Computational and Structural Biotechnology Journal 13 (2015) 402–406

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

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

Mini Review

The power of operon rearrangements for predicting functional associations

Gabriel Moreno-Hagelsieb

Department of Biology, Wilfrid Laurier University, 75 University Ave. W., Waterloo, ON N2L 3C5, Canada

ARTICLE INFO

Article history:
Received 3 April 2015
Received in revised form 25 June 2015
Accepted 26 June 2015
Available online 2 July 2015

Keywords:
Operons
Operon rearrangement
Operon predictions
Functional associations
Genomic context
Conservation of gene order
Comparative genomics

ABSTRACT

In this mini-review I aim to make the case that operons might be the most powerful source for predicted associations among gene products. Such associations can help identify potential processes where the products of unannotated genes might play a role. The power of the operon for providing insight into functional associations stems from four features: (1) on average, around 60% of the genes in prokaryotes are associated into operons; (2) the functional associations between genes in operons tend to be highly conserved; (3) operons can be predicted with high accuracy by conservation of gene order and by the distances between adjacent genes in the same DNA strand; and (4) operons frequently reorganize, providing further insight into functional associations that would not be evident without these reorganization events.

Contents

1. Introduction

Operons, were first defined as a set of genes transcribed from an operator [1]. By extension, here I define them as two or more adjacent genes in the same strand transcribed into a single messenger RNA (a polycistronic mRNA). It is somewhat expected, as it has been corroborated [2,3], that most genes transcribed into a polycistronic mRNA should code for products that work together. Given the traditionally perceived importance of operons in co-regulating genes whose products functionally interact, they have been central in the field of comparative genomics aiming at predicting functional associations. In this mini-review, I attempt at further justifying this focus. I also attempt at providing evidence that predicted operons in one organism can give clues to functional associations in another organism. Because of the potential transference of functional associations from operons in one organism into genes found in another organism, the power of predicted operons for providing potential associations expands exponentially.

This review is not intended to be a comprehensive view on the methods for predicting functional associations, nor is it intended as a comprehensive view at methods for predicting operons. For further learning about predicting functional associations by genomic context,

E-mail address: gmoreno@wlu.ca.

http://dx.doi.org/10.1016/j.csbj.2015.06.002

and derived methods, the reader can consult such works as [4–8]. For methods on operon predictions the reader can check [9–12] among others.

2. Genes without functions and the panorama of potential interactions

Since the very first genome sequences became available, researchers noticed that a large amount of genes could not be functionally annotated by looking for homologues (see for example [13]). Case in point, a third of the genes in the model organism *Escherichia coli* K-12 MG1655 remain functionally uncharacterized [14] (this is still true today). Inspired by this fact, scientists started proposing methods for predicting operons by methods other than those based in direct homology (for example [15] and references therein).

Predicting functions by methods other than direct homology involves the finding of interactions with the expectation that interactions between unannotated genes and genes with characterized functions (or functionally-annotatable by direct homology), would help predict the functions of the uncharacterized genes. The idea behind transference of functions has been aptly called "guilt by association" [16]. Three main ideas for predicting functions by association appeared: (i) phyletic patterns or phylogenetic profiles [17,18], based on the expectation that if the products of two genes functionally interact, then the genes should co-occur, since the product of one gene would be expected to be useless without the product of the other; (ii) conservation of adjacency [19,20,21], where genes remaining next to each other across genomes are expected to functionally interact; and (iii) gene fusions [22,21], where, if two separate genes in one genome appear as a single fused gene, they might functionally interact.

To put the above ideas in perspective, it is useful to think of the problem of predicting functional interactions as the problem of finding actual interacting pairs among the maximum number of pairs available for exploration in a genome. This exploratory space (E) can be calculated from the total number of annotated genes (N) as:

$$E = \frac{N(N-1)}{2} . \quad (1)$$

Let us consider the case of *E. coli* K12 MG1655 as an illustration. The version of the genome available by November 2014 contains 4138 coding genes. This translates into an exploratory space of 8,559,453 pairs. Considering that the genome consists of a circular chromosome, the maximum number of pairs that could be explored by conservation of gene order would be 4138 (the same as the number of genes), less than 5% of the exploratory space. In theory, the exploratory potential would be much larger for gene fusions, since genes do not have to be adjacent in a genome of interest in order to find them fused in another genome. However, in practice we have found few fused genes (Fig. 4B). The potential for phylogenetic profiles would appear to be the largest. After all, there is no need for the genes to be adjacent in any of the genomes analyzed. However, co-occurrence analyses seem to produce few high-quality annotations (Fig. 4B), perhaps precisely because the background is the total exploratory space, which might consist of a large fraction of true negatives. Thus the question becomes: is it possible to expand on high-quality functional interactions and avoid the enormous number of potential negatives in the exploratory space? The answer seems to be the analyses of operon rearrangements.

