Available online at www.sciencedirect.com
ScienceDirect
Gene 410 (2008) 53–66
GENE
www.elsevier.com/locate/gene

Dynamic covariation between gene expression and genome characteristics

Teemu Kivioja [a], Timo Tiirikka [a], Markku Siermala [a], Mauno Vihinen [a,b,*]

[a] Institute of Medical Technology, FI-33014 University of Tampere, Finland
[b] Research Unit, Tampere University Hospital, FI-33520 Tampere, Finland

Received 2 January 2007; received in revised form 13 November 2007; accepted 29 November 2007
Available online 8 December 2007
Received by M. Di Giulio

Abstract

Gene and protein expression is controlled so that cells can react to changing intra- and extracellular signals by modulating biochemical networks and pathways. We have previously shown that gene expression and the properties of expressed proteins are dynamically correlated. Here we investigated correlations between gene related parameters and gene expression patterns, and found statistically significant correlations in microarray datasets for different cell types, organisms and processes, including human B and T cell stimulation, cell cycle in HeLa cells, infection in intestinal epithelial cells, *Drosophila melanogaster* life span, and *Saccharomyces cerevisiae* cell cycle. Our method was applied to time course datasets individually for each time point. We derived from sequence information numerous parameters for nucleotide composition, two-base composition, codon usage, skew parameters, and codon bias. In addition to coding regions, we also investigated correlations for complete genes and introns. Significant dynamic correlations were identified for each of the analyses. Our method also proved useful for detecting dynamic shifts in gene expression profiles, such as in the *D. melanogaster* dataset. Detection of changes in the properties of expressed genes and proteins might be useful for predicting or following biological processes, responses, growth, differentiation and possibly in related disorders.

Keywords: Codon usage; Nucleotide composition; Gene expression patterns; Properties of genes; Coexpressed genes

1. Introduction

Gene expressivity is dependent on and related to many factors such as codon usage, the promoter region, and mRNA stability. Codon usage, i.e. the biased use of synonymous codons, appears in all organisms and can be calculated in different ways (Ikemura, 1981b; Bennetzen and Hall, 1982; Ikemura, 1985; Wright, 1990). Codon bias is thought to result from the balance between mutation and selection of synonymous codons. Codon usage correlates with gene expression both in unicellular and multicellular organisms (Ikemura, 1981b; Gouy and Gautier, 1982; Ikemura, 1985; dos Reis et al., 2003). Codon usage also has a correlation with the abundance of tRNAs (Ikemura, 1985) and the tRNA gene number (Duret, 2000).

Codon usage is a fundamental property of an organism. In addition to gene expression, it is correlated to tissue-specific expression in human (Plotkin et al., 2004; Plotkin et al., 2006) and to the subcellular location of the encoded proteins (Chiapello et al., 1999). In *Escherichia coli*, codon usage of expressed genes, along with tRNA abundance, is correlated and varies at different growth rates (Dong et al., 1996). Further, codon usage is related to the function of the protein. Different gene classes have been shown to have different codon usages in bacteria (Fuglsang, 2003), in *E. coli* (Karlin et al., 1998), *Arabidopsis thaliana* (Chiapello et al., 1998), *Oryza sativa* (Liu et al., 2005), and in mammals (Ma et al., 2002). In *A. thaliana* A/T-biased codon usage exhibits a strong tissue-specific expression pattern.

Abbreviations: BCR, B cell receptor; CBI, The Codon Bias Index; CUTG, Codon Usage Tabulated from GenBank; DNA, Deoxyribonucleic acid; EST, Expressed sequence tag; GExCo, Gene Expression Correlation; NCBI, The National Center for Biotechnology Information; RNA, Ribonucleic acid; SOM, Self-Organizing Map; TCR, T cell receptor.

* Corresponding author. Institute of Medical Technology, FI-33014 University of Tampere, Finland. Tel.: +358 3 3551 7735; fax: +358 3 3551 7710.
E-mail address: mauno.vihinen@uta.fi (M. Vihinen).
URL: http://bioinf.uta.fi (M. Vihinen).

0378-1119/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2007.11.018

Codon usage is also correlated with several other factors, including gene length (Moriyama and Powell, 1998; Duret and Mouchiroud, 1999) and mRNA structural stability (Duan and Antezana, 2003). A biased codon distribution also correlates to gene structure and intron location (Comeron and Kreitman, 2002), gene density (Hey and Kliman, 2002), recombination levels in chromosomes (Kliman and Hey, 1993), and location within the DNA strand (Lafay et al., 1999) or chromosome (Daubin and Perriere, 2003).

