Journal of Molecular Evolution (2019) 87:93–105
https://doi.org/10.1007/s00239-019-09889-1

ORIGINAL ARTICLE

Conserved Critical Evolutionary Gene Structures in Orthologs

Miguel A. Fuertes[1] · **José R. Rodrigo**[2] · **Carlos Alonso**[1]

Received: 19 July 2018 / Accepted: 13 February 2019 / Published online: 28 February 2019

Abstract
Unravelling gene structure requires the identification and understanding of the constraints that are often associated with the evolutionary history and functional domains of genes. We speculated in this manuscript with the possibility of the existence in orthologs of an emergent highly conserved gene structure that might explain their coordinated evolution during speciation events and their parental function. Here, we will address the following issues: (1) is there any conserved hypothetical structure along ortholog gene sequences? (2) If any, are such conserved structures maintained and conserved during speciation events? The data presented show evidences supporting this hypothesis. We have found that, (1) most orthologs studied share highly conserved compositional structures not observed previously. (2) While the percent identity of nucleotide sequences of orthologs correlates with the percent identity of composon sequences, the number of emergent compositional structures conserved during speciation does not correlate with the percent identity. (3) A broad range of species conserves the emergent compositional stretches. We will also discuss the concept of *critical gene structure*.

Keywords Molecular evolution · Triplet-composon · Gene structure · Human–mouse orthologs

Introduction

The apparent lack of patterns resulting from the distribution of bases in gene sequences has been for a long time one of the most interesting and intriguing question marks of molecular biology. In fact, intense research efforts have been made to identify whether and how these patterns are distributed along the length of genomic sequences and how these patterns can be visualized. Actually, the search for DNA patterns in genes and genomes requires comparing the compositional characteristics that genes show in and among species. Between others, the methods for the identification of DNA patterns have been based on genetic and evolutionary considerations (Sueoka 1962), on linear regression models (Dai et al. 2007), on the analysis of nucleotide (NT) asymmetries (Arnold et al. 1988) and on graphical representations of DNA sequences (Gates 1986; Leong and Morgenthaler 1995; Nandy 2009; Roy et al. 1988) among others. Thus, an important and significant step forward toward the identification of genomic patterns will be based on the detection of genes and signals able to rule the function of genes across species (Notebaart et al. 2005). In addition, it is of the outmost relevance identifying the functional information contained in genomes and the way in which the informational content is arranged in chromosomes and in the high order chromatin organization (Gingeras 2009; Takeda 2012). This debate exemplifies the relevance of the identification and analysis of DNA patterns and the clarification of their functional significance. Thus, the analysis of the structure of genes and of the linear organization of nucleotides along the DNA sequence constitutes an active and wide field of investigation (Zhu et al. 2009). For example, while mouse and humans have maintained a substantial divergence at the DNA level (Yue et al. 2014), their genomic differences

Handling editor: **Hideki Innan**.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00239-019-09889-1) contains supplementary material, which is available to authorized users.

✉ Miguel A. Fuertes
mafuertes@cbm.csic.es

José R. Rodrigo
jrrodrigo@yahoo.es

Carlos Alonso
calonso@cbm.csic.es

[1] Centro de Biología Molecular "Severo Ochoa" (CSIC-UAM), Universidad Autónoma de Madrid, c/Nicolás Cabrera 1, 28049 Madrid, Spain

[2] Telefónica de España S.A., Gran Vía 28, Madrid, Spain

remain to be fully characterized. In fact, it has been recently realized that intron patterns may harbour a variety of elements that regulate transcription (Comeron 2001; Gazave et al. 2007), a variety of elements that may harbour the role of non-coded RNAs (Mattick and Gagen 2001) and elements that regulate the splicing control (Majewski and Ott 2002). Thus, while the function of exon sequences would be to encode proteins (Gilbert 1978) the function of intron sequences might be to encode gene regulation information (Costas et al. 2004; Robart et al. 2007; Robart and Zimmerly 2005; Rogozin et al. 2005; Wang et al. 2005). In order to maintain an efficient gene function (Amit et al. 2012; Parmley et al. 2007) it is critical to retain the structural information of the coding and non-coding DNA. The fine regulation of the splicing machinery (Gelfman et al. 2012; Keren et al. 2010; Rogozin et al. 2002) suggests that to maintain the structural DNA patterns there might be strong compositional constraints acting at the DNA and chromatin level (Louie et al. 2003; Schwartz et al. 2009).

Recently, based on the composon concept (Fuertes et al. 2011), a new method revealing emerging DNA patterns not observed previously has been proposed. This outcome results from the application of the tCP-methodology to human–mouse orthologs (Fuertes et al. 2016b). The frequency-usage of sets of DNA triplets having the same gross composition, called triplet-composons (or briefly, tCPs), is the measurement parameter of the tCP-methodology. It is recognized that the use of this method results to be useful for categorizing by their tCP-usage the coding and intron sequences from a large number of orthologs in a few number of families. By using this method it has been shown that exons and introns of human–mouse orthologs do not evolve independently (Fuertes et al. 2016a). In addition, the analysis of a large number of genes from some of those categorizations revealed that the bulk of these orthologs share common functional similarities (Fuertes et al. 2016c) despite having dissimilar NT-sequences. For example, while at the NT-level some of the genes contained in the Rhodopsin G-protein coupled receptor superfamily were considered not to be orthologs, they have a high degree of similarity when we look at their tCP-usage (Fuertes et al. 2016b). The main difference between the data and the analyses presented in this manuscript and those previously reported is that to elaborate the data previously reported we counted NT-triplets in the DNA sequence in a fully overlapping way analysing the tCP-usage frequency. The tCP-usage gives information about how many and what tCPs are contained in the entire DNA sequence but not about their distribution along the length of the sequence. In this paper, however, we counted tCPs focussing on their distribution along the length. By the analysis of the tCP distribution, the NT-compositional architecture of orthologs emerges. In this context, we will also discuss the concept of NT- and tCP-homology.

Materials and Methods

DNA Sequence Acquisition and Pre-processing

Human–mouse orthologs were downloaded from Ensembl gene database release 86—Oct 2016 © WTSI/EMBL-EBI (Yates et al. 2016). The criteria to collect the orthologs takes into account that all selected orthologs must belong in mouse and human to different tCP-clusters (Online Resources 1 and 2) (Fuertes et al. 2016b). Consequently, the human–mouse orthologs selected will diverge in their tCP-sequences. In the study we only consider the genes coding for the longest transcripts (Fuertes et al. 2016c). We identified the genes of the dataset by the gene name and their corresponding protein description. Online Resource 1 also indicates how many genes are in each cluster. In parallel, a sample of human–mouse orthologs contained in the same tCP-cluster (named sample 2) were also analysed (see Online Resource 3). In this paper, we analysed the coding sequences of all human–mouse orthologs from samples 1 and 2.

