Amino Acids (2013) 44:597–606
DOI 10.1007/s00726-012-1380-1

ORIGINAL ARTICLE

Genome-wide computational identification of bicistronic mRNA in humans

Yiming Lu · Yanchun Zhang · Xingyi Hang · Wubin Qu · Gert Lubec · Changsheng Chen · Chenggang Zhang

Received: 23 June 2012 / Accepted: 26 July 2012 / Published online: 4 September 2012

Abstract Mammalian bicistronic mRNA is a recently discovered mammalian gene structure. Several reported cases of mammalian bicistronic mRNA indicated that genes of this structure play roles in some important biological processes. However, a genome-wide computational identification of bicistronic mRNA in mammalian genome, such as human genome, is still lacking. Here we used a comparative genomics approach to identify the frequency of human bicistronic mRNA. We then validated the result by using a new support vector machine (SVM) model. We identified 43 human bicistronic mRNAs in 30 distinct genes. Our literature analysis shows that our method recovered 100 % (6/6) of the previously known bicistronic mRNAs which had been experimentally confirmed by other groups. Our graph theory-based analysis and GO analysis indicated that human bicistronic mRNAs are prone to produce different yet closely functionally related proteins. In addition, we also described and analyzed three different mechanisms of ORF fusion. Our method of identifying bicistronic mRNAs in human genome provides a model for the computational identification of characteristic gene structures in mammalian genomes. We anticipate that our data will facilitate further molecular characterization and functional study of human bicistronic mRNA.

Keywords Bicistronic mRNA · Computational identification · Domain–domain interaction · Open reading frame fusion · Support vector machine

Introduction

It is widely accepted that polycistronic mRNA is a characteristic of prokaryotes where operons are a common form of gene organization (Lawrence 2002). By contrast, the genes of eukaryotes are organized more individually, and their transcripts are generally considered to be monocistronic; however, it recently became clear that not all eukaryotic genes are transcribed monocistronically. Numerous instances of polycistronic transcription in

Electronic supplementary material The online version of this article (doi:10.1007/s00726-012-1380-1) contains supplementary material, which is available to authorized users.

Y. Lu · Y. Zhang · X. Hang · W. Qu · C. Zhang (✉)
Beijing Institute of Radiation Medicine, State Key Laboratory of Proteomics, Cognitive and Mental Health Research Center,
Beijing 100850, China
e-mail: zcgweb@yahoo.com; zhangcg@bmi.ac.cn

Y. Lu
e-mail: luym.science@gmail.com

Y. Zhang
e-mail: zhangyanchun17@163.com

X. Hang
e-mail: xingyi.hang@gmail.com

W. Qu
e-mail: quwubin@gmail.com

G. Lubec
Department of Pediatrics, Medical University of Vienna,
Währinger Gürtel 18, 1090 Vienna, Austria
e-mail: lubec@gmx.at

C. Chen (✉)
Department of Health Statistics, School of Military Preventive Medicine, Fourth Military Medical University,
17 Changle Xilu, Xi'an 710032, China
e-mail: chencs@fmmu.edu.cn

C. Zhang
School of Life Science, Anhui Medical University,
Anhui 230032, China

eukaryotes have been reported [reviewed in (Blumenthal 2004)]. The first case of polycistronic transcription in eukaryotes was found in 1988 in trypanosomes. Widespread operons in an animal were first discovered in the nematode *Caenorhabditis elegans* (Spieth et al. 1993). More recently, the operons were also discovered in flatworms and primitive chordates (Davis and Hodgson 1997; Ganot et al. 2004).

The operon was first described by Jacob et al. (1960) and was defined as a cluster of genes that are under the control of a single regulatory signal or promoter. We now know that bacterial operons are generally transcribed from a single promoter, and that they result in the formation of a polycistronic mRNA that is translated by ribosomes that re-initiate translation at the 5′ ends of downstream genes, having terminated translation at the 3′ ends of upstream genes [reviewed in (Blumenthal and Gleason 2003)]. Unlike bacterial operons, nematode operons are transcribed to produce polycistronic initial transcripts, and then these transcripts are processed by *trans*-splicing procedure to create monocistronic mRNAs that are transported to the cytoplasm and translated (Spieth et al. 1993). Another type of polycistronic (bicistronic) mRNA was first found in mouse nervous system (Lee 1991). Soon after that, several genes were found to be able to produce bicistronic mRNAs in mammalian genomes (Reiss et al. 1998; Stallmeyer et al. 1999; Gray et al. 1999; Hayward et al. 2003; Autio et al. 2008). Distinct from the polycistronic pre-mRNAs in *C. elegans*, the bicistronic transcripts in mammals were mature mRNAs, from which two distinct proteins could be directly translated. Intuitively, they are very similar to the bacterial operons. A small number of studies have also shown that proteins encoded by mammalian bicistronic mRNAs are functionally related. For instance, the bicistronic mRNA of human gene *MOCS2* encodes the small and large subunits of the molybdopterin synthase (Stallmeyer et al. 1999), and a human bicistronic mRNA encodes *MFRP* and *C1QTNF5*, which are proteins that co-localize and interact with each other (Mandal et al. 2006). However, a large-scale study of the relationships between bicistronic mRNA-produced proteins is still lacking. In addition, since the bicistronic mRNAs found in mammals were extremely rare, the frequency of them in mammalian genomes also needs to be estimated.

In order to clarify these issues, we utilized a systematic in silico approach to identify the polycistronic mRNAs in mammals. We chose human genomes as our initial data set owing to its ability to provide reliable mRNA sequences, gene structures, and functional information. To make our prediction more reliable, we used a comparative genomics strategy to find the evolutionarily conserved polycistronic mRNAs. By using these methods, we successfully identified a set of conserved human polycistronic mRNAs. To validate the final bicistronic mRNAs set, we used a fivefold cross-validation support vector machine (SVM) learning method, which can precisely distinguish the two classes of open reading frames (ORFs): protein-coding ORF and UTR pseudo ORF. We found that most of the mRNAs' ORFs in the final set were assigned as protein-coding ORFs by the SVM model. We next studied the functional relationships of these polycistronic mRNA-produced proteins using a graph-based method. Interestingly, during these studies, we also found the appearance of another phenomenon in our final set: several mRNAs fused their two ORFs into one and produced single proteins with multiple domains. Literature studies suggested that different mechanisms might take part in these fusion processes.