3. Operons can be predicted

The problem of predicting operons could be conceptualized as the problem of finding transcription unit (TU) boundaries within a stretch of adjacent genes in the same strand with no intervening genes in the opposite strand. We call these stretches of genes in the same strand "directons" [2] (Fig. 1A).

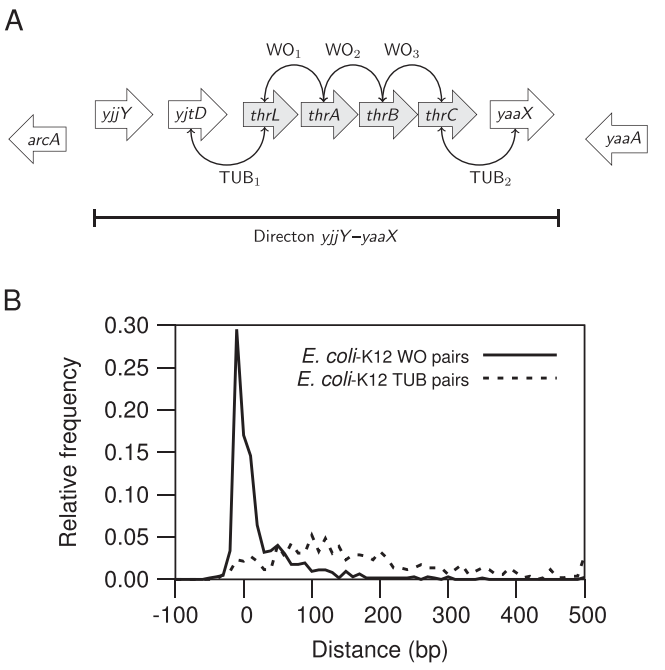

Fig. 1. Intergenic distances. (A) Representation of a directon, a stretch of adjacent genes in the same strand with no intervening gene in the opposite strand. The figure shows an operon within the directon, pairs of genes in operons (WO) and transcription unit boundaries (TUB). (B) The distances between genes in operons tend to be short compared to those between genes in different transcription units. The distances were binned at ten base intervals to calculate relative frequencies.

3.1. Predicting operons by intergenic distances

An initial assumption about genes in operons was that, since there is no need for signals between co-transcribed genes, the distances between genes in the same operon would be shorter than those between genes in different TUs (Fig. 1). The assumption was first confirmed using known operons gathered from the literature as found in RegulonDB [23], mapped into the genome of *Escherichia coli* K12 to find boundaries between TUs [2]. The finding was key in the success of operon predictions from the first time it was used [2,24]. Intergenic distance continues to be the most informative feature for operon predictions [25–27,12].

3.2. Predicting operons by conservation of gene order

Another initial assumption was that operons would have a tendency to be conserved across prokaryotic organisms. Accordingly, some early results in comparative genomics found that adjacent genes in the same strand tend to be better conserved next to each other across genomes than adjacent genes in opposite strands [19,28]. Furthermore, the comparison of conservation of genes in the same strand against that of genes in different strands allowed for high-confidence prediction of operons in genomes with no experimental information on TU organization [29], and for the confirmation that genes in operons have the same tendencies for short intergenic distances among prokaryotes as that observed in *Escherichia coli* [30,24,31].

4. Most genes in prokaryotes are in operons

Some years ago, Cherry [32] published operon estimates based on very simple assumptions. For example, if TUs can be found on either DNA strand, then approximately one fourth of all TUs should be in a strand by themselves. That is, their neighboring TUs would be found in the opposite strand (Fig. 2A). Since there is no reason to expect the length of the TU to influence which ones would be found in a directon