Computers in Biology and Medicine 43 (2013) 738–743

Contents lists available at SciVerse ScienceDirect

Computers in Biology and Medicine

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

Computational identification of operon-like transcriptional loci in eukaryotes

Kishore Nannapaneni [a,b,*], Yehuda Ben-Shahar [e,f,h], Henry L. Keen [a,e,f], Michael J. Welsh [d,e,f], Thomas L. Casavant [a,b,c,g], Todd E. Scheetz [a,b,g]

[a] Center for Bioinformatics and Computational Biology, University of Iowa, Iowa City, IA 52242, USA
[b] Department of Biomedical Engineering, University of Iowa, Iowa City, IA 52242, USA
[c] Department of Electrical and Computer Engineering, University of Iowa, Iowa City, IA 52242, USA
[d] Howard Hughes Medical Institute, University of Iowa, Iowa City, IA 52242, USA
[e] Department of Internal Medicine, University of Iowa, Iowa City, IA 52242, USA
[f] Physiology and Biophysics, University of Iowa, Iowa City, IA 52242, USA
[g] Opthalmalogy and Visual Sciences, University of Iowa, Iowa City, IA 52242, USA
[h] Department of Biology, Washington University, St Louis, MO 63130, USA

ARTICLE INFO

Article history:
Received 9 December 2011
Accepted 15 March 2013

Keywords:
Operon
Trans-spliced
Drosophila
Poly-cistronic
Eukaryote
Computational

ABSTRACT

Operons are primarily a bacterial phenomenon, not commonly observed in eukaryotes. However, new research indicates that operons are found in higher organisms as well. There are instances of operons found in *C. elegans, Drosophila melanogaster* and other eukaryotic species. We developed a prototype using positional, structural and gene expression information to identify candidate operons. We focused our efforts on "trans-spliced" operons in which the pre-mRNA is trans-spliced into individual transcripts and subsequently translated, as widely observed in *C. elegans* and some instances in *Drosophila*. We identify several candidate operons in *Drosophila melanogaster* of which two have been subsequently molecularly validated.

1. Introduction

Transcription in eukaryotes is traditionally described as proceeding one gene at a time. In this model RNA polymerase docks at the promoter of each gene and transcribes them individually. In contrast, polycistronic regulation was, until recently, thought to exist only in prokaryotes. An operon is a group of adjacent genes coordinately regulated by a single promoter. Frequently, genes in operons act cooperatively in a critical biological pathway, and thus the timing of their expression pattern needs to be tightly co-regulated [14]. An operon usually contains two or more genes transcribed as a single transcriptional unit called a polycistronic message. This type of regulation in prokaryotes may compensate for their small genome size by not requiring a promoter in between each pair of genes in an operon [1]. In addition, the clustering of functionally related genes in operons helps to efficiently regulate the genes involved in a common physiologically important pathway. Genes present in operons are adjacent to each other, are usually involved in the same biological pathway, and often share similar expression patterns.

This highly efficient transcriptional mechanism is a characteristic of bacterial regulation. The most popular and well-studied example is the Lac operon in *E.coli* containing the lacZ, lacY and lacA genes all transcribed as a single polycistronic unit from a single promoter and involved in the same biological pathway (lactose metabolism). The ribosomes dock at the translational start site of each gene in the poly-cistronic message and translate them individually.

Recently, the possibility of operon-like transcription in eukaryotes has become a topic of interest. Evidence of operons in eukaryotes was first found in *Caenorhabditis elegans*, a nematode, by [2] where 15% of its genome or approximately 2600 genes are organized into ~1000 operons [3,4]. However, unlike in bacteria, the genes in *C.elegans* operons are transcribed as a single transcription unit and post-transcriptionally spliced into two individual transcripts with a splice leader sequence attached to the 5′ end of both the transcripts. These genes are then translated into protein individually [4].

* Corresponding author at: Center for Bioinformatics and Computational Biology, University of Iowa, 7200, NW 62nd Ave, Iowa City, IA 52242, United States. Tel.: +1 319 400 1505.
E-mail addresses: kishore@eng.uiowa.edu (K. Nannapaneni), benshahary@wustl.edu (Y. Ben-Shahar), henry-keen@uiowa.edu (H.L. Keen), michael-welsh@uiowa.edu (M.J. Welsh), tomc@eng.uiowa.edu (T.L. Casavant), tscheetz@eng.uiowa.edu (T.E. Scheetz).