Results

Identification of functional bicistronic mRNAs in human genome

We developed a systematic, in silico, predictive approach that uses evolutionary information to estimate the frequency of bicistronic mRNAs in the human genome. We used the NCBI RefSeq mRNA database (Pruitt et al. 2007) as the initial mRNA sequence data. All of the ORFs were derived from the entire human mRNAs in three potential frames. To filter the non-functional degraded pseudo ORFs which would contain in-frame stop codons (Harrison et al. 2002), all ORFs were subsequently filtered using a minimum ORF length of 50 codons, which can filter out more than 91.5 % pseudo ORFs and can retain 99.8 % protein-coding ORFs on the basis of previously known coding ORF and pseudo ORF length (see "Methods"). Because the structural and functional domains are the most important characteristics of functional proteins, we examined the functionalities of these ORFs by assessing the existence of protein domains, using publicly available protein domain databases. These public databases enabled us to detect domains with high accuracy and broad coverage (Basu et al. 2009). Six sequence- and/or structure-based protein signature databases in the InterPro suite (Zdobnov and Apweiler 2001) were employed to comprehensively and reliably identify protein domains. The resulting domain information was parsed by Python scripts to make sure it is integrated and non-redundant for each ORF.

In our studies, a bicistronic mRNA candidate was defined as an mRNA with two domain-containing ORFs, which is consistent with the former description of bicistronic mRNA (Blumenthal 2004). According to this definition, we identified 1,813 total human bicistronic mRNA candidates that were distributed across 1,037 genes (one gene may produce more than one bicistronic mRNA isoforms) (Supplementary Table 1). Considering that a portion of these domain-containing ORFs may still not be functional, this number is probably an overestimate of the frequency of bicistronic mRNAs in human genome. To address this issue, we used a comparative genomics approach to further filter the potential non-functional ORFs and make our final set more conservative. We assumed that the probability of randomly generating an ORF with functional domains, and within the reading frame of an mRNA sequence, is very low. Furthermore, we assumed that the detection of ORF pairs with identical domains in the similar positions of two orthologous mRNAs strongly suggests that they are functional bicistronic mRNAs. We used a similar approach on the entire mouse mRNAs as we did on human mRNAs and found 1,871 bicistronic mRNA candidates in 1,389 genes (Supplementary Table 2). We then compared these candidates with the human bicistronic mRNA candidates using an integrated human–mouse orthologous gene data. According to several conservation criteria (see "Methods"), we identified 45 human bicistronic mRNAs from 31 distinct genes that are highly conserved between human and mouse (Supplementary Tables 3, 4). Their conserved protein domains as well as gene structures suggest that they are subject to natural selection, and therefore are very likely to be functional bicistronic mRNAs.

Confirmation of human bicistronic mRNAs set using an SVM model and literature analysis

The conservation criteria and thresholds employed in our method may introduce some bias in the result. Therefore, to confirm the result and to eliminate the possible bias, we adopted a new parameter-threshold-free machine learning method by using an SVM classifier model. Since this model is completely independent of the former identification method, it can be used to further verify the result. The SVM was designed to classify two different kinds of ORFs: protein-coding ORFs and UTR pseudo ORFs. We used two different classes of ORFs to train the SVM model: the protein-coding ORFs and 3′-UTR pseudo ORFs. To robustly classify the two kinds of ORFs, we selected three mutually independent features of ORFs: ORF length, domain number, and codon composition. These features were selected because their value distributions in two ORFs classes are significantly different, ensuring that the model can achieve high accuracy. Using these features, we constructed a 63-dimension SVM classifier model. In these 63 dimensions, 61 dimensions are contributed by 61 codon compositions (removing three stop codons), one dimension is contributed by domain number, and another dimension is contributed by ORF length, so this machine learning method is very independent of the former method. A fivefold cross-validation procedure was repeated 50 times to examine the performance of this SVM classifier (see "Methods"). After that, we found that the overall sensitivity and specificity of this model are 99.52 and 99.29 %, and the average accuracy is up to 99.13 %. The performance of this SVM classifier demonstrates that it has been well trained and can classify protein-coding ORFs and UTR pseudo ORFs very precisely.

We next applied this SVM model to classify the 25 ORFs in human bicistronic mRNAs which have not been confirmed as encoding proteins. In this procedure, the SVM model will assign each ORF a class label and give the corresponding probability based on the training result in the previous step. The result shows that most of the ORFs were assigned as protein-coding ORFs with very high probabilities (Table 1). For example, 17 ORFs were assigned as protein-coding ORFs with probabilities larger than 0.9, among which 10 ORFs were assigned with probabilities larger than 0.99. Despite no ORF being assigned as pseudo ORF, two ORFs in *PRKCB* and *ZNF808* were assigned as protein-coding ORFs with related low probabilities (0.6215 and 0.6704). By removing these two less reliable ORFs, we finally acquired 43 highly confident human bicistronic mRNAs in 30 genes.

Although we did not validate the protein expression directly in experiments, we found that an analysis of current literature could provide some support for the result. We did a thorough literature analysis by searching for the previously known bicistronic mRNAs that had been experimentally confirmed. We found that so far there are altogether six experimentally confirmed human genes producing bicistronic mRNAs, which are *GDF1* (*LASS1*) (Lee 1991), *MOCS1* (Reiss et al. 1998), *MOCS2* (Stallmeyer et al. 1999), *SNRPN* (*SNURF*) (Gray et al. 1999), *MFRP* (*C1QTNF5*) (Hayward et al. 2003), and *RPP14* (Autio et al. 2008). Although they were found individually in different experiments, all of them were successfully recovered by our method and included in the final set, which indicates that our method has a good coverage as well as a high accuracy in detecting functionally bicistronic mRNAs.

