Available online at www.sciencedirect.com
ScienceDirect
Genomics 91 (2008) 243–248
GENOMICS
www.elsevier.com/locate/ygeno

Minireview

Coexpression, coregulation, and cofunctionality of neighboring genes in eukaryotic genomes

Pawel Michalak *

Department of Biology, University of Texas Arlington, P.O. Box 19498, Arlington, TX 76019, USA

Received 28 September 2007; accepted 10 November 2007
Available online 21 December 2007

Abstract

Accumulating evidence indicates that gene order in eukaryotic genomes is not completely random and that genes with similar expression levels tend to be clustered within the same genomic neighborhoods. The mechanism behind these gene coexpression clusters is as yet unclear. In this article, plausible biochemical, genetic, evolutionary, and technological determinants of this pattern are briefly reviewed.

Keywords: Gene expression; Genome regulation; Gene coexpression; Gene clusters; Eukaryotes

Introduction

Unlike tightly packed and highly organized prokaryotic genomes, eukaryotic genomes were long believed to be relatively messy. With a few exceptions of conserved gene clusters formed by gene duplications, such as Hox and β-globin genes, average eukaryotic genes were believed to be randomly distributed in genomes and expressed independent of their neighbors. However, it has become increasingly evident that, apart from being controlled individually through promoter sequences and sequence-specific transcription factors, eukaryotic genes are subject to expression regulation dependent on their location within the genome as well. There have been several lines of evidence demonstrating the effects of genomic location on gene expression. First, expression patterns of transgenes are known to vary due to insertion site. For example, an experiment with the assembly of two transgenes controlled by different promoters in an artificial chromatin domain produced highly coordinated expression in tobacco [1]. Second, adjacent duplicated genes in budding yeast exhibit similar patterns of expression [2,3]. Finally, genome-wide expression studies in several organisms, such *Drosophila* [4–6], nematode [7–9], mouse [10–14], human [11–17], and *Arabidopsis* [18–21], have recently showed that genes with similar expression levels are nonrandomly distributed within genomes and tend to cluster within genomic neighborhoods.

This local coexpression, typically measured as a correlation between the expression levels of genes positioned close to each other, can be explained by multiple biochemical, genetic, evolutionary, and technological factors. Such elements of genomic structure as overlapping genes, tandemly duplicated genes, homologous genes, and operons come first to mind as logical candidate determinants of coexpression. Although these elements seem to enrich coexpression clusters in some cases [7], they do not account for the remaining part of the coexpression pattern. Therefore, it has been hypothesized that coexpression of neighboring genes can be determined by chromatin domains, or multigene segments of DNA, which, in a given cell at a given moment, are consistently either euchromatin or heterochromatin [22]. When chromatin opens during gene expression this may simultaneously facilitate expression of genes from neighborhoods of the open region [23–26]. Genomes would thus be compartmentalized into chromatin domains, with their location possibly varying between cell types to deliver tissue-specific chromatin conformation and concerted transcriptional activity [22]. Alternatively, coexpressed gene clusters are formed through local sharing regulatory elements such as transcription

* Fax: +1 817 272 2855.
E-mail address: michalak@uta.edu.

0888-7543/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2007.11.002

factors, promoters, and enhancers. In the human genome, for example, more than 10% of genes form head-to-head pairs that may be subject to bidirectional expression mediated by common promoter sequences [27].

By a heuristic generalization known as "guilt by association" (rather different from its original meaning as a logical fallacy type), it has been reckoned that functionally related genes are organized into coexpression networks, in practice assisting functional annotation of uncharacterized genes. Indeed, physically interacting proteins in yeast tend to be encoded by coexpressed genes [28,29] and expression levels of interacting proteins seem to exhibit coordinated changes across species [30]. This raises the question of whether observed clusters of coexpressed genes are of functional significance. Lee and Sonnhammer [31] observed that genes involved in the same biochemical pathways tend to be clustered together in a number of eukaryotic genomes. In *Arabidopsis thaliana*, genes involved in root development and mitochondrial functions tend to form distinct clusters [32,33]. These observations lead to a tempting generalization that all coexpression of neighboring genes (due to either chromatin domains or common regulatory elements) originates from functional similarities of genes involved in the same metabolic pathways or in the same biological processes. A similar explanation emphasizing functional significance is that coexpression clusters are formed mainly by housekeeping genes that are constitutively expressed across multiple tissues [16,34]. As this review shows, however, the hypothesis of functional significance falls short of providing a universal explanation of the mechanism underlying coexpression clusters.

