Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Biotechnology

Exploiting transcriptomic data for metabolic engineering: toward a systematic strain design

Minsuk Kim[1], Beom Gi Park[2], Joonwon Kim[2], Jin Young Kim[2] and Byung-Gee Kim[2]

Transcriptomics is now recognized as a primary tool for metabolic engineering as it can be used for identifying new strain designs by diagnosing current states of microbial cells. This review summarizes current application of transcriptomic data for strain design. Along with a few successful examples, limitations of conventionally used differentially expressed gene-based strain design approaches have been discussed, which have been major reasons why transcriptomic data are considerably underutilized. Recently, integrative network-based approaches interpreting transcriptomic data in the context of biological networks were invented to provide complimentary solutions for metabolic engineering by overcoming the limitations of conventional approaches. Here, we highlight recent pioneering studies in which integrative network-based methods have been used for providing novel strain designs.

Addresses
[1] Institute of Engineering Research, Seoul National University, Seoul 08826, Republic of Korea
[2] School of Chemical and Biological Engineering, Institute of Molecular Biology and Genetics, and Bioengineering Institute, Seoul National University, Seoul 08826, Republic of Korea

Corresponding author: Kim, Byung-Gee (byungkim@snu.ac.kr)

Current Opinion in Biotechnology 2018, **54**:26–32

This review comes from a themed issue on **Analytical biotechnology**

Edited by **Hiroshi Shimizu** and **Fumio Matsuda**

https://doi.org/10.1016/j.copbio.2018.01.020

Introduction

In recent years, microbial cell factories for sustainable production of fuels and chemicals are in great demand to combat environmental crises caused by excessive use of fossil resources. Enhancing strain performance to make a process economically viable is a daunting task which requires enormous cost, time, and effort. However, recent revolution in metabolic engineering strategies has dramatically reduced resource requirements and accomplished development of successful processes [1–3]. Without a doubt, systems biology lies at the heart of the revolution by providing a holistic and quantitative view of cellular systems to be engineered thus enabling identification of novel strain designs [4,5]. Systems biology tools for metabolic engineering include data-driven (or top-down) methods which rely on high-throughput data (also known as 'omics' data such as transcriptomic, proteomic, metabolomic, and fluxomic data) to fully characterize different cellular components, and model-driven (or bottom-up) methods which use mathematical models to predict or prioritize effective strain designs [6,7]. Both methods are complementary by nature, as we can see from some examples in which both methods were used in a combinatorial way to create successful microbial cell factories [8,9].

Our ability to generate various omics data has been drastically improved in recent years based on advances in sequencing technology and mass spectrometry methods. As a result, a wealth of such data is now available for a variety of biotechnologically important microorganisms [10,11], and even more multi-omics data are increasingly generated to more comprehensively understand physiology of a cell [12–14]. Nevertheless, transcriptomics is the most frequently used omics tool due to its well-established experimental protocols, easiness for obtaining genome-wide data, and straight-forwardness of interpretation [5]. Therefore, analysis of transcriptomics-driven metabolic engineering examples could still provide a decent view of how data-driven systems biology approaches can be used for development of microbial cell factories. In this review, we will discuss the current use of transcriptomics data for metabolic engineering and possible ways to better utilize the data through highlighting some pioneering works.

Underutilized state of transcriptomic data

To investigate how transcriptomic data is being used in metabolic engineering, we first conducted a PubMed search using a keyword '"Metabolic Engineering" AND ("Transcriptomic" OR "Transcriptome")'. The keyword retrieves only 260 publications (as of October 2017) while actual metabolic engineering had not been performed in many of them. This number is quite surprisingly small considering that more than 5000 series of transcriptomic data for all bacterial species (and more than 800 series of data only for *Escherichia coli*) are currently available in NCBI Gene Expression Omnibus

database, the largest repository of transcriptomic data. Such underutilized state of transcriptomic data (and also other types of omics data) can be more clearly seen in the LASER (Learning Assisted Strain EngineeRing) database recently developed to provide standardized records for metabolic engineering designs of *E. coli* and *Saccharomyces cerevisiae*, the two most important chassis strains [15,16•]. Despite the prevalence of systems biology tools for metabolic engineering, still most of the designs are made by human intuition, and there are only 13 papers (less than 3% of the total listed papers) in which transcriptomic data were used for making strain designs among 450 papers catalogued in the LASER database.