A typical example of highly conserved human bicistronic mRNAs

One example of the highly conserved human bicistronic mRNA set is *CHTF8* (chromosome transmission fidelity factor 8 homolog), which produced three potential bicistronic mRNA transcripts through alternative splicing. These three transcripts only differ in the 5′ untranslated region (5′ UTR) and/or the first alternative exon. Two of the transcripts encode the same 121 amino acid (aa) protein, whereas the other one encodes an isoform of 102 aa. These proteins form a part of the Ctf18 replication factor C (RFC) complex that plays a role in sister chromatid

Table 1 A summary of 43 identified human bicistronic mRNAs

Gene	mRNA	ORF length (aa)		Probability
		ORF1	ORF2	
CDKN2A	NM_058195	173	105	0.991798
DIO1	NM_000792	125	120	–
DIO2	NM_000793	132	51	–
	NM_013989	132	51	–
DIO3	NM_001362	169	131	–
GNAS	NM_016592	245	363	0.9999878
GPX1	NM_000581	74	145	–
GPX3	NM_002084	72	53	–
GPX4	NM_001039847	72	99	–
	NM_001039848	109	69	–
	NM_002085	72	69	–
GRM1	NM_001114329	906	378	0.999943
MOCS1	NM_001075098	385	249	0.9987236
	NM_005943	385	249	0.9987236
MOCS2	NM_176806	88	188	–
SNRPN	NM_003097	71	240	–
	NM_022805	67	240	0.8258974
	NM_022806	67	240	0.8258974
	NM_022807	67	240	0.8258974
	NM_022808	67	240	0.8258974
SNURF	NM_005678	71	240	–
LDB2	NM_001130834	331	74	0.8619329
LASS1	NM_021267	350	372	–
RPP14	NM_001098783	124	168	–
	NM_007042	124	168	–
PEG10	NM_001040152	325	357	0.9970375
	NM_015068	325	357	0.9970375
PLCB1	NM_182734	1,173	119	0.9525629
LRRC29	NM_001004055	132	223	0.9199278
TBX22	NM_001109879	127	400	0.9744221
DUSP13	NM_001007271	188	327	0.9999813
CHTF8	NM_001039690	121	524	–
	NM_001040144	102	524	–
	NM_001040146	121	524	–
ZFP2	NM_030613	76	461	0.902209
MFRP	NM_031433	579	243	–
C1QTNF5	NM_015645	579	243	–
COG8	NM_032382	612	108	0.9586719
XIRP2	NM_152381	3,549	527	0.9999194
ZNF827	NM_178835	1,077	259	0.9999476
ZNF780A	NM_001010880	641	320	0.8209796
	NM_001142577	642	320	0.8209796
	NM_001142578	641	320	0.8209796

– in 'probability' column means encoding known proteins

cohesion and DNA replication and repair. Interestingly, an additional ORF was identified by our method in each 3′ UTR of these three transcripts, the protein sequences of which are identical. This novel ORF is 1,575 bp long and is predicted to encode a 524 aa protein (Fig. 1). A BLAST search of this 524 aa ORF in the UniProt Consortium

(2010) database produced a hit with a 524 aa protein named 'chromosome transmission fidelity protein 8 homolog isoform 2′ (UniProt ID: P0CG12) with 100 % identity. Sequence analysis shows that this protein has a high sequence similarity (83 %) with the known 533 aa protein encoded by the rat *Chtf8* gene. A TBLASTN search of the reference genomic sequences revealed that this ORF is also highly conserved in several other mammals, but not in birds or reptiles. More interestingly, we also found that a four-nucleotide overlap **AUG**A (the start codon is in bold and the stop codon is underlined) forms a stop-start codon between these two consecutive ORFs, which joins the downstream ORF to the upstream ORF in the −1 frame phase. The overlapping stop-start codon is essential for bicistronic mRNA translational coupling in non-long terminal repeat (non-LTR) retrotransposons, which are widely distributed in mammals (Kojima et al. 2005). In addition, this protein contains a domain that belongs to the SWI/SNF-related, matrix-associated, actin-dependent regulator of chromatin-related gene family. The domain present in this protein suggests that both proteins may play roles in the process of chromatin structure modification. Further experimental studies are needed to confirm protein expression and to fully characterize the function of this novel protein (Fig. 1).

Functional relationship between bicistronic mRNA-produced proteins

Although a small number of studies have shown that mammalian bicistronic mRNA-produced proteins are functionally related, in most cases the functions of both proteins are unknown, which makes their functional relationships difficult to determine. Since protein domains are the functional units of proteins, the interactions between proteins are often reflected by the interactions between their domains (Pandey et al. 2008). Therefore, we further examined the domain interactions of the bicistronic-encoded proteins by constructing a domain–domain interaction (DDI) network using a publicly available DDI database named DOMINE (Raghavachari et al. 2008), which is a comprehensive database of protein domain interactions collating known and predicted DDI from 10 different sources. In our DDI network, each node represents a kind of protein domain, each edge that links two nodes represents an "interaction" between two kinds of domains, and the distance between two nodes is defined as the number of edges in the shortest path between them. The closeness of the relationship between two kinds of domains can be measured by the graph distance of the two corresponding nodes which can be calculated by a graph-based algorithm (see "Methods").