Calculation Methods

To analyse the similarities and dissimilarities of human–mouse orthologs in relation to the gene length we set up an analysis based on the tCP-methodology (Fuertes et al. 2011, 2016b). We agreed to use this method to investigate because it is very difficult to obtain the conserved DNA structures here in described by other existing methods. The justification for the tCP-methodology is theoretical and based on the existence of exclusionary multiplets characterized by the presence or absence of particular bases. It was demonstrated that the number of such groups of exclusionary multiplets (later called "triplet-composons" or tCPs), resulted to be 14 and the minimum length of the multiplet was 3 NTs (a triplet) (Fuertes et al. 2011). The 14 tCPs contain all possible nearest-neighbours that can be formed with four DNA bases, forming in this way a close system (see Table 1). The "fully overlapping reading" guarantees that all triplets of a DNA sequence are considered in the study avoiding the lack of information. Finally, we consider that the lost triplets, when reading the DNA in non-overlapping way (as is the case of the genetic code), gives relevant evolutionary information. The notation of tCPs and their associated sets of NT-triplets are in Table 1. To evaluate the distribution of tCPs along the gene length we estimated the cumulative tCP-usage. The cumulative tCP-usage would be then the sum of all previous tCP events up to the current length. To facilitate the visualization of the distribution of tCPs along

Table 1 tCP-code

Composon	Triplets associated to composons
<A>	AAA
<T>	TTT
<G>	GGG
<C>	CCC
<AC>	AAC, CAA, ACA, CCA, ACC, CAC
<AT>	AAT, TAA, ATA, TTA, ATT, TAT
<AG>	AAG, GAA, AGA, GGA, AGG, GAG
<CG>	CCG, GCC, CGC, GGC, CGG, GCG
<GT>	GGT, TGG, GTG, TTG, GTT, TGT
<CT>	CCT, TCC, CTC, TTC, CTT, TCT
<AGC>	AGC, GCA, CAG, ACG, CGA, GAC
<AGT>	AGT, GTA, TAG, ATG, TGA, GAT
<ACT>	ACT, CTA, TAC, ATC, TCA, CAT
<TCG>	TCG, CGT, GTC, TGC, GCT, CTG

List of all composons and their associated nucleotide triplets Fuertes et al. (2011, 2016b)

the DNA sequence we applied the methodology described to the human gene CYP1A2. The gene that codifies for Cytochrome P450 1A2, is a typical gene from the dataset responsible for the metabolism of estrogens and many exogenous compounds (Hong et al. 2004). The cumulative graph of the tCP <AC> of CYP1A2 and its corresponding trend line are shown in Fig. 1A. For comparative purposes, it may be observed that the tCP-profile of <AC> is the projection of the cumulative data of Fig. 1A on the length axis (Fig. 1B). This is equivalent to subtract the cumulative tCP-usage frequency from the trend line. Note that there are gene stretches having <AC>-usages higher and lower than the trend line giving precise visual and numerical information of the distribution of <AC> with the gen length. Thus, the <AC> distribution shows a characteristic compositional structure along the gene sequence. Each one of the 14 tCPs distributions (Table 1) displays in each human–mouse ortholog a characteristic compositional structure, the tCP-profile. We represent the distribution of all tCPs of each ortholog by a panel containing the fourteen different tCP-profiles.

Criteria Used to Highlight the Conserved tCPs Along the Gene Length

The criteria used to highlight the conserved tCPs were determined as follow: (i) we translated into tCP-sequences all NT-sequences from each ortholog by using Table 1 (samples 1 and 2). (ii) We aligned the tCP-sequences by means of a dynamic algorithm used for the global alignment of two sequences (Kruskal 1983; Needleman and Wunsch 1970). (iii) each tCP and the place occupied in the aligned tCP-sequences was recorded and (iv) to compare the tCP-profile of each ortholog a graphical representation of the data was done. We determine the level of similarity between tCP-sequences of each ortholog using a restrictive cut-off for the Pearson correlation coefficient of $r \geq 0.850$. We choose a restrictive cut-off to guarantee that the tCP-profiles of each ortholog have a high level of resemblance. We should be aware that in general correlation does not imply causation (Aldrich 1995) unless we are dealing with ortholog genes. In that case, we may infer that a causal relationship exists between the correlation index and orthology.

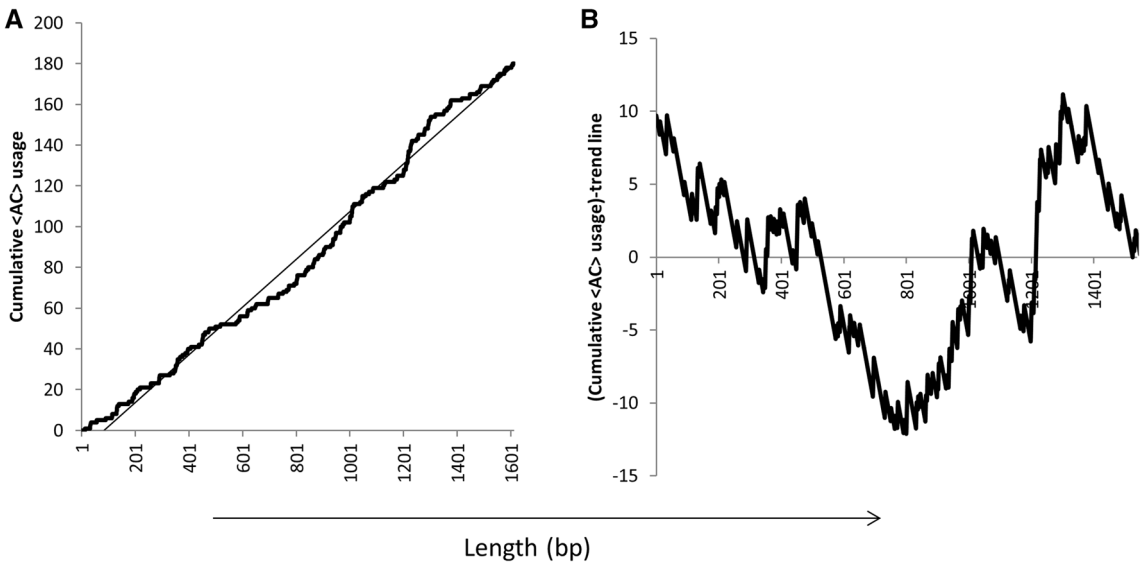
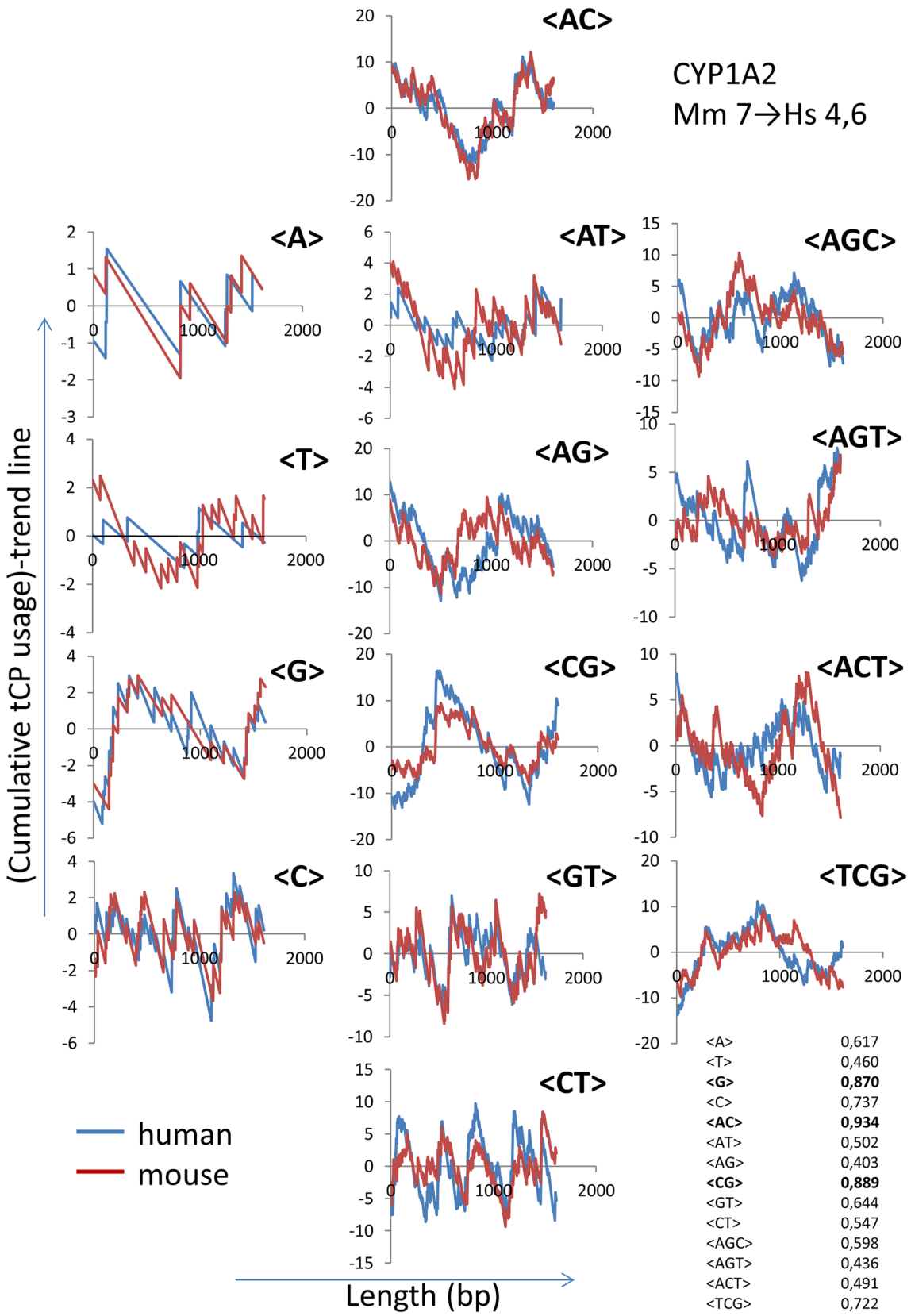