0010-4825/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.compbiomed.2013.03.004

There is also evidence of the existence of an operon-like transcriptional locus in the fruit fly (*Drosophila melanogaster*). The regulatory mechanism in the reported *Drosophila* operonic locus *CheB42a/Llz* appears to be more similar to that of *C. elegans* than those in prokaryotic operons [5]. The genes *cheB42a* and *llz* in *Drosophila melanogaster*, are transcribed as a polycistronic message from a single promoter and subsequently processed into independent mRNAs [5]. Importantly, no known promoter elements are located in the 97 nt between *CheB42a* and *llz* (i.e., the intercistronic sequence). Moreover, in cell culture, co-immuno precipitation studies provided evidence of a direct physical interaction between these two proteins [5]. This gene pair is similar to the *C.elegans* operons in that the polycistronic pre-mRNA is post-transcriptionally cleaved into two individual mRNA transcripts for each gene. These individual transcripts are polyadenylated and translated individually by ribosomes. One important thing to note is that the downstream transcript for *llz* does not contain a 5′ cap necessary for translation but a single nucleotide guanosine is observed attached to the 5′ end of *llz* which is not present in the genome, leading us to speculate that this single guanosine nucleotide may be essential for translation of the uncapped downstream open reading frame (ORF).

The large number of genes organized as operons in *C. elegans* combined with the presence of a polycistronic locus in *Drosophila melanogaster* is intriguing enough to substantiate the possibility of widespread operon-like transcription in eukaryotes. Discovery of operons in eukaryotes would usher in a new dimension in the way eukaryotic transcription is perceived and the involvement of genes in pathways. In *C. elegans* the genes in operons are functionally related, which is a fundamental characteristic of an operon. The genes in the reported *Drosophila* operon also appear to be functionally related as they both have an important influence on male courtship behavior of the flies [6]. The arrangement of the genes in operons and their involvement in a related function compels us to think in a novel way about their involvement in pathways vis-à-vis their genomic localization. The emerging research from *Drosophila* provides direct experimental evidence suggesting the possible existence of other operons in eukaryotes.

The aim of this reported work is to develop a prototype for a system that can computationally identify operon candidates in eukaryotes on a genomic scale and to demonstrate this prototype for *Drosophila melanogaster*. The prototype is used to computationally generate a prioritized list of candidate operons using the characteristics of the existing operons in the organism, thus allowing for validation of the results. This study provides a list of candidate operons in *Drosophila melanogaster* using the attributes of the *CheB42a /Llz* operon including (1) intra-chromosomal location, (2) strand information, (3) presence of stop codons and polyadenylation signal sequence, (4) evidence of involvement in pathways using publicly available gene ontologies and (5) correlation of expression among genes within the candidate operon. The final set of candidates is provided along with the attributes used in their identification. This allows further refinement and prioritization for subsequent validation

2. Methods

The prototype is comprised of two main parts, the Candidate Selection phase and the Candidate Prioritization phase. A system level figure of the entire process is provided in Fig. 1. The candidate selection phase involves three steps (positional, structural and regulatory) that all candidate operons must satisfy. The candidate prioritization process also involves three modalities. The first modality is annotation based upon Gene Ontology terms associated with genes in a candidate operon. Specifically, shared terms are of interest as an indicator of shared or related function. The second modality is correlated gene expression, which is often used to predict similar gene function or related processes [7]. The final modality is a species-specific feature that is characteristic of the operons in that particular species. The prototype described in this article has been used to identify and prioritize potential operons in *Drosophila melanogaster*.

