Journal of Microbiological Methods 116 (2015) 44–52

Contents lists available at ScienceDirect

Journal of Microbiological Methods

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

A global network-based protocol for functional inference of hypothetical proteins in *Synechocystis* sp. PCC 6803

Lianju Gao, Guangsheng Pei, Lei Chen, Weiwen Zhang *

Laboratory of Synthetic Microbiology, School of Chemical Engineering & Technology, Tianjin University, Tianjin 300072, PR China
Key Laboratory of Systems Bioengineering, Ministry of Education of China, Tianjin 300072, PR China
SynBio Platform, Collaborative Innovation Center of Chemical Science & Engineering, Tianjin, PR China

ARTICLE INFO

Article history:
Received 19 April 2015
Received in revised form 24 June 2015
Accepted 25 June 2015
Available online 3 July 2015

Keywords:
Hypothetical proteins
Network
Global algorithm
Operon
Synechocystis

ABSTRACT

Functional inference of hypothetical proteins (HPs) is a significant task in the post-genomic era. We described here a network-based protocol for functional inference of HPs using experimental transcriptomic, proteomic, and protein–protein interaction (PPI) datasets. The protocol includes two steps: *i*) co-expression networks were constructed using large proteomic or transcriptomic datasets of *Synechocystis* sp. PCC 6803 under various stress conditions, and then combined with a *Synechocystis* PPI network to generate bi-colored networks that include both annotated proteins and HPs; *ii*) a global algorithm was adapted to the bi-colored networks for functional inference of HPs. The algorithm ranked the associations between genes/proteins with known GO functional categories, and assumed that the top one ranked HP for each GO functional category might have a function related to the GO functional category. We applied the protocol to all HPs of the model cyanobacterium *Synechocystis*, and were able to assign putative functions to 122 HPs that have never been functionally characterized previously. Finally, the functional inference was validated by the known biological information of operon, and results showed that more than 70% HPs could be correctly validated. The study provided a new protocol to integrate different types of OMICS datasets for functional inference of HPs, and could be useful in achieving new insights into the *Synechocystis* metabolism.

1. Introduction

Hypothetical proteins (HPs) are proteins predicted from nucleic acid sequences and that have not been demonstrated by any experimental evidence (Lubec et al., 2005). They constitute about 25–40% of all open reading frames in well-studied model microorganisms, such as *Escherichia coli* (Galperin and Koonin, 2004). However, this proportion could be as high as 50% in the genomes of less-studied microorganisms (Doerks et al., 2012). Since HPs compose a considerable fraction of proteomes in microbial genomes, it is possible that they are serving important and even novel biological roles (Adams et al., 2007; Desler et al., 2009; Eisenstein et al., 2000). In fact, it becomes clear that the existing of a large number of HPs in the genomics data has significantly restricted the effort in deciphering microbial metabolism, raising urgent needs to develop new experimental and computational methods to decode or infer functions of HPs (Mazandu and Mulder, 2012).

Cyanobacteria are autotrophic prokaryotes that can perform oxygenic photosynthesis, similar to that performed by higher plants (Rippka et al., 1979). According to a recent survey of CyanoBase (http://genome.microbedb.jp/cyanobase), 30–60% of the putative proteins are HPs in various cyanobacterial genomes. The model cyanobacterial species, *Synechocystis* sp. PCC 6803 (here after *Synechocystis*) is the first phototrophic organism sequenced (Kaneko et al., 1995, 1996), and significant researches have been conducted on it (Govindjee, 2011). However, even with many years' efforts in improving its genome annotation, the *Synechocystis* genome still contains a large proportion of HPs, with nearly 33% of all putative proteins are still annotated as HPs (Qiao et al., 2013a). Meanwhile more evidence is emerging that HPs may play important physiological roles in *Synechocystis*. For example, HP Slr1799 was found to be involved in response to salt stress (Karandashova et al., 2002) and chloroplast HP Slr0374 is involved in the regulation of CO_2 utilization in *Synechocystis* (Jiang et al., 2015).

