Gene 518 (2013) 26–34

Contents lists available at SciVerse ScienceDirect

Gene

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

YGA: Identifying distinct biological features between yeast gene sets

Darby Tien-Hao Chang, Wen-Si Li, Yi-Han Bai, Wei-Sheng Wu *

Department of Electrical Engineering, National Cheng Kung University, Tainan 70101, Taiwan

ARTICLE INFO

Available online 22 December 2012

Keywords:
Yeast
Enrichment analysis
Biological features

ABSTRACT

The advance of high-throughput experimental technologies generates many gene sets with different biological meanings, where many important insights can only be extracted by identifying the biological (regulatory/functional) features that are distinct between different gene sets (e.g. essential vs. non-essential genes, TATA box-containing vs. TATA box-less genes, induced vs. repressed genes under certain biological conditions). Although many servers have been developed to identify enriched features in a gene set, most of them were designed to analyze one gene set at a time but cannot compare two gene sets. Moreover, the features used in existing servers were mainly focused on functional annotations (GO terms), pathways, transcription factor binding sites (TFBSs) and/or protein–protein interactions (PPIs). In yeast, various important regulatory features, including promoter bendability, nucleosome occupancy, 5′-UTR length, and TF–gene regulation evidence, are available but have not been used in any enrichment analysis servers. This motivates us to develop the Yeast Genes Analyzer (YGA), a web server that simultaneously analyzes various biological (regulatory/functional) features of two gene sets and performs statistical tests to identify the distinct features between them. Many well-studied gene sets such as essential, stress-response, TATA box-containing and cell cycle genes were pre-compiled in YGA for users, if they have only one gene set, to compare with. In comparison with the existing enrichment analysis servers, YGA tests more comprehensive regulatory features (e.g. promoter bendability, nucleosome occupancy, 5′-UTR length, experimental evidence of TF–gene binding and TF–gene regulation) and functional features (e.g. PPI, GO terms, pathways and functional groups of genes, including essential/non-essential genes, stress-induced/-repressed genes, TATA box-containing/-less genes, occupied/depleted proximal-nucleosome genes and cell cycle genes). Furthermore, YGA uses various statistical tests to provide objective comparison measures. The two major contributions of YGA, comprehensive features and statistical comparison, help to mine important information that cannot be obtained from other servers. The sophisticated analysis tools of YGA can identify distinct biological features between two gene sets, which help biologists to form new hypotheses about the underlying biological mechanisms responsible for the observed difference between these two gene sets. YGA can be accessed from the following web pages: http://cosbi.ee.ncku.edu.tw/yga/ and http://yga.ee.ncku.edu.tw/.

1. Introduction

With the advance of high-throughput experimental technologies, biologists easily have two complementary gene sets according to a specific biological feature (e.g. essential vs. non-essential genes (http://www-sequence.stanford.edu/group/yeast_deletion_project/), TATA box-containing vs. TATA box-less genes (Basehoar et al., 2004), head-to-head vs. tail-to-tail genes (Chang et al., 2012a), and induced vs. repressed genes under certain biological conditions or treatments (Gasch et al., 2000)). To investigate the underlying biological mechanisms that cause the difference, researchers have to propose a hypothesis based on their experiences, collect the required data for analysis and then check whether the analysis results concur with the hypothesis. This is a long and tedious try-and-error process. Thus in recent years, gene ontology (GO) enrichment analysis was getting popular in the literature because it can provide functional annotations that help researchers to propose possible hypotheses. A server that can identify distinct biological features between two gene sets largely expedites the analysis process.

More than 70 servers have been developed to find the enriched biological features in the input gene sets (Huang et al., 2009a). However, most of them were designed for identifying enriched biological features in a single gene set. Therefore, these servers cannot be used to compare two gene sets to identify the distinct features between them. Only a small portion of the existing enrichment analysis servers can accept two input gene sets. Unfortunately, the features being tested in most of these servers (e.g. ProfCom (Antonov et al., 2008), BayGO (Vêncio et al., 2006) and GOEAST (Zheng and Wang, 2008)) were mainly focused on functional annotations such as GO terms,

Abbreviations: TF, transcription factor; ChIP-chip, chromatinimmunoprecipitation-chip; GO, gene ontology; UTR, untranslated region.