We divided the identified human bicistronic mRNA into two sets: the bicistronic mRNA candidates set and the high-quality bicistronic mRNA set (43 highly conserved bicistronic mRNAs), and domain pairs were derived from the two ORFs of these mRNAs. We deleted the identical domains presenting in the same mRNA to avoid the potential impact of ORFs generated due to tandem duplication. The control data also contained two sets: 1,000,000 randomly sampled domain pairs from the 3,915 nodes in the DDI network and 3,022 genomically adjacent (<20 kb), non-bicistronic domain pairs from 5,000 randomly selected gene pairs. We then calculated the graph distances of these domain pairs in different sets. We find that the distances in two bicistronic mRNA sets (highly conserved set and

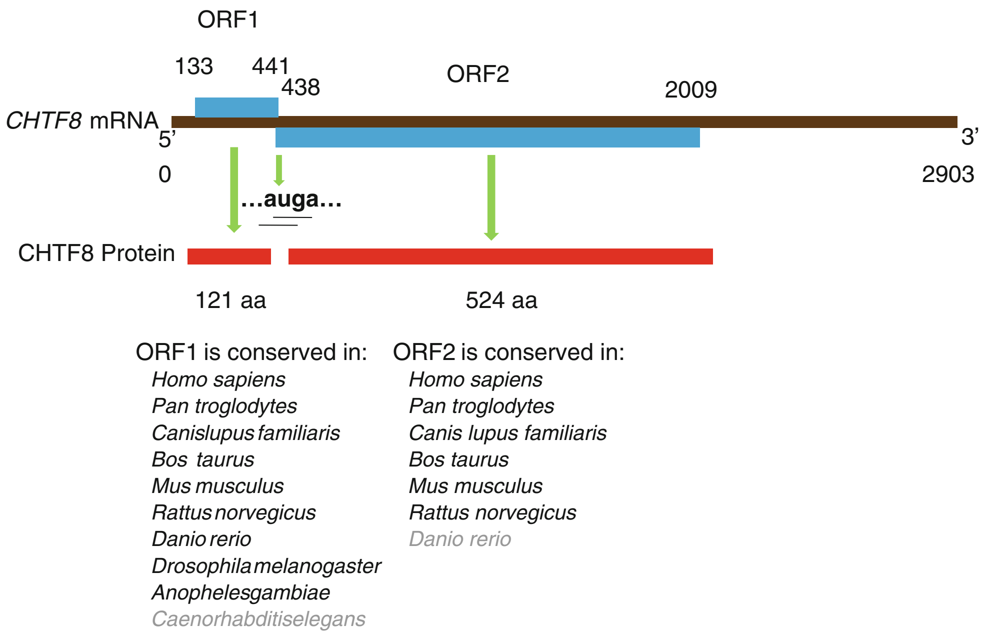

Fig. 1 Bicistronic structure of one human CHTF8 gene transcript. The two consecutive ORFs encode two proteins of 121 aa and 524 aa, respectively. The previously known 121 aa protein is highly conserved in mammals, birds, reptiles, invertebrates, and drosophila; the predicted 524 aa protein is highly conserved in mammals, but not in birds or reptiles

candidates set) are significantly different from those of the random domain pair set. The average distances of the random domain pair set and genomically adjacent, non-bicistronic set are 4.0158 and 2.443, whereas those of the high-quality set and candidates set are 1.2824 and 1.4348 (Fig. 2). This result suggests that the functional relationship between the proteins produced by a bicistronic mRNA is much closer than what is expected between two random proteins. In addition, the average distances of the bicistronic mRNA sets are close to 1, and more than 70 % of domain pairs of the high-quality set have graph distances equal to 1 (more than 60 % in bicistronic candidates set), indicating that the relationship between proteins produced by bicistronic mRNA is prone to be a direct interaction, for only those domain pairs having direct interactions in the DDI network have graph distances equal to 1. Interestingly, the average distance of genomically adjacent, non-bicistronic ORF pairs was also significantly smaller than the completely random ORF pairs, although still much larger than that of bicistronic ORF pairs. This phenomenon may be related to the existence of gene clusters in human genome, for in gene clusters functionally related genes are prone to be arranged in adjacent locations in the chromosome (Hurst et al. 2004).

To further confirm this result, we used gene ontology (GO) (Ashburner et al. 2000) to analyze those protein domains' functional categories. GO analysis of these domains shows that 73.6 % of the bicistronic mRNAs contain domain pairs that belong to the same GO terms, and this observation is consistent with our former result in the DDI network. Consequently, we can conclude that human bicistronic mRNAs are prone to produce different yet closely functionally related proteins. This is understandable from an evolutionary point of view: although having two ORFs in one mRNA is a costly arrangement under evolutionary pressure, this arrangement may be advantageous for the co-expression of functionally related/interacting proteins in mammals.