Fig. 1 Distribution of the tCP <AC> along the human gene CYP1A2. (**A**) Cumulative graph (thick line) and the corresponding trend line (thin line), for CYP1A2. (**B**) tCP-profile of CYP1A2 obtained by the projection on the length axis of the trend line


<AC>
CYP1A2
Mm 7→Hs 4,6
<A>
<AT>
<AGC>
<T>
<AG>
<AGT>
<G>
<CG>
<ACT>
<C>
<GT>
<TCG>
<CT>
(Cumulative tCP usage)-trend line
Length (bp)
human
mouse
<A> 0,617
<T> 0,460
<G> 0,870
<C> 0,737
<AC> 0,934
<AT> 0,502
<AG> 0,403
<CG> 0,889
<GT> 0,644
<CT> 0,547
<AGC> 0,598
<AGT> 0,436
<ACT> 0,491
<TCG> 0,722

◂**Fig. 2** Panel comparing the fourteen tCP-profiles with the length of the human–mouse ortholog CYP1A2. Red and blue lines correspond to distributions of tCPs along the trend line of the mouse and the human cumulative tCP-usages, respectively. The insets display in the upper right corner the name of the ortholog, the mouse (7) and the human (4 and 6) clusters containing the ortholog, and in the bottom right corner a table with the correlation (r) found between the human–mouse tCP-profiles for numerical comparison. In bold the r values higher than the cut-off. (Color figure online)

As control of non-random distribution of tCPs with the length, we analysed the tCP-profiles of randomly generated NT-sequences having a NT-composition identical to that of the ortholog under study. To generate these sequences, we use the utility *shuffleseq* that shuffles a set of sequences maintaining the NT-composition. This tool is in the EMBOSS explorer, a graphical user interface of the EMBOSS suite of bioinformatics tools (Mullan and Bleasby 2002; Olson 2002). We confirmed that identical NT-sequences have identical tCP-profiles as expected (data not shown).

Results

tCP-Profile of CYP1A2, a Typical Human–Mouse Ortholog

To ascertain that the tCP compositional patterns are conserved along the gene length, in a first instance, we studied the tCP-profiles of all pairs of human–mouse orthologs described in the dataset (samples 1 and 2). As an example, we analysed the tCP-profile of the human–mouse ortholog CYP1A2. In mouse, the CYP1A2 gene is present in the tCP-cluster 7. In human, however, the gene is present in tCP-clusters 4 and 6 (see Online Resource 1). The percent identity of the NT- and tCP-sequences of the human–mouse ortholog is 80% and 61%, respectively. The data are listed in Online Resource 2.

Figure 2 shows the panel representing the tCP-profiles of each one of the 14 tCPs along the length of CYP1A2. When the tCP-sequences of the ortholog were compared, three tCPs (<G>, <AC> and <CG>) display a high level of similarity showing a correlation higher than the cut-off ($r \geq 0.850$). The numerical inset of Fig. 2 lists the correlation values of each tCP between both species. The correlation values ranged from $r = 0.403$ for <AG> to $r = 0.934$ for <AC>. We also observed the existence of short correlated stretches in some tCP-profiles of the panel (Fig. 2). These short stretches were not taking into account since they not contribute significantly to the correlation of the profile. Other tCP-profiles show a moderate correlation as it would be the case of <C> and <TCG> with positive correlations of $r = 0.737$ and 0.722, respectively. Since these correlations were near the cut-off but lower, they were discarded from the analysis. Low correlation levels ranging from $0.403 < r < 0.644$ were also observed for <A>, <T>, <AT>, <AG>, <GT>, <CT>, <AGC>, <AGT> and <ACT>. It is worthwhile to point out that despite the high percent identity of NT-sequences of the ortholog CYP1A2, only three tCPs show correlations higher than the cut-off. The fact that only 3 out of 14 tCPs shows correlations higher than the cut-off could partially explain the high differences in percent identity between tCP-sequences and NT-sequences (Online Resources 2 and 3). We have obtained in a similar way to that for CYP1A2, the conserved tCP-profiles of all human–mouse orthologs described in the dataset.

Correlations Between NT- and tCP-Sequences Relative to the Gene Length

As we previously reported, there is a mathematical expression linking NT- and tCP-composition of DNA sequences (Fuertes et al. 2016b). Now, in order to know whether there is along the gene length a correlation between both the NT- and the tCP-sequences we align the NT- and the tCP-sequences of samples 1 and 2. The numerical results of these alignments are listed in Online Resources 2 and 3. Online Resource 2 (sample 1) lists those orthologs that differ between mouse and human in the tCP-cluster. The table also shows the percent identity and of gaps obtained from the alignments in addition to the number of conserved tCPs per ortholog. We have obtained similar data from Online Resource 3 (sample 2) that lists orthologs belonging to the same tCP-cluster in mouse and human (Fuertes et al. 2016b).

Figure 3A displays the data obtained from the alignments of samples 1 and 2 giving the percent identity and of gaps of NT- and tCP-sequences. In average, the genes from sample 1 display higher dispersion in percent identity and of gaps than the genes from sample 2. A detailed inspection of Fig. 3A indicates that there is also a certain resemblance between the NT- and tCP-percent identities and of gaps. Figure 3B shows that independently of the cluster to which samples 1 and 2 pertain, there is a notable correlation between the profiles shown in Fig. 3A (Fuertes et al. 2016b). The inset of Fig. 3B shows the distribution of identities in sequence alignments of NTs and tCPs showing, in all cases, higher identities for NT- than for tCP-sequences. Figure 3C shows the correlation between the percent of gaps in NT- and tCP-sequences. The inset of Fig. 3C shows a high accumulation of short gaps near the origin in both NT- and tCP-sequences. For clarity, we represent the data in a semi-logarithmic plot.