Interestingly, codon usage also has a correlation with protein structure. The first codon position has been linked to biosynthetic pathways, the second to hydrophobicity, and the third to protein secondary structures and molecular weight (Taylor and Coates, 1989; Siemion and Siemion, 1994). Codon usage and amino acid composition are different in differently folded proteins, whereas in humans only amino acid composition varies (Gu et al., 2004). Several papers have described a correlation between codon usage and protein secondary structures (Adzhubei et al., 1996; Oresic and Shalloway, 1998; Xie and Ding, 1998; Gupta et al., 2000) or protein hydropathy (D'Onofrio et al., 1999). CG3 values for C+G content at the third codon position have a positive correlation with hydropathy, both in prokaryotes and eukaryotes (D'Onofrio et al., 1999), including human (Jabbari et al., 2003). α-Helices are preferentially coded by fast codons, whereas slow codons are utilized in β-strands (Thanaraj and Argos, 1996a; Thanaraj and Argos, 1996b). These reports provide controversial, even conflicting results; however, they indicate that codon usage and gene expression also have correlation with protein structural organization and properties. Common to all these studies is that they present correlations for static systems.

Recently we developed a method to correlate in a dynamic way gene expression and numerous parameters characterizing the encoded proteins (Sharabiani et al., 2005). The results indicated clear yet distinct statistically significant correlations in different cells, organisms and treatments that were analysed using microarray methodology. Here we have extended the method to the DNA level to explore the correlations between codon usage, numerous codon and nucleotide derived parameters within coding regions, full gene sequences, and introns. We observe significant correlations, which provide further insight into biological processes.

2. Methods

Gene expression data was derived from peripheral T cells subjected to seven treatments: mock (untreated) cells; CD3- or CD28-stimulated cells; CD3- and CD28-costimulated cells; and cells treated with ionomycin along with phorbol 12-myristate 13-acetate (PMA), phytohaemagglutinin, or FK506 at 6 time points from 1 to 48 h (Diehn et al., 2002). We did not investigate the chemical treatment results in detail because they were quantitatively similar to CD28 and CD3 co-treatment. Because only a few genes had a significant change in expression in the mock and CD28-stimulated datasets, we focused exclusively on the CD3-stimulated and CD3/CD28-costimulated datasets. In total, 4359 cDNA elements (of approximately 18,000 genes and ESTs on the chip) representing 2240 genes were significantly altered after CD3 stimulation or CD3/CD28 costimulation (Diehn et al., 2002). The B cell dataset for genes involved in maturation in anti-immunoglobulin M-stimulated Ramos cells indicated that 1268 genes had significantly altered expression, at least at one time point (Ollila and Vihinen, 2002; Ollila and Vihinen, 2003).

2166 genes were identified to show cell cycle dependant gene expression pattern in human HeLa cancer cell line analysis (Whitfield et al., 2002). Protein sequences were available for 648 entries and DNA sequences for 1028 genes. We used the dataset for thymidine and noladozole treated cells in 19 time points.

Infection by *Listeria monocytogenes* was investigated on human cultured epithelial cells (Baldwin et al., 2003). 1649 genes and proteins were analysed. Parallel dataset for wild-type *Listeria* with three time points was used in the calculations.

In the *Drosophila melanogaster* microarray dataset, the transcriptional profiles were investigated throughout the life cycle (Arbeitman et al., 2002). RNA expression levels of 6044 genes in wild-type flies were examined at 66 time periods.

A total of 581 genes had altered gene expression patterns in yeast cultures synchronized by four independent methods (α factor arrest, elutriation, cdc28, and arrest of a cdc15 temperature-sensitive mutant) (Spellman et al., 1998). In addition, the effect of treatment with either the G1 cyclin Cln3p or the B-type cyclin Clb2p was studied.

2.1. Data mining

We used numerous bioinformatics methods to filter and merge information regarding gene annotations in a number of databases and further calculated and predicted a large number of characteristics for each gene. A dedicated BioData database (Siermala et al., in preparation) constructed for gene and protein information was extensively used for annotations and sequence identification. To identify the genes, we used the Locus Link ID numbers of ESTs and genes. FlyBase (Crosby et al., 2007) was used to identify *D. melanogaster* sequences and the genome sequence from NCBI was used for those in yeast.

2.2. Sequence-derived variables

A Perl script was written to calculate DNA sequence characteristics. DNA parameters and nucleotide proportions in each codon position were directly calculated from gene sequences. Calculations were also made for groupings of nucleotides into two-base categories: purines R (A and G) and pyrimidines Y (C and T), keto K (G and T) and amino M (A and C) nucleotides, and strong S (C and G) and weak W (A and T) bases (Karlin and Ghandour, 1985). The length indicates the number of nucleotides. The parameters were determined separately for coding regions, introns, and for whole genes (from the beginning of exon 1 to the end of the last exon). Several additional parameters were calculated for coding regions.