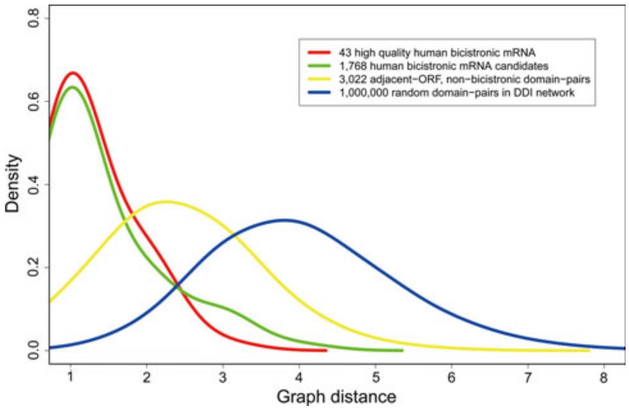

Fig. 2 Graph distance distribution of four groups of domain pairs

ORF fusion events in human bicistronic mRNA set

Thorough investigation of the highly conserved human bicistronic mRNA set shows that 13 of the 43 bicistronic mRNAs (present in 8 genes) fused both their ORFs into a larger ORF (Table 2) (Mandel et al. 1992; Berry et al. 1991; Salvatore et al. 1995, 1996; Mullenbach et al. 1987; Schuckelt et al. 1991; Chu et al. 1992; Gray and Nicholls 2000; Shigemoto et al. 2001). In other words, instead of producing two functionally related proteins each with a start and stop codon, they produce relatively long and multi-domain proteins by overcoming the first ORF's stop codon. Sequence and literature analyses show that three distinct mechanisms are involved in these ORF fusion events: (1) selenocysteine suppression of the UGA stop codon, where six bicistronic mRNAs (DIO1, DIO2, DIO3, GPX1, GPX3, and GPX4) fused their two consecutive ORFs by selenocysteine suppression of the UGA stop codon, which requires sec insertion sequences (SECISs) (Mandel et al. 1992; Berry et al. 1991; Salvatore et al. 1995, 1996; Mullenbach et al. 1987; Schuckelt et al. 1991; Chu et al. 1992); (2) avoidance of a stop codon by alternative splicing where only *MOCS1* fused its two ORFs by employing alternative splicing to skip the non-coding region that contains the stop codon (Gray and Nicholls 2000); and (3) stop codon read-through by a −1 ribosomal frameshift where *PEG10* fused its two ORFs by a −1 ribosomal frameshift mechanism (Fig. 3) (Shigemoto et al. 2001).

We then wanted to see if these ORF junctions including the stop codons were conserved in other species. Therefore, we did multiple sequence alignments (MSA) on the homologous sequences of these bicistronic mRNAs in other organisms, including mouse, rat, cow, zebrafish, fruitfly, nematode, yeast, etc. We found that nearly all the ORF junctions including the stop codons are preserved quite well in mammals, chicken, and zebrafish, but most of them are not conserved in fruitfly, nematode, and yeast (Supplementary Table 5). Part of the result shows that the stop codons were created as a result of mutations (in the third position of the codon) (Supplementary Fig. 2). For example, the 'TGT' of *GPX3* in yeast was mutated to 'TGA' in mammals, the 'TGC' in *GPX4* in fruitfly was mutated to 'TGA' in mammals. In addition, we found that *MOCS1* is the only gene which is also conserved in fruitfly and nematode.

Three conditions must be satisfied before an ORF fusion can occur. The first condition is that the two ORFs cannot overlap. *MOCS2* does not fulfill this criterion because its two ORFs slightly overlap (Stallmeyer et al. 1999). The second condition is that the original functions of two donor proteins must be maintained after ORF fusion. For example, two mRNA transcripts of *MOCS1* fuse their ORFs

Table 2 List of the 13 human ORF-fused bicistronic mRNAs and their corresponding mechanisms

ORF fusion mechanisms	GeneID	Gene	mRNA
Selenocysteine suppression of UGA stop codon	1733	*DIO1*	NM_213593
	1734	*DIO2*	NM_000793
			NM_013989
	1735	*DIO3*	NM_001362
	2876	*GPX1*	NM_000581
	2878	*GPX3*	NM_002084
	2879	*GPX4*	NM_001039847
			NM_001039848
			NM_002085
Avoidance of stop codon by alternative splicing	4337	*MOCS1*	NM_001075098
			NM_005943
Stop codon reading through by −1 ribosomal frameshift	23089	*PEG10*	NM_001040152
			NM_015068

whereas the other two transcripts express the individual ORFs, for the well-conserved C-terminal motif: isoleucine-glycine glycine-stop (IGG*) in different species (Reiss et al. 1998; Wilson et al. 1994) suggests its necessity for the biological functions of the upstream protein. The third condition is the frame phase restriction: ORF fusion by selenocysteine suppression of UGA stop codon requires the two ORFs to be in-frame, and the ribosomal −1 frameshift mechanism fusion requires a −1 frameshift between two ORFs.

Discussion

A small fraction (13/43) of the human bicistronic mRNA set fused two ORFs into a relatively long and multi-domain ORF. This phenomenon suggests that mammalian bicistronic mRNAs are not evolutionarily stable and that there are evolutionary forces that select for the creation of a new multi-domain ORF from two ORFs in one transcript (Enright et al. 1999). The evolutionary advantages of protein fusion have been described in several articles including spatial and temporal co-regulation of related gene expression, protein multi-functionalization and the creation of new genes (Long 2000; Long et al. 2003). ORF fusion is a key step in the evolution of novel fusion proteins from bicistronic transcripts. Two donor genes should be close enough genomically (which could be achieved by gene duplication, gene translocation or de novo origination of a new gene) so that they can be transcribed into one transcript, followed by the removal of the non-coding region between the two ORFs or reading-through this region, so that the ORFs can successfully fuse. Our analysis shows that three distinct mechanisms can lead to ORF fusion and thereafter the creation of multi-domain proteins.

Although protein evolution is a dynamic process, in most cases we can only observe the final product. Many details of this evolutionary process, such as how proteins with new functions originate and evolve, are unknown. Fortunately, the traces that remain in several important genes can still be observed and studied, such as the mammalian bicistronic mRNA structure. As an intermediate evolutionary state of protein fusion, the bicistronic mRNAs may provide a rare opportunity to observe how two individual proteins became one multi-domain protein. Such observations may allow us to understand the molecular mechanisms underlying the creation of novel multi-domain proteins in mammals. Clearly, additional experiments are necessary to study the function of these novel proteins, which will enhance our knowledge of the evolutionary forces that underlie protein evolution. In the future, it will be helpful to combine computational genomic analyses and biochemical characterization of the novel fusion proteins to better understand the molecular mechanisms and patterns of protein evolution.

Conclusions

In this report, a comparative genomics approach was applied to identify the frequency of human bicistronic mRNA. By validating the result using an independent SVM classifier model and performing thorough literature analyses, we finally identified 43 human bicistronic mRNAs in 30 distinct genes. Our graph theory-based analysis and GO analysis indicated that human bicistronic mRNAs are prone to produce different yet closely functionally related proteins. In addition, we also described and analyzed three different mechanisms of ORF fusion.

