BioSystems 139 (2016) 23–28

Contents lists available at ScienceDirect

BioSystems

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

Influence factors on the correlations between expression levels of neighboring pattern genes

Xiang-Jun Cui*, Lu Cai, Yong-Qiang Xing, Xiu-Juan Zhao, Chen-Xia Shi

School of Mathematics, Physics and Biological Engineering, Inner Mongolia University of Science and Technology, Baotou 014010, China

ARTICLE INFO

Article history:
Received 23 March 2015
Received in revised form 7 October 2015
Accepted 23 November 2015
Available online 13 December 2015

Keywords:
Gene expression
Histones
Acetylation
Methylation
Gene Ontology
Semantics

ABSTRACT

Some genes tend to cluster and be co-expressed. Multiple factors affect gene co-expression. In this study, we investigated the relationships between multiple factors and the correlations of expression levels of neighboring genes, which were divided into four kinds of pattern genes and one type of non-pattern gene. Our results indicate that the correlation between expression levels of neighboring non-pattern genes is related to multiple factors with the exception of transcriptional orientations of neighboring genes. The correlation between expression levels of neighboring specific genes or neighboring repressed genes is likely to be dependent on the co-functions of neighboring genes. The correlation between expression levels of neighboring housekeeping genes is associated with histone modifications in intergenic regions.

1. Introduction

In general, eukaryotic gene expression is controlled by physiological influences and chromosomal genetic architecture (Ogawa and Miyake, 2015; Talebzadeh and Zare-Mirakabad, 2014). Genes that exhibit differential expression behaviors are called pattern genes (Pan et al., 2012). Many studies of gene expression have focused on four kinds of pattern genes: housekeeping genes (HKgene), specific genes (Spgene), selective genes (Segene) and repressed genes (Regene). HKgenes comprise a group that is necessary for the maintenance of basic cellular functions. Expression levels of HKgenes remain constant in all cell types of an organism under normal physiological conditions (Butte et al., 2001; Eisenberg and Levanon, 2003; Greer et al., 2010). Spgenes and Segenes are defined as genes that are exclusively or preferentially expressed in response to specific physiological stimuli (Liang et al., 2006). Regenes are expressed under almost all conditions, and play critical roles in cellular functional differentiation (Thorrez et al., 2011).

Gene expression is regulated by a variety of factors. In the case of epigenetic domains, these include histone modifications, chromatin remodeling, nucleosome positioning, and DNA methylation. Histone modifications have received much attention as regulators of gene expression, with a variety of such modifications (methylation, acetylation, phosphorylation, ubiquitination, and sumoylation) occurring in the N-terminal tail domains of core histones (Strahl and Allis, 2000). Lysines and arginines can be targets of methylation. Lysine can undergo mono-, di-, or trimethylation, while arginine can undergo mono- or di-methylation. Methylations are catalyzed by histone methyltransferases, and there are 24 known sites of methylation (17 lysine residues and 7 arginine residues) (Bannister and Kouzarides, 2005). Importantly, methylations are correlated with gene silencing or gene activation (Zhang and Reinberg, 2001). For example, the histone modifications H3K4me3 and H3K27me3 correlate with gene activation and repression, respectively (Lui et al., 2014). H3K9me3 is involved in heterochromatin formation and correlates with gene repression (Schones and Zhao, 2008). Various histone acetyltransferases specifically modify different acetylation sites (Kurdistani et al., 2004; Lee and Workman, 2007). In general, acetylation of core histone tail correlates with transcription activation (Zhang and Reinberg, 2001).

The distribution of genes in eukaryotic genomes is not random. Some genes have a tendency to form gene clusters and co-express (Deng et al., 2010; Lercher et al., 2002). Functionally related genes involved in the same metabolic pathways or in the same biological processes tend to cluster with each other within the genome. Expression levels of genes in close juxtaposition are typically coordinated and occur at similar levels (Deng et al., 2010; Hurst et al., 2004; Lee and Sonnhammer, 2003). Co-expression phenomena have been studied and explained by factors such as the existence of operons, gene-function correlation, inter-gene physical distance,

* Corresponding author.
E-mail address: cuinmgkjdx@126.com (X.-J. Cui).

http://dx.doi.org/10.1016/j.biosystems.2015.11.007
0303-2647/

sharing similar regulatory elements, and modulation of chromatin conformation.

Operons are rarely observed in eukaryotes, suggesting that they play an insignificant role in gene co-expression (Michalak, 2008). Some reports indicate that co-expression of neighboring genes is dependent on intergenic distance (Tsai et al., 2009). Transcriptional directions of neighboring genes may also affect co-expression of neighboring genes. Divergent gene pairs are more likely to share the same regulatory system (Deng et al., 2010), so co-expression levels of divergent gene pairs are predicted to be significantly higher than those of parallel pairs. However, co-expression of neighboring gene pairs is not noticeably related to the shared TFs (Tsai et al., 2007). Neighboring genes in any orientation are likely to be co-expressed (Cohen et al., 2000). Another factor controlling co-expression of neighboring genes is their shared chromatin environment (Hurst et al., 2004; Batada et al., 2007). Propagations along the chromatin fiber of histone-modifying enzymes enable the chromatin to form an extended domain, and neighboring genes within the domain may share similar regulatory elements and co-express (de Wit and van Steensel, 2009).