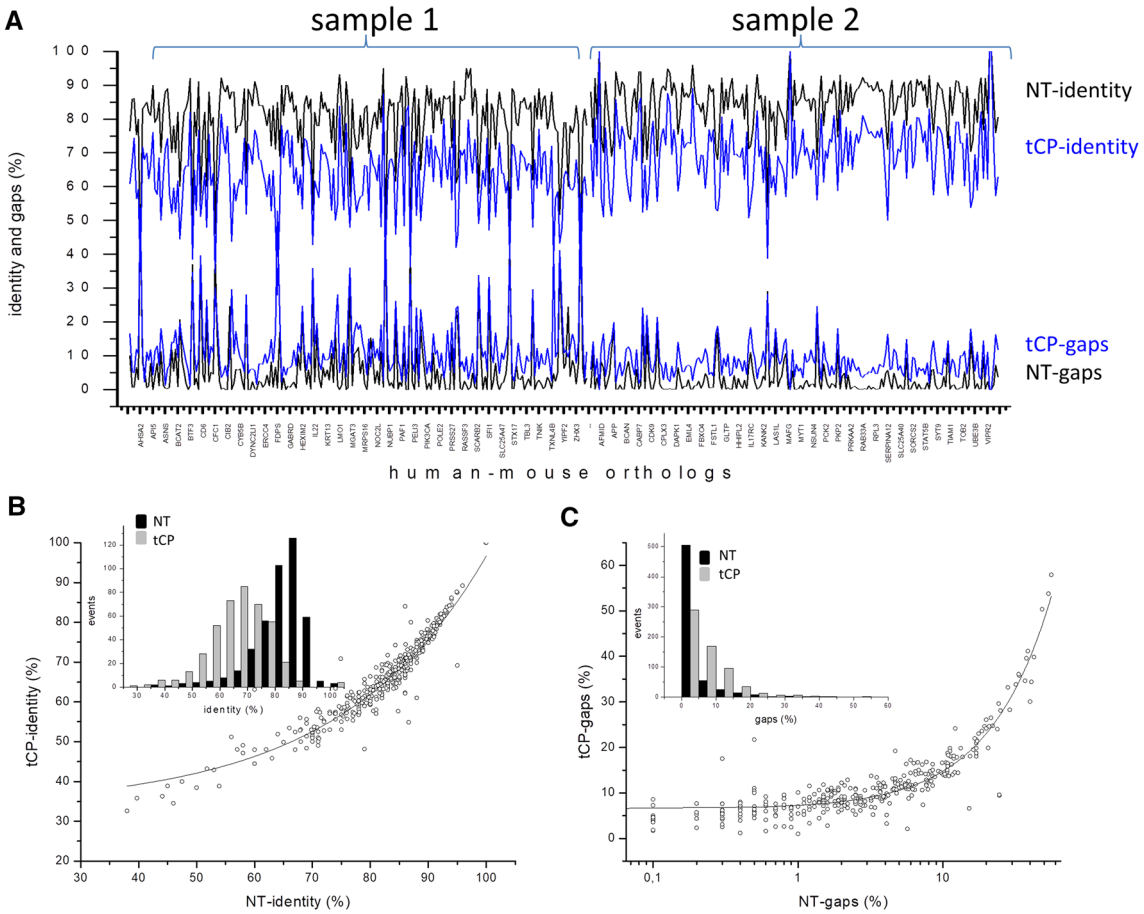


Fig. 3 Percent identity and gaps calculated per human–mouse ortholog in samples 1 and 2 (Online Resources 2 and 3). (**A**) The percent identity in alignments of NT- (upper black line) and tCP-sequences (upper blue line) and the percent of gaps in alignments of NT- (lower black line) and tCP-sequences (lower blue line). (**B**) Correlation observed in the percent identity between NT- and tCP-sequences of samples 1 and 2. (**C**) Correlation observed in the percent of gaps between NT- and tCP-sequences of samples 1 and 2. Insets of (**B, C**) illustrate differences on the percent identity and gaps between NT- (filled black square) and tCP-sequences (filled grey square), respectively. (Color figure online)

Distribution of Conserved tCPs Along Human–Mouse Orthologs

Since the NT- and tCP-sequences are highly correlated, we further will analyse the type and amount of tCPs shared by the orthologs of samples 1 and 2. When representing the number of orthologs sharing the same tCP (Fig. 4A), we observed that the sequences maintain, in most cases, the tCP <AG> and to a lower extent the <CT>. About 73% of the genes from samples 1 and 2 share the <AG>-profile. The <AT>-profile is the least conserved in sample 1 and the <AGC> is the least conserved in sample 2. We detected that human–mouse orthologs share non-degenerated tCPs <A>, <G>, <T> or <C> in a similar extent or even higher than some of the tCPs with degeneracy as <AC>, <AT>, <AGC>, <ACT> or <TCG>. We should notice, then, that non-degenerated tCPs are also under selection constraint. Figure 4B shows how many conserved tCPs are per ortholog in both samples. We noticed that 90% of orthologs from sample 1 and 74% from sample 2 share less than 6 out of the 14 tCPs. Thus, a notable fraction of the orthologs analysed share less than half of all tCPs, independently on the percent identity of both the NT- and tCP-ortholog sequences. Only 10% of sample 1 and 26% of sample 2 share more than six tCPs. We also observed that 10% of orthologs from sample 1 and 6% of sample 2 does not share any tCP probably due to correlations lower than the cut-off ($r \geq 0.850$). We observed, in addition, that there is not a significant correlation between the percent identity of NT-sequences and the number of conserved tCPs per ortholog (Fig. 4C). The point clouds of Fig. 4C from samples 1 and 2 are highly dispersed ($r < 0.575$) despite the high correlation between percent identities of NT- and tCP-sequences as it was observed in Fig. 3B. The same data would be obtained when the tCP-sequences instead of the NT-sequences are analysed. In some cases, as expected, the higher the percent

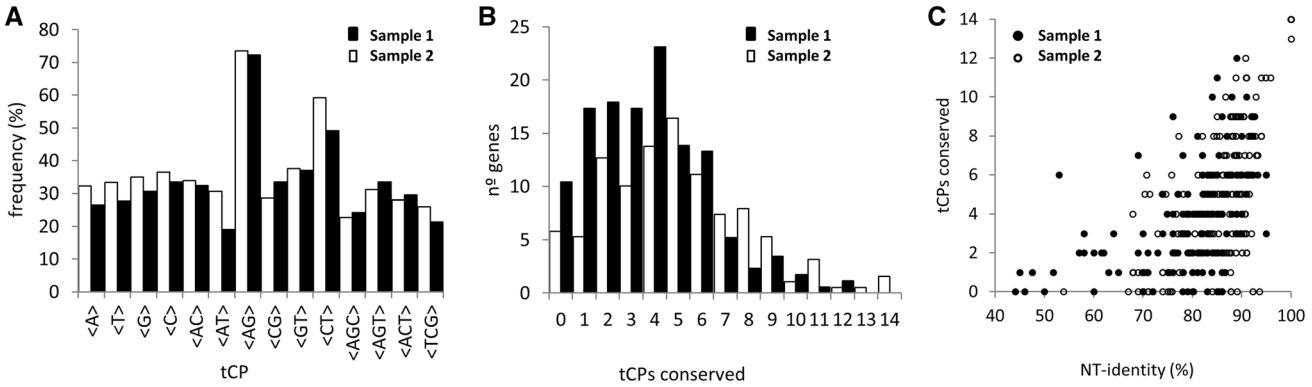


Fig. 4 (**A**) Distribution of orthologs relative to the type of tCP conserved (in %). (**B**) Distribution of orthologs relative to the number of tCPs conserved. (**C**) Graph representing the percent identity of NT-sequences per ortholog relative to the number of conserved tCPs. In all cases, we represent together the sample 1 (black) and 2 (white) for comparison

identity between the NT-sequences the higher the number of conserved tCPs. There are also cases in which high percent identity between NT-sequences is associated with low number of conserved tCPs. The result illustrates a basic difference between the evolutionary information supplied by the NT-sequences and the tCP-sequences. In fact, it has been assumed that the evolutionary information supplied by the NT-sequences is implicit to the high percent identity between orthologs (Pearson 2013). However, we found that in a significant number of cases (Fig. 4C) high percent identity between ortholog NT-sequences does not guarantee the existence of a high number of conserved tCPs. Moreover, in some cases, low percent identity between NT-sequences is associated with a notable number of conserved tCPs. In the light of the neutral theory of evolution, the low percent identity observed in some non-conserved ortholog regions could be attributed to changes in functionally less important regions of the ortholog sequence different from the emergent tCP-conserved structure.