2.1. Candidate operon selection in drosophila melanogaster

The first step in the selection phase involves a positional criterion, requiring candidate operons to consist of two adjacent genes in the same orientation and within a specific distance from each other. Only the closest 20% of adjacent gene pairs are initially selected. Candidate operons can be identified using Ensembl's Perl API to access the genome of the organism from Ensembl. The second criterion utilizes gene structure information requiring that the upstream gene has a stop codon and that both upstream and downstream genes have a polyadenylation signal sequence. The upstream gene in each candidate operon will be required to contain both an annotated stop codon and one of the two canonical polyadenylation signal sequences (AATAAA or ATTAAA) within the last 50 bp of the annotated gene structure in confirmation of the polyadenylation of both the genes. The final criterion applied in candidate operon identification requires the lack of a promoter element associated with the down-stream gene. The intergenic sequence between each candidate pair can be obtained using Ensembl's Perl API. This sequence can then be analyzed with the promoter prediction program from the Berkeley *Drosophila* Genome Project (http://www.fruitfly.org/seq tools/promoter.html) [8] to identify potential promoters from these sequences. Only pairs of genes that satisfied all three criteria were included in the final set of candidate operons.

2.2. Prioritization methods

The candidate prioritization phase uses public annotation and associated data to quantify which of the candidate operons are most likely to be bona fide operons. This phase may use any number of resources to aid in prioritization depending on their availability and relevance. For this analysis, we utilize three specific data modalities: functional annotation, expression correlation and a species specific feature.

2.2.1. Similar functional annotation

Gene Ontology (GO) annotation was used to assess similarity in function and process [9]. Both the molecular function and biological process categories of GO terms were obtained from the Ensembl *Drosophila* genome databases for all genes of the organism under investigation (http://www.ensembl.org; [10]). The significance of two genes sharing a given annotation term will be assessed as being inversely proportional to the prevalence of the term among all genes. For the GO term **i**, GO_i is used to represent the number of genes annotated with the term **i**. Thus $1/GO_i$, when assessed for all terms i shared among a candidate pair, is maximized for the GO term(s) annotated on the fewest number of genes. When no GO terms are shared in common between a candidate pair a score of 1 is utilized.

2.2.2. Expression correlation process

The second criterion used in the candidate prioritization process is the correlation of gene expression. Hybridizations for several expression studies can be obtained from the Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/) [11]. These datasets all should use the same microarray platform. When multiple probesets are available for a single gene, only the highest expressed probeset will be used. This will result in a final set of

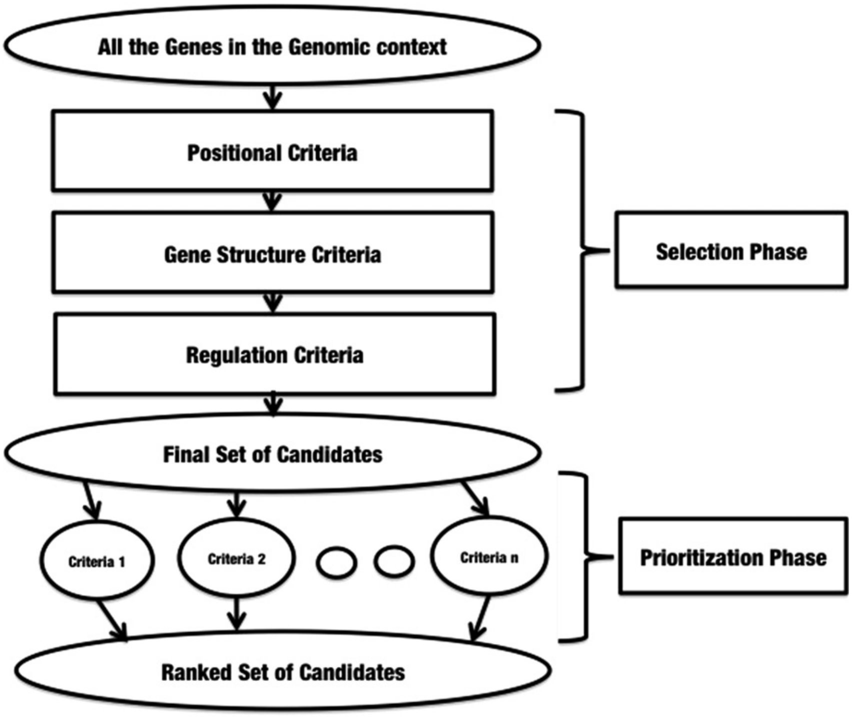

Fig. 1. The generic prototype for identifying candidate operons in eukaryotes.

expression data for which, the Pearson's correlation coefficient will be computed for every pairwise combinations. The significance of a correlation can be determined by the likelihood of a gene pair having a higher correlation.