Methods

Computational identification of human bicistronic mRNAs

We developed a systematic, in silico, predictive approach to reliably identify the bicistronic mRNAs in the human genome. We used the human mRNA reference sequence data (Pruitt et al. 2007) as our initial mRNA sequence data set. Approximately 770,000 ORFs, in three possible reading frames, were derived from the whole human mRNA sequences using the 'sixpack' application in the EMBOSS suite (Rice et al. 2000). To determine a threshold length within which to identify truly functional ORFs, we investigated the length distributions of our ORF set and the

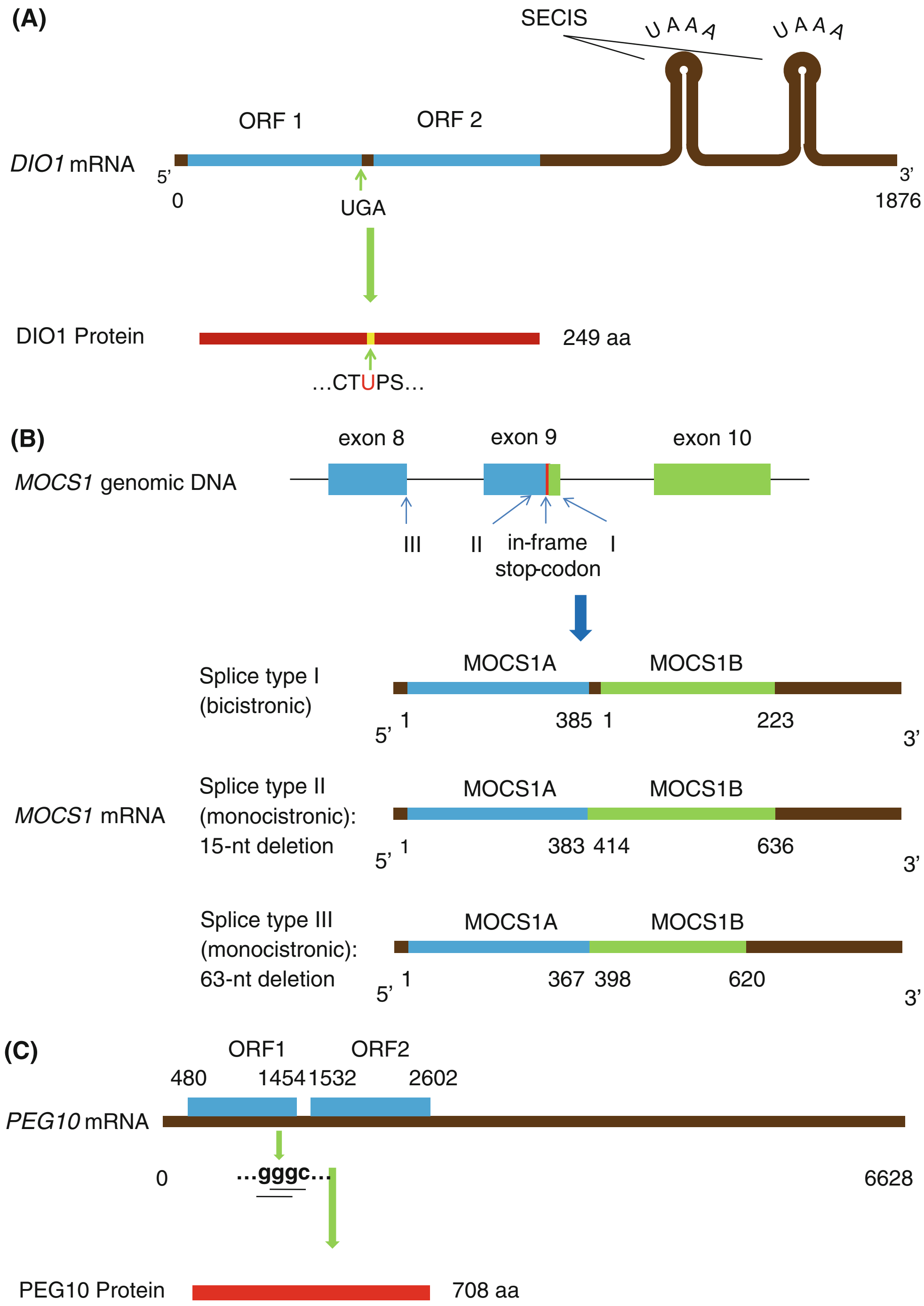

Fig. 3 Diagrams of three possible mechanisms of ORF fusion events. **a** Selenocysteine suppression of the UGA stop codon where the bicistronic transcript of the *DIO1* gene fused its two consecutive ORFs by selenocysteine suppression of the UGA stop codon, which requires the existence of sec insertion sequences (SECISs); **b** avoidance of the stop codon by alternative splicing, where the *MOCS1* protein fused its two ORFs by employing alternative splicing to skip a non-coding region that contains the stop codon; and **c** stop codon read-through by a −1 ribosomal frameshift where the PEG10 protein fused its two ORFs by a −1 ribosomal frameshift mechanism