Since experimental characterization of protein function cannot accommodate the vast amount of HPs already available in the database (Liolios et al., 2010), the computational-based annotation has therefore been proposed as one useful and practical mean in inferring function of HPs, and providing functional clues for further experimental validation (Radivojac et al., 2013). Among various computational methods, the network-based approach has attracted significant attention in deciphering potential function since it is not dependent of sequence similarity (Kourmpetis et al., 2010).

* Corresponding author at: Laboratory of Synthetic Microbiology, School of Chemical Engineering & Technology, Tianjin University, Tianjin 300072, PR China.
E-mail address: wwzhang8@tju.edu.cn (W. Zhang).

http://dx.doi.org/10.1016/j.mimet.2015.06.013
0167-7012/

Two types of network-based approaches have been previously established: direct annotation and module-assisted scheme (Sharan et al., 2007). The direct annotation scheme is guided by the principle that proteins that lie closer to one another in the network are more likely to have a similar function (Deng et al., 2003; Karaoz et al., 2004; Letovsky and Kasif, 2003; Mostafavi et al., 2008; Vazquez et al., 2003), while the module-assisted scheme defines protein modules based on connectivity of different proteins in the network and then assigns possible functions to proteins based on associated member proteins with known functions (Arnau et al., 2005; Becker et al., 2012). So far a number of algorithms have been developed for both direct- and module-assisted function predictions, and a fraction of them have even been implemented and supported with graphical interfaces (Sharan et al., 2007). In a simplistic comparison of the two approaches, Sharan et al. (2007) applied a simple neighbor-counting method and the more involved module-assisted the molecular complex detection algorithm (MCODE) to protein–protein interaction (PPI) datasets, and found that direct method has a higher specificity in predicting functions when compared to the module-assisted one.

Recently a new method of the direct annotation scheme was developed and used to annotate functions of long non-coding RNAs, and results showed that the inferred functions highly matched to those in the literature (Guo et al., 2012). Using this method as a core, in this paper, we described a network-based protocol for functional inference of HPs using experimental OMICS datasets and existing PPI datasets. To integrate different types of datasets (i.e., genomics, transcriptomics and proteomics), bi-colored biological networks were first constructed using either transcriptomics or proteomics, along with PPI data. In addition, a global propagation algorithm was applied to the networks to infer functions of HPs (Guo et al., 2012). The analysis resulted in functional inference of 122 HPs in the *Synechocystis* genome, which were then validated using operon information. The study provided a new protocol to integrate different types of experimental OMICS datasets for functional inference of HPs.

2. Methods

2.1. Data sources

2.1.1. Proteomic data

A total of eleven iTRAQ LC–MS/MS proteomic datasets of *Synechocystis* from previous studies were obtained. For the first five datasets, *Synechocystis* was grown under ethanol (1.50%, *v*/*v*), butanol (0.20%, *v*/*v*), hexane (0.80%, *v*/*v*), salt stress (4%, *w*/*v*) and nitrogen starvation conditions, respectively. For each condition, cells were harvested at two time points (i.e., 24 and 48 h) that were corresponding to middle-exponential and exponential-stationary transition phases in the growth time courses for proteomics analysis (Huang et al., 2013; Liu et al., 2012; Qiao et al., 2012, 2013b; Tian et al., 2013). In the remaining six datasets, knockout mutants in *Synechocystis* grown under conditions of ethanol (1.60%, *v*/*v*), butanol (0.25%, *v*/*v*), cadmium (4.0 μM), pH 5.0, pH 12.0, and salt stress (4%, *w*/*v*). Cells were harvested at either 36 or 48 h (Chen et al., 2014; Qiao et al., 2012, 2013b; Ren et al., 2014; Tian et al., 2013). The mass spectroscopy analysis was performed using an AB SCIEX TripleTOF™ 5600 mass spectrometer (AB SCIEX, Framingham, MA), coupled with an online micro-flow HPLC system (Shimadzu Co, Kyoto, Japan) as described previously (Unwin et al., 2010). For details regarding experimental design and quality control of the data, please refer to the previous publications (Huang et al., 2013; Liu et al., 2012; Qiao et al., 2012, 2013b; Ren et al., 2014; Tian et al., 2013).