NT-Versus tCP-Sequence Homology

A high percent identity between orthologs does not guarantee the existence of a high number of conserved tCPs. This allowed us to hypothesize that a high percent identity of NT-sequence between orthologs may overlook some inner structures biologically relevant and highly informative as those corresponding to a low number of conserved tCPs. Being that so, we believe that the comparison of tCP-profiles between orthologs may be more informative than the comparison of the NT-profiles. For simplicity and in order to prove the consistency of the hypothesis, we have chosen the short and highly identical human–mouse ortholog SAMD12. The gene codifies for the sterile alpha motif domain-containing protein 12 involved in inter-chromosomal and intra-chromosomal rearrangement events (Zhao et al. 2009). The alignment of the NT-sequences of SAMD12 revealed 89% identity between human and mouse. However, human SAMD12 and its mouse-ortholog share only 4 out of 14 tCPs (<T>, <G>, <AT> and <AG>) with correlations higher than the cut-off (Online Resource 5). The high identity of NT-profiles and the low number of conserved tCPs between both species suggests that the analysis of tCPs is, from an evolutionary point of view, more informative than the NT-sequence analysis. When decoding the conserved tCP-sequences we observe the associated NT-stretches corresponding to the conserved tCPs. The conserved structure is interspersed in stretches along the CDS length.

In Fig. 5 we visualize the conserved tCP-structure at the NT-level of the human and mouse ortholog sequences of SAMD12. Decoding the conserved tCP-sequences into NT-sequences (Table 1) we visualize the stretches of the NT-sequences that are associated with the conserved tCPs. We observed the common compositional structure characteristic of this human–mouse ortholog. As each degenerated tCP is associated with six different NT-triplets (see Table 1) it would be possible to find NT-mismatches in the conserved structure of SAMD12. In fact, as it can be observed in Fig. 5 we can find NT-mismatches at locations 69 (T→A), 99 (G→A), 122 (A→G), 423 (G→A) and 429 (A→G). Thus, the conservation of the tCP-sequence does not necessarily imply conservation of the associated NT-sequence.

As indicated above (Fig. 4C), the low number of conserved tCPs of SAMD12 is not linked to the high percent identity (89%) in the NT-sequence. Observe that 40% of coincidences in alignments of SAMD12 (195 in 486 bp) are due to conserved tCPs. The remaining 49% identity could be attributed to several factors: (i) to the number of coincidences in short stretches of the gene; (ii) to coincidences attributable to correlations lower but near the cut-off and (iii) to a certain amount of *at random* coincidences expected to happen by chance or changes due

Fig. 5 NT-sequence alignment of the human–mouse ortholog SAMD12. The grey shaded areas of both sequences represent the conserved NT-triplets (see Table 1) contained in the conserved tCPs <T>, <G>, <AT> and <AG> of SAMD12 in human and mouse. The blue shaded areas represent the common stretches conserved of SAMD12 in mouse and human. A line connecting both sequences (|) represents a match and a point (.) represents a mismatch. (Color figure online)

to fluctuations about the mean. Some of these changes, probably, could correspond to tCPs non-correlated at all or even negative-correlated. The limited number of changes in the conserved sequence can be a consequence of strong evolutionary constrains acting over the conserved tCP stretches. The calculations demonstrated that the number of NT-mismatches for the conserved tCPs <T>, <G>, <AG> and <AT> in the shared emergent structure of SAMD12 relative to the non-shared one results to be 5 out of 195 NTs (2.6%), against 22 out of 53 NTs (41.5%) for the same tCPs. The data indicate that the emergent conserved structure is under strong evolutionary constraints.

Critical Emerging Evolutionary Gene Structures

It has been accepted that the critical processes leading to the evolution of gene structure are the generation of mutations and the filtering of these mutations by natural selection. The existence of conserved emerging gene structures, as those reported in this paper, leads us to ask whether such interspersed structures are particularly prone to mutations. In order to ask whether the compositional structure identified in the human–mouse ortholog SAMD12 is preserved in evolutionary related species, we compared the human–mouse conserved sequence with those of orthologs of SAMD12

from other species. In addition to *Homo sapiens* (humans) and *Mus musculus* (mouse) we have considered to be orthologs of SAMD12 *Pongo abelii* (Sumatran orangutan), *Pan troglodytes* (Chimpanzee), *Ovis aries* (sheep), *Tursiops truncatus* (dolphin), *Canis lupus* (dog), *Choloepus didactylus* (sloth), *Bos Taurus* (cow) and *Oryctolagus cuniculus* (rabbit). We also considered being orthologs of SAMD12 two non-mammalian species as *Anolis carolinensis* (lizard) and *Danio rerio* (zebrafish). The multiple alignment of NT-sequences of SAMD12 (Online Resource 4) generates the phylogenetic tree of Fig. 6. Online Resource 4 shows 42 gene stretches that give rise to the conserved structure of SAMD12 corresponding to the tCPs <T>, <G>, <AG> and <AT> ($r \geq 0.850$) in all species. In the zebrafish 10 out of 42 stretches are conserved in their entirely and 11 are largely conserved despite the long evolutionary distance from mammals (Fig. 6). The lizard shares with mammals 23 out of 42 stretches conserved in their entirely and 11 largely conserved. The bar graphic of the inset of Fig. 6 shows, for each species, the percent of stretches conserved in their entirely and those largely conserved. In mammals, the amount of stretches conserved in their entirely is higher than 79% (more than 33 stretches out of 42). Taking into account the entirely conserved stretches, the tCP <AG> is present in 25 out of 42 (~ 60%) stretches in contrast with <AT> that is present in 8 out of 42 (19%), as can be observed in Online Resource 4. The multiple sequence alignment of SAMD12 (Online Resource 1) reveals that the genomic regions in which all species share the tCP <AG>, a very low number of mismatches are observed relative to that found in regions in which only one of the species has <AG>. This occurs especially in regions of insertions and deletions (as the 3′ and 5′ regions). We concluded that the emergent tCP-pattern of SAMD12 would be under selection constraint for the conserved tCP < AG>, as an indication that such patterns could have importance in fitness.

Discussion

About the Methodology

It has been suggested that other codes besides the genetic code can be used to analyse genomes in and out CDS (Parker and Tullius 2011; Pearson 2006; Trifonov 2011). The search for new codes stems from the suggestion that regulatory