* Corresponding author. Tel.: +886 6 2757575x62426; fax: +886 6 2345482.
E-mail address: wessonwu@mail.ncku.edu.tw (W.-S. Wu).

0378-1119/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.gene.2012.11.089

pathways, and protein–protein interactions (PPIs). Important regulatory features such as transcription factor (TF)-gene binding, TF–gene regulation, and nucleosome occupancy were not considered in these servers. Providing analyses on these regulatory features helps to construct regulation-related hypotheses.

In yeast, various regulatory features have been shown helpful in distinguishing two gene sets. For example, Lawless et al. showed that the 5′-UTR lengths of stress-repressed genes are significantly shorter than those of stress-induced genes (Lawless et al., 2009). Tirosh et al. showed that the DNA region ~100–200 bp upstream of the start codon in TATA box-less genes has low bendability, but not in TATA box-containing genes (Tirosh et al., 2007). Lin et al. showed that long 5′-UTR genes tend to have much higher nucleosome occupancy near the transcriptional start site (TSS) compared to short 5′-UTR genes (Lin et al., 2010). Wu found that essential genes have a sharply peaked transcription factor binding site (TFBS) distribution, whereas non-essential genes have a dispersed one (Wu, 2011). These observations revealed the importance of providing analyses on regulatory features. Furthermore, many regulatory features in yeast were complete (generated from genome-wide experiments) and have been well organized into several databases (Chang et al., 2011; Hong et al., 2008; Monteiro et al., 2008). Yeast is also the only organism of which the comprehensive (covered more than 200 TFs) ChIP-chip data and TF knockout microarray data are available (Harbison et al., 2004; Hu et al., 2007). The ChIP-chip data, denoted TF–gene binding evidence in the YPA database (Chang et al., 2011), provide experimental evidence showing that a gene could be bound by a TF in vivo. The TF knockout microarray data, denoted TF–gene regulation evidence in the YPA database (Chang et al., 2011), provide experimental evidence showing that the expression of a gene changes significantly owing to the knockout of a TF.

In addition to regulatory features, yeast genes have rich functional annotations such as GO terms, pathways and PPIs. Several important functional groups of genes also have been identified in yeast. First, essential genes for growth on rich glucose media have been identified in the *Saccharomyces* Genome Deletion Project (http://www-sequence.stanford.edu/group/yeast_deletion_project/). The deletion of any one of these genes is sufficient to confer a lethal phenotype. It is estimated that 17.8% of the yeast genome is essential. On the other hand, non-essential genes are those genes when deleted may have some fitness effects but the yeast still can survive. Second, the microarray analysis conducted in Gasch et al. (2000) has identified approximately 900 environmental stress response (ESR) genes, which respond in a stereotypical manner to various environmental stresses. These ESR genes can be divided into two clusters according to whether the genes are repressed or induced due to the stresses. There are about 600 repressed ESR (rESR) and about 300 induced ESR (iESR) genes. Many rESR genes are housekeeping genes, while iESR genes are usually involved in various stress defense mechanisms (Gasch et al., 2000). Third, Basehoar et al. have identified 2114 TATA box-containing genes in the yeast genome (Basehoar et al., 2004) and shown that the fraction of TATA box-containing genes related to stress is higher than that of TATA box-less ones. Fourth, Tirosh and Barkai have identified two classes of genes according to the patterns of nucleosome occupancy in the promoters (Tirosh and Barkai, 2008). The first class of genes exhibited low occupancy close to the TSS and high occupancy at the more distal region, denoted as depleted proximal-nucleosome (DPN) genes. The DPN genes were characterized as having low transcriptional plasticity and low sensitivity to disruption of chromatin regulators (Tirosh and Barkai, 2008). By contrast, the genes that exhibited relatively high nucleosome occupancy close to the TSS coupled and low occupancy at the more distal region are denoted as occupied proximal-nucleosome (OPN) genes. The OPN genes were characterized as having high transcriptional plasticity and sensitivity to chromatin regulation (Tirosh and Barkai, 2008), as well as high level of stochastic fluctuations (Newman et al., 2006) and evolutionary divergence (Tirosh et al., 2006). Fifth, the microarray analysis conducted in Spellman et al. (1998) has identified approximately 800 cell cycle genes whose expressions change periodically during the cell cycle.