$\text{signif}(g1,g2)=P(\text{corr}(i,j)>\text{corr}(g1,g2);\ \forall i,\ j\ \text{where}\ i\neq j)$

2.2.3. Species-specific feature

This relates to the mechanism that a particular species utilizes for the transcription of a polycistronic message. If such a mechanism exists and is well understood, this information is very valuable to the identification of operons in that species with greater accuracy. Our prototype utilizes such a mechanism or mechanisms which we call the species-specific features in the identification of operons in eukaryotes. The canonical example of this in eukaryotes is the SL2 splice-leader sequence utilized in the trans-splicing of *C. elegans operons* [2]. Another example is the presence of a leading, untemplated guanine nucleotide incorporated onto the 5′ end of the downstream gene(s) in *Drosophila melanogaster* [5].

2.2.4. Integrated final score

For each candidate operon pair, an objective score can be calculated as shown in the equation below

$\text{score}(c)=-\sum_{\text{criteria}}\log_{10}(\text{sign}_{\text{criteria}}(c))$

Briefly, the significance of each individual criteria is incorporated by adding the negative log of the significance of each criteria for a given candidate operon. The negative log is chosen to normalize the vast differences in the magnitude of significance of each individual criterion. This metric can be utilized to prioritize the final set of candidate operons.

3. Results

We applied the prototype devised for the identification of candidate operons in eukaryotes involving the trans-splicing to *Drosophila melanogaster*. The design of the prototype shown in Fig. 1 considers all the genes in the genome of the species. The prototype follows important criteria for genes to be identified as operons as listed in the Methods section. Following the selection phase, sets of different criteria are applied which depend on available resources for the species in question to prioritize the selected candidate pairs.

3.1. Application of the prototype to identify operons in drosophila

To illustrate this process, we consider the application of the generic prototype shown in Fig. 1 to identify potential operons in *Drosophila melanogaster*. The process starts with the complete set of 14,383 Ensembl-annotated *Drosophila* genes and follows the stepwise procedure outlined in the Methods section above. In the first step, 2788 gene pairs were identified in the same orientation, adjacent to each other and separated by less than 700 bp (the closest 20% of gene pairs are approximately 700 bp apart). Applying the gene structure and regulation criteria further refined this list of candidates to sets of 730 and 410 candidate operons, respectively. The application of these criteria in any order will finally result in the same set of 410 candidates. Of these 410 candidates, 13 pairs listed in Additional Table 2 were found to be tandem duplicates identified by Quijano et al. [15]. These 13 pairs were not eliminated from the final candidate set as they could still be transcribed as a single pre-mRNA. This final set of 410 candidate operons represents ~15% of the 2788 positionally defined gene pairs we initially started with, and only 2.8% of all 14.383 genes. This substantially reduced the number of putative operons by leveraging the known characteristics of an operon.

3.1.1. GO terms

The prioritization scores for the 410 candidate operons are listed in Additional Table 1. Each row represents a distinct candidate; the two constituent genes are listed in the first two columns. The priority scores for each candidate for the Molecular function and Biological process ontologies are provided in the 4th and 5th columns respectively. As indicated earlier in the methods section, GO terms are controlled and curated vocabulary for the function of genes and their products. Our metric quantifies the GO term with least number of genes shared by both the genes for Molecular function and Biological process.

The majority of the candidates did not have a shared GO annotation. For those candidate pairs with mutually consistent annotations, the shared term is often commonly occurring. For example, CG3642 and CG3662 share the term "Protein Binding" (GO:0005515). Note that 4949 other genes are also thus annotated. However, several candidate pairs share very rarely used GO terms. For example, the GO term "intramolecular oxidoreductase activity" (GO:0016860), is associated with only two of the genes identified as candidate operons. Similarly that same pair of genes is uniquely annotated with the term "indole derivative biosynthetic process" (GO:0042435). While such cases increase confidence in our methods, it does not imply that lack of common terms eliminates a candidate pair. As the richness and depth of the GO resource improve, the ability of our method to implicate candidate operons will only be enhanced.

