Analysis of operon genes using a compendium of expression data

Brilli M, Fondi M, Fani R

Dipartimento di Biologia Animale e Genetica dell'Università di Firenze, via Romana 17/19, 50125 Firenze, Italia

Motivation

The origin and evolution of operons is still under debate. One of the most accepted hypothesis concerning the molecular mechanisms and forces responsible for their assembly suggests operons origin and evolution be driven by the co-transcription of the genes they contain, as well as by the spatial co-localization of the products they code for. In principle, the transcription of operons into a single polycistronic mRNA implies an equal transcription levels of all genes. However, it is known that some operons, especially the longest ones, may contain internal promoters responsible for the transcription of distal genes. In this work we tried to assess the degree of correlation existing among the transcription levels of genes belonging to the same operon and to check whether the distance of a gene from the transcription start point might influence the degree of its transcription, by using a compendium of published expression data.

Methods

Expression data were downloaded as supplemental material to published papers, corresponding to 10 different publications and a total of 79 different experimental conditions. Only normalized and filtered datasets were used. All the operons (over 1400 genes) from Escherichia coli, retrieved from the regulonDB website (http://regulondb.ccg.unam.mx/index.html) were used. Specific java classes were written to calculate the Pearson correlation coefficients of expression patterns of the first gene in the operon vs all the downstream genes. A set of 10000 pseudo-randomly generated pair of genes was used as a background. We explored the distributions of correlation coefficients and performed linear regression analyses with the distance of a gene from the transcription start as predictor of the correlation among the expression patterns of that gene and the first of the operon.

Results

We checked our expression compendium to detect the significance of the correlation coefficients calculated for operon genes vs random pairs, by analysing the corresponding distributions. Data obtained revealed that the squared-Rs calculated for operon genes are, on average, greater than the corresponding values calculated for pseudo-randomly generated pairs of genes (Wilcoxon Rank Sum test, p0.72; for the random dataset the value is <50%) confirmed that the operon is very effective in maintaining the coexpression of genes. We then explored the relationship existing between gene location in the operon and the correlation coefficients by performing a linear regression analysis using as predictor variable the distance of a gene from the first of the same operon, and dependent variable the correlation coefficient in expression patterns. A statistically significant negative correlation was found (R-squared=0.087, p<5500 nt in length). Nevertheless, very long operons exist with high expression correlations among all genes, and they do not fit well with the model. For this reason, we extracted information concerning the presence of internal promoters (IP) (source: RegulonDB) and we found a positive correlation among operon size and number of IPs. To check the effect of IPs on the expression levels inside operons, we normalized them with respect to the expression level of the first gene in the operon. This revealed that 62% of the genes in operons without IP have a normalized expression between 0 and 1, suggesting that most of them have a reduced expression with respect to the first gene. On the contrary, only 45% of genes in operons with IPs are in the same range, revealing that internal promoters increase the fraction of operon genes with an expression level greater than that of the first gene. In conclusion, the analysis of the E. coli operon dataset suggested that the size may be a limiting parameter during operon evolution. This can explain why very often operons are extremely compact and that the greater the operon length the greater the number of Ips. Therefore, it is possible that the longest operons might have been assembled by a piece-wise mechanism, according to which they have been constructed by the fusion of shorter pre-existing mini-operons rather than the sequential addition of single genes, as suggested for the histidine biosynthetic operons (Fani et al. 2005). This example

reveals that using compendia of expression data can reveal useful information when large datasets are analysed.

Contact email: r_fani@dbag.unifi.it

Supplementary informations
References to microarray data are available upon request, they were not explicited for shortness.