2.1.2. Transcriptomic data

Five RNA-seq transcriptomic datasets of *Synechocystis* from previous studies were obtained. Cells were collected from five stress conditions: 0.2% butanol (*v*/*v*), 1.5% ethanol (*v*/*v*), 0.8% hexane (*v*/*v*), 4.0% NaCl (*w*/*v*), and nitrogen-starvation treatments. For each condition, cells were harvested at three time points (i.e., 24, 48 and 72 h) for transcriptomic analysis. For details regarding experimental design and quality control of the data, please refer to the previous publications (Huang et al., 2013; Liu et al., 2012; Qiao et al., 2013b; Wang et al., 2012; Zhu et al., 2013).

2.1.3. PPI dataset and annotation information

A PPI dataset of *Synechocystis* was downloaded from the STRING database (http://www.string-db.org/) (Jensen et al., 2009). In the STRING database, several types of evidences for the association, including genomic context, high-throughput experiments, conserved coexpression and previous biological knowledge were used to calculate a single combined_score for each gene in the genome. In this study, the combined_scores indicative of a higher confidence than other single evidence, were applied to construct the PPI network to cover potential protein–protein connections (Szklarczyk et al., 2011).

To describe protein function, we used the classification scheme provided by the biological process (BP) of the Gene Ontology (GO) Consortium (Ashburner et al., 2000; Lægreid et al., 2003). The known 'gene2go' associations in the *Synechocystis* genome were downloaded from CyanoBase database (Nakamura et al., 1998).

2.2. Missing value estimation

To improve the quality of imputation of missing proteomic values, three imputation methods were first implemented and evaluated, they were: the method based on *K* nearest neighbors (KNN) algorithm (Thirumahal and Patil, 2014), the local least squares imputation (LLSimpute) method (Kim et al., 2005), and the imputation method based on chained equations named predictive mean matching (PMM) (Souverein et al., 2006). To compare error rates for each method, a set of values were randomly chosen from a proteomic dataset and removed to generate an incomplete proteomic dataset at certain missing rates. Because the real values are known, the estimation error can be calculated. These methods were evaluated according to normalized root-mean-square error (NRMSE) values:

$$\mathrm{NRMSE} = \frac{\sqrt{\sum_{\mathrm{j}=1}^{\mathrm{N}} \left(y_j - \hat{\mathrm{y}}_{\mathrm{j}}\right)^2 / \mathrm{N}}}{\sigma_{\mathrm{y}}}$$

where y_j is the real value, $\hat{\mathrm{y}}_\mathrm{j}$ is the estimated value, and σ_y is the standard deviation for the N true values (Kim et al., 2005). The value of NRMSE is between 0.0 and 1.0. A smaller NRMSE value means a higher accuracy. The evaluation criteria have been applied in several previous studies (Feten et al., 2005; Meng et al., 2014).

2.3. Construction of bi-colored networks

The procedure to construct co-expression network includes:*i*) data normalization was performed with the raw proteomic or transcriptomic data converted into a ratio of condition versus its control, and then log2 transformed; *ii*) correlation values were calculated for all possible pairs, in which correlation was defined as the Pearson correlation coefficient for all pairwise genes/proteins; *iii*) normalized the correlation coefficient using a Min–Max linear normalization algorithm developed previously (Guo et al., 2012), and then the co-expression network was constructed in which the weight of edge represents Pearson correlation coefficients, and the node of network represents genes/proteins (Pei et al., 2014); *iv*) a correlation coefficient cutoff was applied to the co-expression network, where only gene/protein pairs with a correlation coefficient higher than the cutoff were considered connected. As the biological networks behave like a scale-free network (Tsoi et al., 2014), the distribution of connections follows power-law relationship. To select suitable correlation coefficient cutoff for co-expression networks,