3.1.2. *Gene expression analysis*

The second prioritization metric also assesses the likelihood of shared gene function. This was performed using Pearson's correlation coefficient, which is commonly used to estimate related function [7]. The Affymetrix DrosGenome1 array (GPL72) contains more than 13,000 unique probe sets for the *Drosophila* Genome. In all, a collection of 99 hybridizations was obtained from the following GEO datasets (GDS516,GDS667,GDS653,GDS1068,GDS602,GDS732, GDS664,GDS192). The significance scores based upon the correlation of gene expression for each candidate pair are provided in the third column of Additional Table 1.

The relative distribution of all pair-wise correlation values is presented in Fig. 2. As expected, the distribution appears to be normal, although not centered at 0. This is an expected result given the broad variety of experiments that were utilized. Systematic differences among the tissues and conditions surveyed may lead to apparent correlations. The correlation of the candidates is also shown in Fig. 2.

To assess the similarity of the expression profiles between genes in the candidate operons, we computed the Pearson correlation coefficient using publicly available microarray data. The distribution of correlation coefficients for the candidate operons and all gene pairs is shown in Fig. 2. The X-axis is the Pearson correlation coefficient of the genes in the gene pairs. The Y-axis is the relative frequency metric we developed which is the number of genes with a particular Pearson correlation coefficient divided by the total number of gene pairs in that set. The motivation for the relative frequency is the great disparity between the number of gene pairs in the candidate operons and all gene pairs. The relative frequency is used as a means to normalize for the number of gene pairs in each set and to derive objective estimates for the fraction of gene pairs with a particular Pearson correlation coefficient. The solid line in Fig. 2 indicates the set of pairwise correlations from our set of 410 gene pairs. It is important to note that only 284 out of the 410 candidate gene pairs have probes for both genes on the chip and hence the apparent periodicity at low correlations. Since the relative frequency depended on the number of gene pairs available at each correlation interval and this varied for the candidate gene pairs, this may have resulted in the crests and troughs in the dashed line indicating the candidate gene pairs in Fig. 2. The dotted line is the pairwise correlations of all the probes on the chip. Although the overall curve for the candidate operon pairs appears to be significantly different from the All gene pairs, there is an enrichment in highly correlated pairs in the candidate operon group compared to all gene pairs as shown in Fig. 2. As only a subset of the candidate set is expected to be operons, these highly correlated pairs might constitute a better-prioritized set. The Pearson correlations of the candidate operons are listed in column 3 of Additional Table 1.

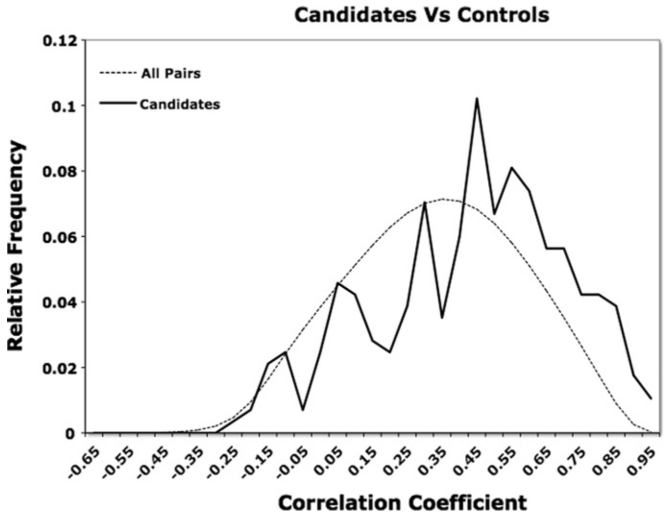

Fig. 2. The comparison of correlation coefficients of the candidate gene pairs and all gene pairs in the *Drosophila Melanogaster* Genome.

3.1.3. *Species-specific feature*

The final criteria used in prioritizing the candidate operons is the presence of a single, untemplated guanine nucleotide at the 5′-most position of the downstream gene unique to *Drosophila melanogaster*. To identify the set of candidates with an untemplated G in the 5′-most position, a hypothetical version of the requisite sequence was derived by prepending a G to the first 100 nt of each downstream gene in the candidate set. These sequences were then blasted versus the database of all *Drosophila* ESTs from dbEST [12]. Hits with perfect alignment to all 101 nt with alignments beginning at the first nt of the EST were counted as evidence of bona fide leading G addition. This procedure required that the preceding position based upon the genome assemble was not a G. The significance of this result was calculated as the number of candidates with an observed leading G, divided by the total number of candidates. Every candidate having a leading G is given this significance value. A significance value of 1.0 was assigned to those candidate pairs in which no G was detected or for which a G is located in the genome.