currently known functional ORFs in humans. We found that a threshold of 50 codons could provide a good coverage of functional ORFs and efficiently filter the randomly generated ORFs. Using this threshold, we selected about 200,000 functional ORF candidates. To comprehensively and reliably identify protein domains in these functional ORF candidates, we employed a locally installed sequence- and/or structure-based protein signature database named InterPro (Zdobnov and Apweiler 2001), which provides an integrated layer on top of the most commonly used signature databases by creating a unique, non-redundant characterization of a given protein family, domain, or functional site. Using our previous testing data, we found six databases (hmmpfam, hmmsmart, hmmpanther, superfamily, profilescan, and fprintscan) that could accurately identify the majority of the previously annotated protein domains. We chose to use these six InterPro databases to examine the significant protein domains of the functional ORF candidates. Each database gives *E* values to represent the significance of the domains; we chose an *E* value threshold of 0.01 to remove the insignificant protein domains. The InterPro suite also enabled us to look up the corresponding GO terms of a known domain using the non-redundant InterPro terms. Additionally, each ORF sequence was used as a query to search the human protein reference sequence data. The BLASTP program was used to examine whether an ORF encodes a previously known protein. Only when an ORF and a protein were equal in length and 100 % matched did we consider that this ORF encoded a previously known protein. We subsequently integrated the domain-detecting results, the BLASTP results, and the gene information using Python scripts. The human and mouse bicistronic mRNA candidates were identified according to the integrated information and our definition of bicistronic mRNA candidates. We then compared these human and mouse bicistronic mRNA candidates according to the human–mouse orthologous gene data integrated from the NCBI Homologene database (http://www.ncbi.nlm.nih.gov/homologene/) and the MGI Mammalian Orthology database (http://www.informatics.jax.org) (Bult et al. 2008). We only selected the orthologous bicistronic mRNA pairs that satisfied the following three conditions: (1) each mRNA in the orthologous mRNA pair contains two ORFs (called an ORF pair); (2) in one certain orthologous mRNA pair, each ORF of an ORF pair contains the same domain(s) with the corresponding ORF in another pair; (3) ORFs containing the same domains should be in the similar position of the corresponding mRNA. The entire workflow can be seen in Supplementary Fig. 1.

Application of SVM model

We used the ‘e1071’ package (Chang and Lin 2001; Meyer 2006) in R (Ihaka and Gentleman 1996) for SVM modeling. The training data set contains two classes: the protein-coding ORFs and 3′-UTR pseudo ORFs, both of which were derived from RefSeq mRNAs. Altogether, we derived 21,878 protein-coding ORFs and 444,597 3′-UTR pseudo ORFs. In order to balance the two classes of the training data, we randomly sampled 15,000 ORFs in each class. To robustly classify the two kinds of ORFs, three mutually independent features of ORFs, namely ORF length, domain number, and codon composition, were selected as the training features. Using these features, we constructed a 63-dimension SVM classifier. In these 63 dimensions, 61 dimensions are contributed by 61 codon compositions (removing three stop codons), one dimension is contributed by domain number, and another dimension is contributed by ORF length. A fivefold cross-validation procedure was repeated 50 times to robustly estimate the performance of this SVM classifier.

Calculation of the domain–domain interaction network distances based on graph theory

We constructed a DDI network using a compiled domain interaction database which integrated 10 protein domain interaction databases. The resulting network is a 3,915-node graph, in which each node represents a kind of protein domain, each edge that links two nodes represents an “interaction” between two kinds of domains, and the distance between two nodes is defined as the number of edges in the shortest path between them. We then derived domain pairs from the ORF pairs of mRNAs in human bicistronic mRNA candidates set and high-quality bicistronic mRNA set, and we deleted the identical domains presenting in the same mRNA to avoid the potential impact of ORFs generated as a result of tandem duplication. To generate the random domain pair set, 1,000,000 domain pairs were randomly sampled from the entire 9,204,891 domain pairs that can be linked by paths from the graph. To generate the genomically adjacent, non-bicistronic ORF pairs, we randomly selected 5,000 adjacent gene pairs in human chromosomes within a window of 20 kb. The Floyd–Warshall algorithm (Floyd 1962) was employed to find the shortest path between each pair of domains and calculate the corresponding distances.

Acknowledgments This work was supported by the National Basic Research Project (973 program) (2012CB518200), the General Program (30900830, 81172770) of the Natural Science Foundation of China, the State Key Laboratory of Proteomics of China (SKLP-O201104, SKLP-K201004, SKLP-O201002), and the Special Key Programs for Science and Technology of China (2012ZX09102301-016).

References

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. The gene ontology consortium. Nat Genet 25:25–29

Autio KJ, Kastaniotis AJ, Pospiech H, Miinalainen IJ, Schonauer MS, Dieckmann CL, Hiltunen JK (2008) An ancient genetic link between vertebrate mitochondrial fatty acid synthesis and RNA processing. FASEB J 22:569–578

Basu MK, Poliakov E, Rogozin IB (2009) Domain mobility in proteins: functional and evolutionary implications. Brief Bioinform 10:205–216

Berry MJ, Banu L, Chen YY, Mandel SJ, Kieffer JD, Harney JW, Larsen PR (1991) Recognition of UGA as a selenocysteine codon in type I deiodinase requires sequences in the 3’ untranslated region. Nature 353:273–276

Blumenthal T (2004) Operons in eukaryotes. Brief Funct Genomic Proteomic 3:199–211

Blumenthal T, Gleason KS (2003) Caenorhabditis elegans operons: form and function. Nat Rev Genet 4:112–120

Bult CJ, Eppig JT, Kadin JA, Richardson JE, Blake JA (2008) The mouse genome database (MGD): mouse biology and model systems. Nucleic Acids Res 36:D724–D728

Chang C, Lin C (2011) LIBSVM: a library for support vector machines. ACM Trans Intell Syst Technol 2:1–27

Chu FF, Esworthy RS, Doroshow JH, Doan K, Liu XF (1992) Expression of plasma glutathione peroxidase in human liver in addition to kidney, heart, lung, and breast in humans and rodents. Blood 79:3233–3238

Davis RE, Hodgson S (1997) Gene linkage and steady state RNAs suggest trans-splicing may be associated with a polycistronic transcript in Schistosoma mansoni. Mol Biochem Parasitol 89:25–39

Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA (1999) Protein interaction maps for complete genomes based on gene fusion events. Nature 402:86–90

Floyd RW (1962) Algorithm 97: shortest path. Commun ACM 5:345

Ganot P, Kallesoe T, Reinhardt R, Chourrout D, Thompson EM (2004) Spliced-leader RNA trans splicing in a chordate, *Oikopleura dioica*, with a compact genome. Mol Cell Biol 24:7795–7805