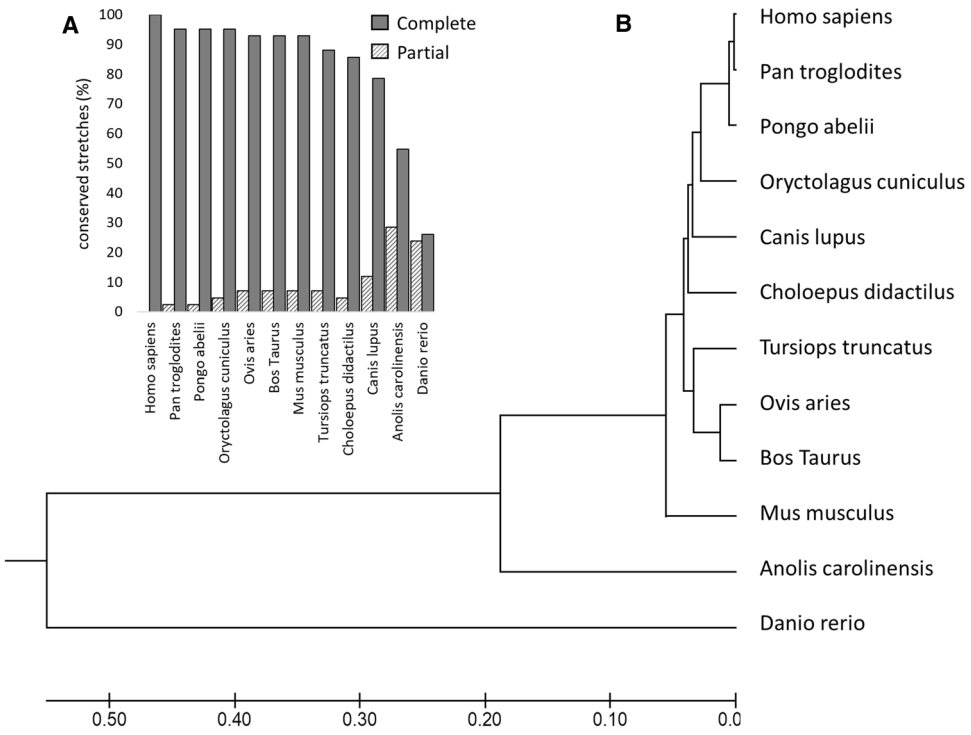


Fig. 6 Evolutionary relationships of taxa and conserved gene structures. (**A**) The histogram shows the percent of complete structures (open square) and the partial conserved structures (filled black square) in each species. (**B**) The evolutionary history of SAMD12 was inferred using the UPGMA method (Sneath and Sokal 1973). The optimal tree with their corresponding branch lengths is shown. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the Maximum Composite Likelihood method (Tamura et al. 2004) and are in the units of the number of base substitutions per site. The analysis involved 12 NT-sequences. All positions containing gaps and missing data were eliminated. There was a total of 434 positions in the final dataset. Evolutionary analysis was conducted in MEGA7 (Kumar et al. 2016)

mutations would have a larger biological impact than would mutations in CDS (King and Wilson 1975). The methodology herein used to reveal emergent compositional structures use the concept of *triplet-composon* or tCP that provides useful evolutionary information in and out CDS (Fuertes et al. 2011, 2016a, b). Although the tCP-code is close to the genetic code in many aspects, it is very different in others. We think, however, the tCP-code is a code in a strict sense (Table 1) because an input generates an output. Actually, there are properties shared by both codes: (i) to decode genomic information both codes operate reading NT-triplets and (ii) both codes are degenerated. However, the genetic code and the tCP-code differ in some basic aspects: (i) in the genetic code the DNA is read in a non-overlapping way in the correct reading frame (CDS), while the tCP-code is read in fully overlapping way, and consequently, in any reading frame. (ii) The genetic code translates NT-sequences (CDS) into protein sequences while the tCP-code translates NT-sequences into tCP-sequences. Alignments of tCP-sequences are the basis to obtain useful evolutionary information from ortholog genes since they clearly manifest, between others, the existence of highly conserved stretches. NT- and tCP-sequences are connected by compositional parameters as the tCP-usage and the NT-composition (Fuertes et al. 2016b). In this paper, we demonstrated that there are local conserved structures inside CDS (see Fig. 3). We show that orthologs share common compositional structures not observed previously which are interspersed along the length of CDS. To do that, we have use a modified version of the original tCP-method consisting in the identification of the position of each tCP along the gene length. The data presented indicate that evolution may take advantage of such inner structures suggesting in some cases shared biology and in other cases revealing differences that could shed light on what makes species different.

We conclude by indicating the strengths of analysing tCP-sequences instead of nucleotide sequences: (i) the "fully overlapping reading" guarantee that all triplets of a DNA sequence are taken into account, avoiding the lack of relevant contextual evolutionary information. (ii) The independence on the reading frame allows to study not only CDS but also introns, intergenic sequences, etc., that could be of interest in mutational analysis, helping to research on gene structure and function, species identification and in the study of genetic diseases. (iii) The simplicity of the method allows an easy access to both the numerical and the visual data in a fast way. (iv) The control of the correlation coefficient allows tuning the level of constriction of the study. (v) The concept of "composon" as a group of triplets with the same gross composition (see Table 1) brings a new perspective to the study of compositional aspects not considered previously revealing innovative properties of DNA sequences as those reported in this paper.

About the Correlations Between tCP and NT-Sequences

In addition to the fact that there is a correlation between the NT-composition of a DNA sequence and its tCP-usage (Fuertes et al. 2016b) the data presented in this paper show that there are correlations between percent of tCP- and NT-identities and gaps (Fig. 3 and Online Resources 2 and 3). Gaps are due to insertions and deletions (INDELs) into the sequences compared (Bhangale et al. 2005; Weber et al. 2002). We noticed in addition that the identities obtained from alignments of ortholog NT-sequences were always higher that those observed in alignments of their corresponding tCP-sequences (Fig. 3B). This is because a single substitution in a NT-sequence implies 1–3 changes in the corresponding tCP-sequence because of the fully overlapping reading during translation of NT- to tCP-sequences. Moreover, when INDELs are present the number of differences between NT- and tCP-sequences increases faster in tCP-sequences than in NT-sequences (Fig. 3C). A notable decrease in the number of INDELs in NT-sequences indicates that the percent identity observed would be higher in NT-sequence alignments that in tCP-sequence alignments. In agreement with our data, simulations carried out on human–mouse orthologs indicate (with an error of 2%) that INDEL rates are up to twice higher than it would be deduced from the NT-sequence alignments (Lunter 2007). In human, small INDELs are found often in exons. Some lines of evidence point to such variations as a major determinant of the human biological diversity (Mills et al. 2011).

About the New Conserved tCP-Structure

We know that an accepted strategy for finding functional sequences is the search for conserved nucleotide stretches through species (Hardison et al. 1997; Kellis et al. 2003; Woolfe et al. 2005). The accumulation of sequenced genomes and the increase and the improvement of tools that permit whole genomes to be aligned (Blanchette et al. 2004; Bray and Pachter 2004) lead us to search for conserved sequences and to use the data obtained to create new hypotheses and experiments (Frazer et al. 2001; Woolfe et al. 2005). We think that the main outcome of this paper highlights when we align and compare ortholog tCP-sequences. The outcome deals with the existence of highly conserved compositional structures of conserved tCPs. The number of conserved compositional structures will depend on the tuning of the cut-off. As indicated in Material and Methods the tuning was fixed to $r \geq 0.85$. Decoding the tCP-sequences into NTs (using Table 1) allowed us to find the corresponding conserved NT-stretches. This conserved stretches would be suitable for functional studies, elucidate new genetic structures or assist in mutational studies, among others.

tCP structures partially conserved can also be detected in certain stretches of the ortholog CYP1A2 (Fig. 2) as it occurs, in particular, with tCPs < TCG > and < GT > between 100 and 900 bp and 0–700 bp, respectively. The conservation of these short fragments does not significantly increase the correlation of the tCP-profile when the entire sequences were aligned. We think, however, that the method proposed here is suitable for evolutionary analysis of these partial coincidences endowing it with a great analytical power by either tuning the cut-off or considering only such fragments in the study. We are conscious that the analysis of partially conserved structures could have a great potential to extract evolutionary information from the alignment. In summary, and in order to reinforce the concept of *tCP*, as an informative evolutionary parameter, we stress the point that the tCP-sequence analysis of orthologs reveals the existence of common inner compositional structures, not described before as far as we know.