Of the 410 candidates, 25 were found with evidence of an untemplated leading G at the 5′ position of the downstream gene. Five such gene pairs with a leading G supported by EST evidence are listed in Additional Table 3 as examples. The prioritization score based upon this data is presented in the sixth column of Additional Table 1. As a control for the leading G observation, downstream genes with leading nucleotides were also assessed. These were observed at most once for each of A, C and T across the entire set of 410 candidates. The significance of this result was determined by assessing the prevalence of leading G in sets of genes randomly selected, but specifically not from the set of candidates. This assessment yielded a p-value <0.0005 based upon Fischer's exact test. Given the data from the controls, it is evident that there is enrichment for a leading G in the set of candidate operons. If upon further experimental analysis, it is confirmed that the leading G plays a similar role as the SL sequences in *C.elegans*, this feature could turn out to be the most important prioritization criteria for identifying operon-like loci in *Drosophila*.

3.1.4. Integrative final score

Finally, the prioritization scores are merged and used to prioritize the set of candidate operons. This score was used to sort the candidates in Additional Table 1, and is presented in the seventh column.

In Additional Table 1, columns 1 and 2 are the Ensembl gene ids of 410 candidate operons, column 3 is the significance of the correlation of the genes in the candidate operons, columns 4 and 5 are the GO metrics developed for the candidate operons in biological process and molecular function respectively, column 6 gives the significance of having a leading untemplated G in the downstream gene of the candidate operon and finally column 7 is the integrated score of the significance of the prioritization methods. Additional Table 1 has 410 rows indicative of the 410 gene pairs surviving the identification phase.

4. Discussion

The major finding of the present study is that there are a substantial number of gene pairs in the *Drosophila melanogaster* genome that exhibit several of the characteristics expected to occur in operons. This finding is consistent with the recent molecular evidence suggesting that this type of regulation occurs in *Drosophila* [5]. Moreover, by combining annotation data with results from publicly available microarray studies, we were able to generate a highly prioritized list of candidate operons. Two of the candidate operons *Tektin-C/CG10542* and *Cdk5/flotillin* from this list were tested and both were found to behave like operons [5]. The molecular methods are presented in Ben-Shahar et al. [5]. These gene pairs are conserved across all species of *Drosophila* and also appeared to be involved in similar pathways [5]. These two genes are highlighted with yellow color in Additional Table 1. This set of gene pairs exhibiting operon-like characteristics suggest that there might be more operons in eukaryotes than previously believed, and provides a list of candidates for experimental verification in a molecular biology laboratory. The relationship between *Cheb42a* and *llz* would not have been realized without recognizing their underlying operon structure.

4.1. Data challenges

One of the greatest challenges in this project was the incomplete availability of data. For gene ontology terms, this is a common occurrence. Many genes are incompletely annotated for function and process annotation. Similarly, some genes have much better EST coverage than others. This impacts the ability to detect bona fide untemplated leading Gs. The leading G analysis is also confounded with the presence of a G in the genome. For microarray data, the challenge is broader; a gene may not have a probeset on the given array, or it may have several probesets. As noted in the Methods section, our strategy has been to utilize the probeset with the greatest signal. Although one would expect the frequency histogram all pairwise correlations (i.e., the dotted line in Fig. 2) to be a normal curve centered at zero, it is not the case, perhaps because the hybridizations ($n=99$) used in the calculation were from disparate experiments involving different laboratories, tissues, and conditions. Importantly, data from each experiment was normalized separately. One final component to data integrity is to assess the genomic structure of the operons. Of particular interest are those identified candidates that share significant homology. For such elements, it is often difficult to distinguish between segmental duplications and bona fide operons.