Although co-expression of neighboring genes has been studied from multiple perspectives, little is known about the influence factors on the correlations between expression levels of neighboring pattern genes. In this paper, we investigated the factors including transcriptional orientation, intergenic physical distance, correlation of identical histone modifications in neighboring coding regions, modification levels in intergenic regions, and functional similarity of neighboring genes. The results show that there are some differences in influence factors for different pattern genes and non-pattern gene.

2. Materials and methods

2.1. Filtering pattern genes

Pattern gene datasets were obtained from the Pattern Gene database (PaGenBase) (Pan et al., 2013). Because of current limitations of molecular technologies able to identify pattern genes, Pan et al. differentiated the pattern genes by using four statistical parameters which contained a Specificity Measure (SPM), a Dispersion Measure (DPM), a Contribution Measure (CTM), and a Repression Measure (RPM), with these being based on the following public high-throughput microarray repositories: NCBI GEO (Barrett et al., 2011), GNF BioGPS (Wu et al., 2009) and EBI ArrayExpress (Parkinson et al., 2011). SPM, DPM, CTM and RPM were used to measure the specificity of gene expression in a sample, the standard deviation of the specificity, the enhancement of gene expression levels, and the significance of expression levels of repressed genes, respectively.

PaGenBase contains 906,599 pattern genes collected from 11 model organisms. There are 27,008 human pattern genes in the database, including 11,868 HKgenes, 5122 Spgenes, 6021 Segenes and 3997 Regenes. There are also some repetitive genes in the four pattern gene datasets: with a gene being considered repetitive when assigned to as two or more types of pattern genes, for example A1CF. We excluded repetitive genes from our study to limit interpretational bias. Genes that are neither pattern genes nor repetitive genes are referred as general genes herein. After data filtering, we constructed a database containing 7128 HKgenes, 1719 Spgenes, 1807 Segenes, 1276 Regenes and 12,544 general genes.

2.2. Histone modification data source

Location-specific histone modification data detected by ChIP-Seq experiments in human CD4$^+$ T cells were downloaded from the human histone modification database (HHMD) (http://202.97.205.78/hhmd/Download.jsp) (Zhang et al., 2010). HHMD contains experimental human histone modification data. The histone modifications we studied included 18 kinds of acetylations (H2AK5ac, H2AK9ac, H2BK5ac, H2BK12ac, H2BK20ac, H2BK120ac, H3K4ac, H3K9ac, H3K14ac, H3K18ac, H3K23ac, H3K27ac, H3K36ac, H4K5ac, H4K8ac, H4K12ac, H4K16ac, H4K91ac) and 20 kinds of methylations (H2BK5me1, H3K4me1, H3K4me2, H3K4me3, H3K9me1, H3K9me2, H3K9me3, H3K27me1, H3K27me2, H3K27me3, H3K36me1, H3K36me3, H3K79me1, H3K79me2, H3K79me3, H3R2me1, H3R2me2, H4K20me1, H4K20me3, H4R3me2). Histone modification data on the X and Y chromosome were eliminated from our modification dataset.

2.3. Human genome data

The human genome database (release hg18) contains 32,609 genes (NCBI Map Viewer (http://www.ncbi.nlm.nih.gov/projects/mapview/map_search.cgi?taxid=9606&build=105.0), and was used to identify the neighboring genes studied herein.

In this work, neighboring genes were defined based on genomic coordinates. Two consecutive genes on same or different chromosomal strand were considered to be neighboring genes.

2.4. Measuring the functional similarity of neighboring genes

Functional similarity (co-functions) of neighboring genes can be measured via semantic similarity based on GO annotations of neighboring gene products. The GO database contains three ontologies: cellular component, molecular function, and biological process. Each ontology is structured as a directed acyclic graph (DAG), wherein nodes and edges represent terms describing components of gene product function and defined relationships between terms, respectively (Ashburner et al., 2000).

To measure the semantic similarity of GO terms, we used the shortest semantic differentiation distance (SSDD) of Xu et al. (2013). They proposed the method based on the concept of semantic totipotency. The capacity of semantic differentiation is called semantic totipotency. In Eq. (1), the semantic totipotency of a given term t is calculated. Here T-value ($T(t)$) represents the semantic totipotency.

$$T(t) = \begin{cases} 1 & \text{if } t = r \\ \underset{t_p \in \text{parent of } (t)}{\text{mean}} (\omega \cdot T(t_p)) & \text{if } t \neq r \end{cases} \quad (1)$$

where r and ω represent a root term and the degree of semantic differentiation of a edge linking term t to its parent t_p, respectively. ω can be calculated using Eq. (2):

$$\omega = \frac{Dst(t)}{Dst(t_p)} \quad (2)$$

$Dst(t_p)$ and $Dst(t)$ represent the number of parents of the term t and its descendants including itself, respectively. The semantic similarity is defined by Eq. (3):

$$\text{Sim}_{ssdd}(t_A, t_B) = 1 - \frac{\arctan\left(\min\left\{\sum_{t \in path(t_A, t_B)} T(t)\right\}\right)}{\pi/2} \quad (3)$$

Here we first determine the shortest path linking t_A to t_B. The path must pass through lowest common ancestors of t_A and t_B which are nearer to the root. In Eq. (3), path(t_A, t_B) represents a set of terms on the shortest path.