It might be intuitive that in general, a high correlation must exist between the number of conserved tCPs and the percent identity. However, this is not always the case (Fig. 3C). The low correlation observed between the percent identity and the number of conserved tCPs in some orthologs from samples 1 and 2 ($r = 0.153$ and $r = 0.251$, respectively) shows that in some cases high sequence identities and a low number of conserved tCPs can coexist as the neutral theory of evolution predicts. This result is interesting because it allows identifying a fact that have been discussed (Morrison 2009, 2015) having to do with the observed percent identity between alignments of gene sequences and their connection with homology (Pearson 2013). Thus, our data could fuel the debate that, frequently, high identities in orthologs overlook inner gene structures biologically relevant.

Differences Between tCP and NT-Sequence Homologies

The concept of homology (Pearson 2013) is central to the analyses of gene sequences. In our dataset, orthology (a type of homology) is established as the selection criterion. All genes selected in this study are orthologs. In no case, we based our selection criteria on percent identity, on similarity or both. In the case considered in this manuscript, the percent identity can be considered a measure of similarity because we have selected orthologs genes (Pearson 2013). We considered that sometimes, however, high percent identity from NT-alignments could mask valuable evolutionary information.

We observed that some human–mouse orthologs in this study have high percent identity in the NT-sequence but low number of shared tCPs. For example, in the case of the human–mouse orthologs CYP1A2 or SAMD12, 3 and 4 out of 14 tCPs, respectively, are conserved in each ortholog (Fig. 2 and Online Resource 5) despite the high percent identity of NT-sequences (see Online Resource 2). In fact, 90% of orthologs from sample 1 and 74% of sample 2 share less than 6 tCPs (Fig. 4B). We have to underline here that negative-correlated tCP-profiles are more common than we might think and occur when high values of one variable tend to be associated with low values of the other. The fact of having negative-correlated profiles gives as a result that some coincidences in the sequence alignments may mainly be due to fluctuations about the mean or at random fits in many NTs of the aligned sequences. The fluctuation and random fits observed would affect the global computation of the percent identity between both ortholog NT-sequences.

A notable difference between NT-homology and tCP-homology has to do with the fact that in the NT-ortholog sequences, generated by the conserved tCP-ortholog sequences, there is some NT-mismatches (see Fig. 5) due to the redundancy of the tCP-code (Table 1). These mismatches at NT-level do not affect the integrity of the conserved tCP-structure because those changes correspond to NT-triplets that are included in some of the conserved tCPs (see Table 1). Consequently, the conserved stretch of a tCP-sequence might not be conserved in the associated NT-sequence. These mismatches at NT-level may imply, in some cases, alterations in non-synonymous codons having an important role in the fitness of the organism because they are conserved at tCP-level during speciation.

Critical Gene Structures and Evolution

How relevant is in orthologs the high percent identity of NT-sequences of non-conserved fragments as those detected by the tCP-method? In CYP1A2 and SAMD12, the high percent identity might be due to the sum of several factors: (i) to highly conserved tCP-profiles ($r \geq 0.85$) as expected; (ii) to tCP-profiles with correlations lower but near the cut-off; (iii) to non-correlated tCP-profiles far from the cut-off and finally, (iv) to negative-correlated tCP-profiles. For example, in CYP1A2 the tCPs < C> ($r = 0.737$) and < TCG> ($r = 0.722$) have correlations lower but near the cut-off. The same occurs in SAMD12 with < A> ($r = 0.720$), <GT> ($r = 0.738$), <AGC> ($r = 0.819$) and < AGT> ($r = 0.830$). Thus, in both cases, the tCPs with correlations lower but near the cut-off contribute significantly to the percent identity observed in their associated NT-sequences (Online Resource 5). It is important to take in mind that the tuning of the cut-off correlates with evolutionary time (see Fig. 6). Thus, the tuning of the cut-off would be a powerful tool to analyse the evolution of conserved structures of orthologs. Our data suggest that mutations in conserved tCP-sequences could be critical for the fitness of the organisms since a simple substitution in the NT-sequence involves 1 to 3 changes in the tCP-sequence.

In summary, the high sequence homology detected in the structure of orthologs interspersed all along the gene sequence are frequently masked by a variety of NT coincidences attributable to local fits, to low correlated tCPs or even to fluctuations about the mean in the case of negative-correlated tCPs (data not shown). The highly conserved regions free of the noise of spurious identities, as indicated above, could be potential immunological and/or pharmacologic targets for treatment and control. The conservation of these highly conserved motifs suggests that they have a biological significance from the functional standpoint. The identification of such common and highly conserved structures in mammals allows us introducing the concept of *evolutionary critical gene structures* whose change would seriously affect the fitness or gene expression.

Funding This work was funded by a program of the Instituto de Salud Carlos III-Redes Temáticas de Investigación Cooperativa en Salud (ISCIII-RETIC RD06/0021/0008 program) and Laboratorios LETI. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. An institutional grant from Fundación Ramón Areces is also acknowledged.

References

Aldrich J (1995) Correlations genuine and spurious in pearson and yule. Stat Sci 10:364–376

Amit M et al (2012) Differential GC content between exons and introns establishes distinct strategies of splice-site recognition. Cell Rep 1:543–556. https://doi.org/10.1016/j.celrep.2012.03.013

Arnold J, Cuticchia AJ, Newsome DA, Jennings WW, Ivarie R (1988) Mono-through hexanucleotide composition of the sense strand of yeast DNA: a Markov chain analysis. Nucleic Acids Res 16:7145–7158

Bhangale TR, Rieder MJ, Livingston RJ, Nickerson DA (2005) Comprehensive identification and characterization of diallelic insertion-deletion polymorphisms in 330 human candidate genes. Hum Mol Genet 14:59–69. https://doi.org/10.1093/hmg/ddi006

Blanchette M et al (2004) Aligning multiple genomic sequences with the threaded blockset aligner. Genome Res 14:708–715. https://doi.org/10.1101/gr.1933104

Bray N, Pachter L (2004) MAVID: constrained ancestral alignment of multiple sequences. Genome Res 14:693–699. https://doi.org/10.1101/gr.1960404

Comeron JM (2001) What controls the length of noncoding DNA? Curr Opin Genet Dev 11:652–659

Costas J, Pereira PS, Vieira CP, Pinho S, Vieira J, Casares F (2004) Dynamics and function of intron sequences of the wingless gene during the evolution of the Drosophila genus. Evol Dev 6:325–335. https://doi.org/10.1111/j.1525-142X.2004.04040.x

Dai Q, Liu XQ, Wang TM, Vukicevic D (2007) Linear regression model of DNA sequences and its application. J Comput Chem 28:1434–1445. https://doi.org/10.1002/jcc.20556

Frazer KA et al (2001) Evolutionarily conserved sequences on human chromosome 21. Genome Res 11:1651–1659. https://doi.org/10.1101/gr.198201

Fuertes MA, Perez JM, Zuckerkandl E, Alonso C (2011) Introns form compositional clusters in parallel with the compositional clusters of the coding sequences to which they pertain. J Mol Evol 72:1–13. https://doi.org/10.1007/s00239-010-9411-6

Fuertes MA, Rodrigo JR, Alonso C (2016a) Do intron and coding sequences of some human–mouse orthologs evolve as a single unit? J Mol Evol 82:247–250. https://doi.org/10.1007/s00239-016-9746-8

Fuertes MA, Rodrigo JR, Alonso C (2016b) A method for the annotation of functional similarities of coding DNA sequences: the case of a populated cluster of transmembrane proteins. J Mol Evol 84:29–38. https://doi.org/10.1007/s00239-016-9763-7

Fuertes MA, Rodrigo JR, Zuckerkandl E, Alonso C (2016c) The chromosomal and functional clustering of markedly divergent human–mouse orthologs run parallel to their compositional features. J DNA RNA Res 1:1–31

Gates MA (1986) A simple way to look at. DNA J Theor Biol 119:319–328

Gazave E, Marques-Bonet T, Fernando O, Charlesworth B, Navarro A (2007) Patterns and rates of intron divergence between humans and chimpanzees. Genome Biol 8:R21. https://doi.org/10.1186/gb-2007-8-2-r21