4.2. Prioritization methods

The prioritization methods specified in the Systems and Methods section are by no means the only ways to prioritize candidate operons. We can use as many resources as we might want depending on their availability, reliability and relevance to the particular species in question. Other potential prioritization methods include preservation of the gene order across species. If one or more adjacent genes are tightly coupled in the same order across multiple species, this might suggest that their gene order is necessary for a crucial pathway, one of the primary characteristics of an operon. One can also use publicly available protein–protein interaction databases as a measure of functional similarity. Interaction of proteins from adjacent genes may indicate a functional rationale for their proximity. These criteria, if used along with the identification methods, could prove to be a formidable way to prioritize candidate operons. One resource that has been largely ignored is the literature correlation. Functional annotation of genes in candidate operons can be obtained by scouring through the titles and abstracts of the vast amount of medical literature electronically available on the Internet. This information can be used as a way to prioritize genes depending on the denseness of the resource. One should be careful when using literature correlation as a prioritization process as the amount of information that can be gleaned in this way depends on the species in question as we would expect that human and selected model organisms to be more studied than other species resulting in a potential bias.

The prioritization methods used in the Methods section are not exhaustive by themselves. The choice of the prioritization methods and their precedence depends on the species in question, availability of resources and knowledge of the process. This is a decision that rests solely in the hands of the investigator who wishes to use this procedure for identification of operons. For example, an investigator may choose the distance between genes in a operon in base pairs as having highest priority instead of other methods based on his domain knowledge.

4.3. Multi-gene operons

Although the procedure described in this manuscript to identify candidate operons only identifies candidate pairs, these may be trivially extended to longer operon sets. For example, if genes A and B are in one operon set, and B and C are in another, then through associativity A, B and C make up a larger, three gene operon. The lack of regulatory requirements on the first gene (i.e. no promoter check) allows the upstream gene of one candidate operon to be the downstream gene of another. This “chaining” of operons was observed several times in our data set. Our search for chain operons yielded 26 three-gene, three four-gene and one five-gene operon chains. An attempt was made to identify motifs in the intercistronic region and its neighborhood that might have a major role in the transcription mechanism. Various lengths of the intercistronic region and its surroundings were obtained from the candidate operons and the motif identification algorithm MEME [13] was used to identify conserved motifs in the candidate operons. No motifs were found to be significantly enriched in the set of candidate operons suggesting regulation of operons in *Drosophila* at a different level.

4.4. Future study

Although no transcription of the kind described herein has been reported for either mouse or human, this type of analysis needs to be extended to these and other genomes. Human and mouse genomes may not have traces of the unconventional kind of operons as found in *Drosophila melanogaster* as it appears that this mechanism is unique to the *Drosophila* genome as was a similar trans-splicing mechanism unique to the *C.elegans* genome. Other higher organisms may have evolved lineage-specific mechanism for the transcription and processing of functionally related genes, or possibly there might be no such process at all. However, if a similar unconventional mechanism as that in *Drosophila melanogaster* does exists in humans, it could provide valuable insight into the evolution of gene regulation.

5. Conclusions

Organization of genes in operons in *C. elegans* and *D. melanogaster* not only substantiate the presence of operons in eukaryotes but also reinforces the evolutionary diversity of eukaryotic mechanisms (i.e., trans-splicing) relative to that reported for operon transcription in prokaryotes. An understanding of the regulation of the operon architecture in *Drosophila melanogaster* will bolster efforts to accurately identify all the operons in the *D. melanogaster* genome as current evidence suggests that some gene pairs are organized as operons, but the underlying processes are not completely understood. The identification phase in our computational procedure provides a prototype for identification of operon-like candidates in eukaryotes and the prioritization phase prioritizes candidate operons based on the level of the knowledge of the regulation and available resources for that particular species. We have successfully demonstrated the application of the procedure and identified potential operons, two of which have been shown to be transcribed and processed as eukaryotic operons. In sum, the operon-like mechanism might be just one of the myriads of mechanisms involving gene regulation in eukaryotes.

Conflict of interest statement

None declared.

Acknowledgments

This research was partially supported through a Career Development Award from Research to Prevent Blindness to TES.

YBS was an Associate and MJW is an Investigator of the HHMI.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.compbiomed.2013.03.004.

References

[1] T. Blumenthal Trans-splicing and Operons,in: The C.elegans Research Community (Ed.), WormBook .doi/10.1895/wormbook 2005, 1(1).

