype using the naïve Bayes algorithm for network information propagation, which propagates information of network nodes to the direct neighbors only. The score of propagated information depends on the sum of the edge weight to the virulence genes (Wang and Marcotte 2010). For the given prioritized
genes of each virulence-associated phenotype, we conducted the ROC analysis based on the true-positive rate and false-positive rate of prioritized candidate genes. We then measured the area under the ROC curve (AUC) and used it as a measure of prediction power.

Figure 1. Overview of the construction process of the metagenome-based gene neighborhood network. We aligned Arabidopsis (*A. thaliana*), yeast (*S. cerevisiae*), zebrafish (*D. rerio*), and *S. aureus* genes against the metagenome contig sequences derived from the human and ocean microbiomes. For each species, we defined a co-functional relationship between genes that co-exist in the same contig, which were given a score according to their median distance (in base pairs) and the total number of contigs in which the two genes co-occur.
Results and discussion

Construction of the metagenome-based gene neighborhood network

Previous studies have demonstrated that two genes located in chromosomal proximity in prokaryotic genomes tend to be functionally coupled (Shin et al. 2014). We hypothesized that a similar approach may work for metagenome contigs, which are incompletely assembled genomic DNA fragments. Because of the larger amount of novel sequences available in the metagenome, determining the gene neighborhood in metagenomes may significantly expand the coverage of inferred functional networks.

The overall procedure for the construction of the metagenome-based gene neighborhood network is summarized in Figure 1 as briefly described in the Materials and Method section. First, we obtained the human and ocean microbiome contig sequence from the HMP (Human Microbiome Project 2012) and TARA Ocean Project (Sunagawa et al. 2015) datasets, respectively. To investigate the applicability of the metagenome-based inference of functional links in broad taxonomical groups, we selected A. thaliana, D. rerio, S. cerevisiae, and S. aureus as representative species for plants, animals, fungi, and bacteria, respectively. For each species, we aligned gene sequences to the metagenome contig sequences by DIAMOND (Buchfink

Figure 2. Assessment of gene neighborhood networks based on completely assembled genomes or metagenome contigs in four species: A. thaliana (A), S. cerevisiae (B), S. aureus (C), and D. rerio (D). The precision value represents the proportion of gene pairs that are annotated by the same biological process or pathway terms for a given species.
et al. 2015). Next, we found pairs of genes that co-existed in the same contigs and assigned scores according to their median distance and frequency of co-existence. We defined the final metagenome-based gene neighborhood network for each species by selecting links that passed the threshold of the likelihood score that was calculated based on a Bayes statistics framework as described in our previous work (Lee et al. 2004).

Assessment of metagenome-based gene neighborhood networks

We assessed the accuracy and genome coverage of the metagenome-based gene neighborhood networks. Based on the gold-standard gene pairs of each of four organisms, we calculated the precision and corresponding coding genome coverage for every 1000 links. In *A. thaliana* and *S. cerevisiae*, the metagenome-based networks outperformed the network constructed based on completely assembled prokaryotic genomes in terms of both precision and coverage (Figure 2(A) and (B)). In *S. aureus*, the metagenome-based network showed much better performance than that based on completely assembled prokaryotic genomes (Figure 2(C)). We observed similar overall performance between the two types of gene neighborhood networks in *D. rerio*, although higher precision was obtained for top-ranked functional links in the metagenome-based network (Figure 2(D)). Taken together, we concluded that we can infer functional networks with equal or higher performance by considering the gene neighborhood in metagenomes for broad taxonomical groups. The improved performance of the metagenome-based networks may be attributed to the unculturable microbes in metagenome, which provide substantial increase in both quantity and diversity of prokaryotic genomic sequences. For example, comparison of the edge information between the two different gene neighborhood networks showed that 14.6%, 13.5%, 15.8%, and 12.2% of the top 10,000 links overlapped in *A. thaliana*, *D. rerio*, *S. cerevisiae*, and *S. aureus*, respectively. These results imply that the gene neighborhood in metagenomes provides largely novel functional interactions in all tested species.

**Metagenome-based functional links are predictive of virulence-associated phenotypes in *S. aureus***

Next, we assessed the predictive power of the constructed metagenome-based network for prioritizing genes related to phenotypes of medical interest in the human pathogenic bacterium *S. aureus*. We compiled a total of 253 *S. aureus* subsp. USA300_FPR3757 genes associated with 10 virulence-associated phenotypes from three sources as described in the Materials and Methods. For each virulence-associated phenotype, we measured the predictive power of the metagenome-based gene neighborhood network of *S. aureus* using ROC analysis. We found that the metagenome-based network effectively retrieved genes for seven virulence-associated phenotypes (AUC > 0.65) in *S. aureus* (Figure 3). These results indicate that the metagenome-based functional links could be sufficiently predictive for some phenotypes of medical interest.