Since so many important biological (regulatory/functional) features are available only in yeast, it provides a hotbed for the first analysis server of both regulatory and functional features. This study presents the Yeast Genes Analyzer (YGA), which simultaneously analyzes various biological features of two yeast gene sets and performs statistical tests to identify the distinct features between them. In comparison with the existing enrichment analysis servers, YGA tests more comprehensive regulatory features (e.g. TFBS spatial distribution, promoter bendability distribution, nucleosome occupancy distribution, 5′-UTR length, TF–gene binding and TF–gene regulation) and functional features (e.g. GO terms, PPI, pathways and important functional groups of genes). Including various kinds of biological features raises the challenge of requiring more analysis techniques than only conventional enrichment analysis. Thus, the second contribution of this study is to choose the most appropriate statistical tool for analyzing different biological features according to their mathematical characteristics.

2. Material and methods

In brief, the YGA is an analysis platform over various biological features, which consists of three components. The first component, data collection, is responsible for collecting 15 biological features from databases and articles and preprocessing them for efficient retrieval. The second component, analysis tools, includes a set of programs responsible for statistically analyzing the collected biological features with different mathematical characteristics such as real or Boolean numbers. The last component, web server, is responsible for extracting the data according to users' queries, invoking the analysis tools and presenting the analysis results. The details of each component are described below.

2.1. Data collection

YGA collected 15 biological features from six databases and eight articles. First, the genomic locations of TFBSs of 76 TFs (predicted by the Phylogibbs algorithm) were collected from SwissRegulon (Pachkov et al., 2007). Second, the bending propensity of each tri-nucleotide was retrieved from Brukner et al.'s work (Brukner et al., 1995). They used DNase I digestion experiments to estimate the bending propensity. Third, the nucleosome occupancy at every base pair in the yeast genome was retrieved from Kaplan et al.'s work (Kaplan et al., 2008). The nucleosome occupancy at every base pair is calculated as the log-ratio between the number of reads that cover that base pair and the average number of reads per base pair. Fourth, the genomic locations of the TSSs and 5′-UTRs of 4560 genes were retrieved from Nagalakshmi et al.'s work (Nagalakshmi et al., 2008). Fifth, the binding evidences of 28,838 TF-promoter pairs based on band-shift, footprinting or ChIP assays in the literature were retrieved from YEASTRACT (Monteiro et al., 2008). This feature, denoted as TFB, tells whether a specific TF can bind to a target promoter. Sixth, the regulation evidences of 21,847 TF–gene pairs based on TF knockout assays in the literature were retrieved from YEASTRACT (Monteiro et al., 2008). This feature, denoted as TFR, tells whether the expression of a target gene would change significantly in response to a specific TF knockout. Seventh, 300,617 protein–protein interactions (PPI) with experimental evidence were retrieved from BioGRID (Stark et al., 2011). All the PPIs were identified by biological experiments, though they still contained some false positives (Yu et al., 2010). Eighth, 99 manually curated pathways were retrieved from KEGG (Kanehisa et al., 2012). Ninth, the genomic locations of 2953 TATA boxes in the promoters of 2114 genes were retrieved from Basehoar et al.'s work (Basehoar et al., 2004). Tenth, the list of 544