Gelfman S et al (2012) Changes in exon-intron structure during vertebrate evolution affect the splicing pattern of exons. Genome Res 22:35–50. https://doi.org/10.1101/gr.119834.110

Gilbert W (1978) Why genes in pieces? Nature 271:501

Gingeras TR (2009) Implications of chimaeric non-co-linear transcripts. Nature 461:206–211. https://doi.org/10.1038/nature08452

Hardison RC, Oeltjen J, Miller W (1997) Long human–mouse sequence alignments reveal novel regulatory elements: a reason to sequence the mouse genome. Genome Res 7:959–966

Hong CC, Tang BK, Hammond GL, Tritchler D, Yaffe M, Boyd NF (2004) Cytochrome P450 1A2 (CYP1A2) activity and risk factors for breast cancer: a cross-sectional study. Breast Cancer Res 6:R352–R365. https://doi.org/10.1186/bcr798

Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature 423:241–254. https://doi.org/10.1038/nature01644

Keren H, Lev-Maor G, Ast G (2010) Alternative splicing and evolution: diversification, exon definition and function. Nat Rev Genet 11:345–355. https://doi.org/10.1038/nrg2776

King MC, Wilson AC (1975) Evolution at two levels in humans and chimpanzees. Science 188:107–116

Kruskal JB (1983) An overview of squence comparison. Time warps, string edits and macromolecules: the theory and practice of sequence comparison, Addison Wesley edn. CSLI Publications, Stanford University

Kumar S, Stecher G, Tamura K (2016) MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol Biol Evol 33:1870–1874. https://doi.org/10.1093/molbev/msw054

Leong PM, Morgenthaler S (1995) Random walk and gap plots of DNA sequences. Comput Appl Biosci 11:503–507

Louie E, Ott J, Majewski J (2003) Nucleotide frequency variation across human genes. Genome Res 13:2594–2601. https://doi.org/10.1101/gr.1317703

Lunter G (2007) Probabilistic whole-genome alignments reveal high indel rates in the human and mouse genomes. Bioinformatics 23:i289–296 https://doi.org/10.1093/bioinformatics/btm185

Majewski J, Ott J (2002) Distribution and characterization of regulatory elements in the human genome. Genome Res 12:1827–1836. https://doi.org/10.1101/gr.606402

Mattick JS, Gagen MJ (2001) The evolution of controlled multitasked gene networks: the role of introns and other noncoding RNAs in the development of complex organisms. Mol Biol Evol 18:1611–1630

Mills RE et al (2011) Natural genetic variation caused by small insertions and deletions in the human genome. Genome Res 21:830–839. https://doi.org/10.1101/gr.115907.110

Morrison DA (2009) Why would phylogeneticists ignore computerized sequence alignment? Syst Biol 58:150–158. https://doi.org/10.1093/sysbio/syp009

Morrison DA (2015) Is sequence alignment an art or a science? Syst Bot 40:14–26. https://doi.org/10.1600/036364415X686305

Mullan LJ, Bleasby AJ (2002) Short EMBOSS User Guide. Eur Mol Biol Open Softw Suite Brief Bioinform 3:92–94

Nandy A (2009) Empirical relationship between intra-purine and intra-pyrimidine differences in conserved gene sequences. PLoS ONE 4:e6829. https://doi.org/10.1371/journal.pone.0006829

Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48:443–453

Notebaart RA, Huynen MA, Teusink B, Siezen RJ, Snel B (2005) Correlation between sequence conservation and the genomic context after gene duplication. Nucleic Acids Res 33:6164–6171. https://doi.org/10.1093/nar/gki913

Olson SA (2002) EMBOSS opens up sequence analysis. Eur Mol Biol Open Softw Suite Brief Bioinform 3:87–91

Parker SC, Tullius TD (2011) DNA shape, genetic codes, and evolution. Curr Opin Struct Biol 21:342–347. https://doi.org/10.1016/j.sbi.2011.03.002

Parmley JL, Urrutia AO, Potrzebowski L, Kaessmann H, Hurst LD (2007) Splicing and the evolution of proteins in mammals. PLoS Biol 5:e14. https://doi.org/10.1371/journal.pbio.0050014

Pearson H (2006) Genetic information: codes and enigmas. Nature 444:259–261. https://doi.org/10.1038/444259a

Pearson WR (2013) An introduction to sequence similarity ("homology") searching. Curr Protoc Bioinform. https://doi.org/10.1002/0471250953.bi0301s42

Robart AR, Zimmerly S (2005) Group II intron retroelements: function and diversity. Cytogenet Genome Res 110:589–597. https://doi.org/10.1159/000084992

Robart AR, Seo W, Zimmerly S (2007) Insertion of group II intron retroelements after intrinsic transcriptional terminators. Proc Natl Acad Sci USA 104:6620–6625. https://doi.org/10.1073/pnas.0700561104

Rogozin IB et al (2002) Connected gene neighborhoods in prokaryotic genomes. Nucleic Acids Res 30:2212–2223

Rogozin IB, Sverdlov AV, Babenko VN, Koonin EV (2005) Analysis of evolution of exon-intron structure of eukaryotic genes. Briefings Bioinf 6:118–134

Roy A, Raychaudhury C, Nandy A (1988) Novel techniques of graphical representation and analysis of DNA sequences—a review. J Biosci 23:55–71

Schwartz S, Meshorer E, Ast G (2009) Chromatin organization marks exon-intron structure. Nat Struct Mol Biol 16:990–995. https://doi.org/10.1038/nsmb.1659

Sneath PHA, Sokal RR (1973) Numerical taxonomy. The principles and practice of numerical calssification. A series of books in biology. W. H. Freeman and Company, San Francisco

Sueoka N (1962) On the genetic basis of variation and heterogeneity of DNA base composition. Proc Natl Acad Sci USA 48:582–592

Takeda M (2012) How is the biological information arranged in genome? Am J Mol Biol 2:171–186

Tamura K, Nei M, Kumar S (2004) Prospects for inferring very large phylogenies by using the neighbor-joining method. Proc Natl Acad Sci USA 101:11030–11035. https://doi.org/10.1073/pnas.0404206101

Trifonov EN (2011) Thirty years of multiple sequence codes. Genomics Proteom Bioinform 9:1–6. https://doi.org/10.1016/S1672-0229(11)60001-6

Wang C, Typas MA, Butt TM (2005) Phylogenetic and exon-intron structure analysis of fungal subtilisins: support for a mixed model of intron evolution. J Mol Evol 60:238–246. https://doi.org/10.1007/s00239-004-0147-z

Weber JL, David D, Heil J, Fan Y, Zhao C, Marth G (2002) Human diallelic insertion/deletion polymorphisms. Am J Hum Genet 71:854–862. https://doi.org/10.1086/342727

Woolfe A et al (2005) Highly conserved non-coding sequences are associated with vertebrate development. PLoS Biol 3:e7. https://doi.org/10.1371/journal.pbio.0030007

Yates A et al (2016) Ensembl 2016. Nucleic Acids Res 44:D710-716 https://doi.org/10.1093/nar/gkv1157

Yue F et al (2014) A comparative encyclopedia of DNA elements in the mouse genome. Nature 515:355–364. https://doi.org/10.1038/nature13992

Zhao Q et al (2009) Transcriptome-guided characterization of genomic rearrangements in a breast cancer cell line. Proc Natl Acad Sci USA 106:1886–1891. https://doi.org/10.1073/pnas.0812945106

Zhu L, Zhang Y, Zhang W, Yang S, Chen JQ, Tian D (2009) Patterns of exon-intron architecture variation of genes in eukaryotic genomes. BMC Genom 10:47. https://doi.org/10.1186/1471-2164-10-47