Gray TA, Nicholls RD (2000) Diverse splicing mechanisms fuse the evolutionarily conserved bicistronic MOCS1A and MOCS1B open reading frames. RNA 6:928–936

Gray TA, Saitoh S, Nicholls RD (1999) An imprinted, mammalian bicistronic transcript encodes two independent proteins. Proc Natl Acad Sci U S A 96:5616–5621

Harrison PM, Hegyi H, Balasubramanian S, Luscombe NM, Bertone P, Echols N, Johnson T, Gerstein M (2002) Molecular fossils in the human genome: identification and analysis of the pseudogenes in chromosomes 21 and 22. Genome Res 12:272–280

Hayward C, Shu X, Cideciyan AV, Lennon A, Barran P, Zareparsi S, Sawyer L, Hendry G, Dhillon B, Milam AH, Luthert PJ, Swaroop A, Hastie ND, Jacobson SG, Wright AF (2003) Mutation in a short-chain collagen gene, CTRP5, results in extracellular deposit formation in late-onset retinal degeneration: a genetic model for age-related macular degeneration. Hum Mol Genet 12:2657–2667

Hurst LD, Pal C, Lercher MJ (2004) The evolutionary dynamics of eukaryotic gene order. Nat Rev Genet 5:299–310

Ihaka R, Gentleman R (1996) R: a language for data analysis and graphics. J Comput Graph Stat 5:299–314

Jacob F, Perrin D, Sanchez C, Monod J (1960) Operon: a group of genes with the expression coordinated by an operator. C R Hebd Seances Acad Sci 250:1727–1729

Kojima KK, Matsumoto T, Fujiwara H (2005) Eukaryotic translational coupling in UAAUG stop-start codons for the bicistronic RNA translation of the non-long terminal repeat retrotransposon SART1. Mol Cell Biol 25:7675–7686

Lawrence JG (2002) Shared strategies in gene organization among prokaryotes and eukaryotes. Cell 110:407–413

Lee SJ (1991) Expression of growth/differentiation factor 1 in the nervous system: conservation of a bicistronic structure. Proc Natl Acad Sci USA 88:4250–4254

Long M (2000) A new function evolved from gene fusion. Genome Res 10:1655–1657

Long M, Betran E, Thornton K, Wang W (2003) The origin of new genes: glimpses from the young and old. Nat Rev Genet 4:865–875

Mandal MN, Vasireddy V, Jablonski MM, Wang X, Heckenlively JR, Hughes BA, Reddy GB, Ayyagari R (2006) Spatial and temporal expression of MFRP and its interaction with CTRP5. Invest Ophthalmol Vis Sci 47:5514–5521

Mandel SJ, Berry MJ, Kieffer JD, Harney JW, Warne RL, Larsen PR (1992) Cloning and in vitro expression of the human selenoprotein, type I iodothyronine deiodinase. J Clin Endocrinol Metab 75:1133–1139

Meyer D (2006) Support vector machines: the interface to libsvm in package e1071. Technische University Wien, Austria

Mullenbach GT, Tabrizi A, Irvine BD, Bell GI, Hallewell RA (1987) Sequence of a cDNA coding for human glutathione peroxidase confirms TGA encodes active site selenocysteine. Nucleic Acids Res 15:5484

Pandey J, Koyuturk M, Subramaniam S, Grama A (2008) Functional coherence in domain interaction networks. Bioinformatics 24:i28–i34

Pruitt KD, Tatusova T, Maglott DR (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 35:D61–D65

Raghavachari B, Tasneem A, Przytycka TM, Jothi R (2008) DOMINE: a database of protein domain interactions. Nucleic Acids Res 36:D656–D661

Reiss J, Cohen N, Dorche C, Mandel H, Mendel RR, Stallmeyer B, Zabot MT, Dierks T (1998) Mutations in a polycistronic nuclear gene associated with molybdenum cofactor deficiency. Nat Genet 20:51–53

Rice P, Longden I, Bleasby A (2000) EMBOSS: the European molecular biology open software suite. Trends Genet 16:276–277

Salvatore D, Low SC, Berry M, Maia AL, Harney JW, Croteau W, St Germain DL, Larsen PR (1995) Type 3 iodothyronine deiodinase: cloning, in vitro expression, and functional analysis of the placental selenoenzyme. J Clin Invest 96:2421–2430

Salvatore D, Bartha T, Harney JW, Larsen PR (1996) Molecular biological and biochemical characterization of the human type 2 selenodeiodinase. Endocrinology 137:3308–3315

Schuckelt R, Brigelius-Flohe R, Maiorino M, Roveri A, Reumkens J, Strassburger W, Ursini F, Wolf B, Flohe L (1991) Phospholipid hydroperoxide glutathione peroxidase is a selenoenzyme distinct from the classical glutathione peroxidase as evident from cDNA and amino acid sequencing. Free Radic Res Commun 14:343–361

Shigemoto K, Brennan J, Walls E, Watson CJ, Stott D, Rigby PW, Reith AD (2001) Identification and characterisation of a developmentally regulated mammalian gene that utilises −1 programmed ribosomal frameshifting. Nucleic Acids Res 29:4079–4088

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

Stallmeyer B, Drugeon G, Reiss J, Haenni AL, Mendel RR (1999) Human molybdopterin synthase gene: identification of a bicistronic transcript with overlapping reading frames. Am J Hum Genet 64:698–705

The UniProt Consortium (2010) The universal protein resource (UniProt) in 2010. Nucleic Acids Res 38:D142–D148

Wilson R, Ainscough R, Anderson K, Baynes C, Berks M, Bonfield J, Burton J, Connell M, Copsey T, Cooper J et al (1994) 2.2 Mb of contiguous nucleotide sequence from chromosome III of *C. elegans*. Nature 368:32–38

Zdobnov EM, Apweiler R (2001) InterProScan—an integration platform for the signature-recognition methods in InterPro. Bioinformatics 17:847–848