**Conclusion**

Functional links can be inferred by consideration of the gene neighborhood in not only completely assembled prokaryotic genomes but also in metagenome contigs. The metagenome-based gene neighborhood could effectively retrieve functionally coupled genes in organisms from broad taxonomical groups. We found that the metagenome-based gene neighborhood outperformed that generated based on completely assembled prokaryotic genomes. As the amount of publicly available metagenome sequences grows, we expect that adoption of the metagenome-based network inference method will have a substantial contribution in gaining...
a greater understanding of the functions of genomes in the future.

**Disclosure statement**
No potential conflict of interest was reported by the authors.

**Funding**
This work was supported by a grant from the National Research Foundation of Korea [grant number 2015R1A2A1A15055859], [grant number 2017M3A9B4042581] to I.L.

**References**
Bowers PM, Pellegrini M, Thompson MJ, Fierro J, Yeates TO, Eisenberg D. 2004. Prolinks: a database of protein functional linkages derived from coevolution. Genome Biol. 5:R35. Epub 2004/05/07.

Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 12:59–60.

Caspi R, Billington R, Ferrer L, Foerster H, Fulcher CA, Keseler IM, Kothari A, Krummenacker M, Latendresse M, Mueller LA, et al. 2016. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res. 44:D471–480.

Chen L, Yang J, Yu J, Yao Z, Sun L, Shen Y, Jin Q. 2005. VFDB: a reference database for bacterial virulence factors. Nucleic Acids Res. 33:D325–328.

Dandekar T, Snel B, Huynen M, Bork P. 1998. Conservation of gene order: a fingerprint of proteins that physically interact. Trends Biochem Sci. 23:324–328. Epub 1998/10/27.

Fey PD, Endres JL, Yajjala VK, Widhelm TJ, Boissy RJ, Bose JL, Bayles KW. 2013. A genetic resource for rapid and comprehensive phenotype screening of nonessential Staphylococcus aureus genes. MBio. 4:e00537–00512.

Galperin MY, Koonin EV. 2000. Who’s your neighbor? New computational approaches for functional genomics. Nat Biotechnol. 18:609–613. Epub 2000/06/03.

Harrington ED, Singh AH, Doerks T, Letunic I, von Mering C, Jensen LJ, Raes J, Bork P. 2007. Quantitative assessment of protein function prediction from metagenomics shotgun sequences. Proc Natl Acad Sci U S A. 104:13913–13918.

Human Microbiome Project C. 2012. Structure, function and diversity of the healthy human microbiome. Nature. 486:207–214.

Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. 2012. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res. 40:D109–114. Epub 2011/11/15.

Kim H, Shin J, Kim E, Kim H, Hwang S, Shim J, Lee I. 2014. Yeastnet v3: a public database of data-specific and integrated functional gene networks for Saccharomyces cerevisiae. Nucleic Acids Res. 42:D731–736.

Laabei M, Recker M, Rudkin JK, Aledjawi M, Gulay Z, Sloan TJ, Williams P, Endres JL, Bayles KW, Fey PD, et al. 2014. Predicting the virulence of MRSA from its genome sequence. Genome Res. 24:839–849.

Lee I, Date SV, Adai AT, Marcotte EM. 2004. A probabilistic functional network of yeast genes. Science. 306:1555–1558. Epub 2004/11/30.

Lee T, Yang S, Kim E, Ko Y, Hwang S, Shin J, Shim JE, Shim H, Kim H, Kim C, et al. 2015. Aranet v2: an improved database of co-functional gene networks for the study of Arabidopsis thaliana and 27 other nonmodel plant species. Nucleic Acids Res. 43:D996–1002.

Overbeek R, Fonstein M, D’Souza M, Pusch GD, Maltsev N. 1999. The use of gene clusters to infer functional coupling. Proc Natl Acad Sci U S A. 96:2896–2901. Epub 1999/03/17.

Shim H, Kim JH, Kim CY, Hwang S, Kim H, Yang S, Lee JE, Lee I. 2016. Function-driven discovery of disease genes in zebrafish using an integrated genomics big data resource. Nucleic Acids Res. 44:9611–9623.

Shin J, Lee T, Kim H, Lee I. 2014. Complementarity between distance- and probability-based methods of gene neighbourhood identification for pathway reconstruction. Mol Biosyst. 10:24–29.

Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, Djahanschiri B, Zeller G, Mende DR, Alberti A, et al. 2015. Ocean plankton. Structure and function of the global ocean microbiome. Science. 348:1261359.

Wang PI, Marcotte EM. 2010. It’s the machine that matters: Predicting gene function and phenotype from protein networks. J Proteomics. 73:2277–2289.