Current application of transcriptomic data

To seek for potential limitations of current approaches and ways to fully utilize the myriad of data, the current uses of transcriptomic data for metabolic engineering were surveyed. Although transcriptomic data has diverse applications in metabolic engineering, for example, identification of genetic parts for precise control of target gene expression [17,18] and discovery of natural biosynthetic pathways to be heterologously expressed in chassis cells [19,20], here we focus our discussion on only the use of transcriptomic data for system-level strain design, that is, identification of targets for genetic intervention such as knock-out, overexpression or underexpression. As summarized in Table 1, transcriptomics is used to solve a variety of metabolic engineering problems, but it is especially powerful for engineering complex traits (e.g. stress tolerance) which are very difficult to handle with forward engineering tools.

A general procedure for current transcriptomics-driven metabolic engineering is illustrated in Figure 1a. Firstly, transcriptomic data used for strain design is obtained from two or more contrasting experimental conditions or strains that differ in a property to be engineered. For example, transcriptome of *S. cerevisiae* cells pre-cultured in media with and without weak acids was compared to identify genes conferring tolerance to the acid stress [21], and similarly comparative transcriptome analysis was conducted for low-*p*-coumaric acid-producing and high-*p*-coumaric acid-producing *S. cerevisiae* strains to reveal further genetic intervention targets for improving the production [22]. Then, for most cases, in order to identify genetic intervention targets from the transcriptomic data, differentially expressed genes (DEGs) are identified through some computational workflows (for more details, see [28,29]) based on an assumption that DEGs are basis for a given phenotypic difference. Finally, based on another assumption that further overexpression or underexpression of the DEGs will reinforce the phenotypic difference, a small fraction of the DEGs (or very occasionally all the DEGs) are selected for experimental test based on their fold-changes of expression and/or a prior information on functions of the genes. For example, in the abovementioned weak acid stress study [21], 213 and 223 genes were found to be upregulated by more than twofold in formate-acclimated and acetate-acclimated cells, respectively, and two transcriptional/translational machinery-related genes (*RTC3* and *ANB1*) were selected for experimental test among 44 genes which were upregulated in both acclimated-cells. Yeast strains overexpressing *RTC3* and *ANB1* showed not only improved tolerance to the acids but also increased ethanol production. As another example, in the *p*-coumaric acid production study [22], expression of 652 genes was found to be significantly affected by *p*-coumaric acid production levels. Among the DEGs, 24 genes were subjected to experimental test. As a result, knock-outs of seven deregulated genes, which are all involved in transportation of amino acids, polyamines and sugars, in the high-producing strain showed 20–50% improvements in *p*-coumaric acid titer.

Even though such DEG-based strain design approaches have been successful in several cases (Table 1), the DEG-based approaches have a number of limitations which might limit the use of transcriptomic data for metabolic engineering [30]. First, the number of DEGs usually exceeds the number that is experimentally tractable, especially when complex phenotypes such as stress tolerance are being engineered. Therefore, only a fraction of DEGs selected by literature information or human intuition are experimentally tested in the majority of cases including the studies listed in Table 1. However, in many cases, including the two examples introduced above [21,22], clear and solid rationales for the selection are missing in the publication, thus making it difficult for other researchers to mimic the approaches. Often, multi-omics analysis can be used here to further narrow down the number of potential candidates, but usually not enough [9,22]. Second, not all but substantial part of DEGs might be the results of a phenotype change rather than drivers for the change, and it is not easy to distinguish the driver genes from a large number of DEGs. One possible solution to this issue is to focus on differentially expressed transcription factors (TFs) as TFs are usually considered as drivers for a phenotype change [24,25•]. However, there is a non-negligible possibility such that some of the key drivers do not belong to DEGs. For example, a change in TF activity can occur without a significant change in the level of the transcript for the TF [31]. Therefore, DEG-based approaches do not always guarantee the discovery of key drivers responsible for the phenotype change.

Integrative network-based approaches

To cope with the limitations of DEG-based approaches, integrative network-based approaches which use biological networks to interpret omics data in the context of biological networks can be employed [32,33]. Here, essentially all types of biological networks based on