A critical insight into the significance of coordinated expression can be provided using comparative methods. Evolutionary conservation in the organization of coexpression clusters across various taxa would suggest that this feature has been under selective constraints, indirectly reflecting its functional significance. Such a result may seem to be in conflict with the common intuition that gene expression evolved rapidly despite the high extent of genomic synteny [35,36]. Subsequent studies confirmed the high rate of gene expression evolution but additionally recognized that major reorganizations of regulatory networks could still be adaptive [37–39]. This review makes it evident that to date we lack a comprehensive understanding of the driving force behind formation of eukaryotic coexpression clusters. The picture that emerges here from the reviewed studies suggests that coexpression clusters can be a mosaic of functional (adaptive) and nonfunctional (neutral) genomic domains even within a single genome, in many cases with fuzzy rather than clear-cut boundaries.

From operons to coalitions of neighboring genes

Can coexpression of neighboring genes in eukaryotes be explained by operons? Cotranscription of genes in operons, which is the norm in prokaryotes, is very rare in eukaryotes except for nematodes [40] and trypanosomes [41]. Another exception is provided by the tunicate *Ciona* genome, which contains at least 350 two-gene operons [42]. Operons were also found in flatworms [43] and nucleomorphs of chlorarachniophyte algae [44], and occasional cases of dicistronic units are seen in *Drosophila* [45,46] and humans [47]. In *Caenorhabditis elegans* worms, it is estimated that around 15–25% of genes are contained in operons [48,49]. These operons differ from those in prokaryotes in that polycistronic mRNA is typically translated only at the 5′-proximal cistron and the other cistrons must be activated by *trans*-splicing [50]. In addition to mechanistic differences, nematode operons seem to be distinct from prokaryotic operons by their functional content as well. Mering and Bork [51] estimated that only about 4% of *C. elegans* operons contained two or more genes annotated for the same biological process, compared to 36% in *Escherichia coli* operons.

However, operons do not account for coexpression of all neighboring genes in *C. elegans*. Although genes within operons indeed exhibit the strongest coexpression as measured with microarrays, neighboring genes on the same strand, as well as on opposing strands, show much stronger coexpression compared to genes selected at random over sequence stretches of up to 20 kb, even after duplicate gene pairs were excluded [7]. Another study that focused on expression of muscle-, sperm-, oocyte-, and germ-line-enriched genes in more than 550 diverse microarray experiments also reported coexpression clusters along chromosomes for each of the tissue-specific gene classes [8]. Additionally, Miller and colleagues [9] found that genes expressed during spermatogenesis in *C. elegans* are nonrandomly distributed across the genome and aggregated into three large (48 to 86 genes) clusters on two autosomes.

Higher-level organization of coexpression into clusters of functionally related genes is by no means unique to eukaryotes. Colocalization of functionally related genes outside of operons has been known in bacteria as "uber-operons" [52]. For example, ribosomal genes were found to be clustered into uber-operons in a total of 15 bacterial genomes studied [52]. Genes encoding membrane proteins are located next to another, forming tandem clusters in bacterial, archaeal, and yeast genomes [53].

A correlation between gene proximity and function is evident in the yeast genome [54]. For example, Cho and colleagues [55] found that the expression of adjacent genes is frequently (>25%) initiated in the same phase of the cell cycle. A large fraction of these genes are transcribed from opposite strands, suggesting that a common regulatory system controls the expression of both genes. To test whether the increased correlation of neighboring pairs of yeast genes was due to divergently transcribed promoters, Cohen and colleagues [56] compared the distributions of correlation coefficients for divergent, convergent, and tandem pairs of adjacent genes with a control set of randomized nonadjacent genes. Adjacent genes in all orientations tended to be significantly coexpressed but divergent genes showed the greatest deviation from the control set. However, when the distance between adjacent genes was accounted for, there was no significant difference between the distributions, suggesting that the distance between the genes is a critical factor determining their coexpression. These adjacent genes tended to fall into the same functional categories. Correlated triplets of genes (but not quadruplets) were also