[2] J. Spieth, G. Brooke, S. Kuersten, K. Lea, T. Blumenthal, Operons in C. elegans: polycistronic mRNA precursors are processed by trans-splicing of SL2 to downstream coding regions, Cell 73 (3) (1993) 521–532.

[3] T. Blumenthal, D. Evans, C.D. Link, A. Guffanti, D. Lawson, J. Thierry-Mieg, D. Thierry-Mieg, W.L. Chiu, K. Duke, M. Kiraly, A global analysis of Caenorhabditis elegans operons, Nature 417 (2002) 851–854.

[4] T. Blumenthal, Operons in eukaryotes, Brief. Funct. Genomic. Proteomic. 3 (3) (2004) 199–211.

[5] Y. Ben-Shahar, K. Nannapaneni, T.L. Casavant, T.E. Scheetz, M.J. Welsh, Eukaryotic operon-like transcription of functionally related genes in Drosophila, Proc. Natl. Acad. Sci. U. S. A. 104 (1) (2007) 222–227.

[6] H. Lin, K.J. Mann, E Starostina, R.D. Kinser, C.W. Pikielny, A. Drosophila DEG/ENaC channel subunit is required for male response to female pheromones, Proc. Natl. Acad. Sci. U. S. A. 102 (36) (2005) 12831–12836.

[7] M.B. Eisen, P.T. Spellman, P.O. Brown, D. Botstein, Cluster Analysis and Display of Genome-Wide Expression Patterns, PNAS. 95 (25) (1998) 14863–14868.

[8] M.G. Reese, Application of a time-delay neural network to promoter annotation in the Drosophila melanogaster genome, Comput. Chem. 26 (1) (2001) 51–56.

[9] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin, G. Sherlock, Gene ontology: tool for the unification of biology. The Gene Ontology Consortium, Nat. Genet. 25 (1) (2000) 25–29.

[10] T.J. Hubbard, B.L. Aken, K. Beal, B. Ballester, M. Caccamo, Y. Chen, L. Clarke, G. Coates, F. Cunningham, T. Cutts, T. Down, S.C. Dyer, S. Fitzgerald, J. Fernandez-Banet, S Graf, S. Haider, M. Hammond, J. Herrero, R. Holland, K. Howe, K. Howe, N. Johnson, A. Kahari, D. Keefe, F. Kokocinski, E. Kulesha, D. Lawson, I. Longden, C. Melsopp, K. Megy, P. Meidl, B. Ouverdin, A. Parker, A. Prlic, S. Rice, D. Rios, M. Schuster, I. Sealy, J. Severin, G. Slater, D. Smedley, G. Spudich, S. Trevanion, A. Vilella, J. Vogel, S. White, M. Wood, T. Cox, V. Curwen, R. Durbin, X.M. Fernandez-Suarez, P. Flicek, A. Kasprzyk, G. Proctor, S. Searle, J. Smith, A. Ureta-Vidal, E. Birney, Ensembl 2007, Nucleic Acids Res. 35 (2007) D610–7, Database issue.

[11] T. Barrett, D.B. Troup, S.E. Wilhite, P. Ledoux, D. Rudnev, C. Evangelista, I.F. Kim, A. Soboleva, M. Tomashevsky, R. Edgar, GEO NCBI, mining tens of millions of expression profiles – database and tools update, Nucleic Acids Res. 35 (2007) D760–5, Database issue.

[12] M.S. Boguski, T.M.J. Lowe, C.M. Tolstoshev, dbEST – database forexpressed sequence tags, Nat. Genet. 4 (4) (1993) 332–333.

[13] T.L. Bailey, N. Williams, C. Misleh, W.W. Li, MEME: discovering and analyzing DNA and protein sequence motifs, Nucleic Acids Res. 34 (2006) W369–W373, Web Server issue.

[14] M. Land, A. Islas-Trejo, C.S. Rubin, Origin, properties, and regulated expression of multiple mRNAs encoded by the protein kinase C1 gene of Caenorhabditis elegans, J. Biol. Chem. 269 (20) (1994) 14820–14827.

[15] C. Quijano, P. Tomancak, J. Lopez-Marti, M. Suyama, P. Bork, M. Milan, D. Torrents, M. Manzanares, Selective maintenance of Drosophila tandemly arranged duplicated genes during evolution, Genome Biol. 9 (12) (2008) R176.