doi:10.1016/j.jmb.2006.05.024 J. Mol. Biol. (2006) 360, 258–266

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

The Role of Introns in Repeat Protein Gene Formation

Timothy O. Street, George D. Rose and Doug Barrick*

T.C. Jenkins Department of Biophysics, Johns Hopkins University, 3400 N. Charles Street, Baltimore, MD 21218 USA

Genes composed of tandem repetitive sequence motifs are abundant in nature and are enriched in eukaryotes. To investigate repeat protein gene formation mechanisms, we have conducted a large-scale analysis of their introns and exons. We find that a wide variety of repeat motifs exhibit a striking conservation of intron position and phase, and are composed of exons that encode one or two complete repeats. These results suggest a simple model of repeat protein gene formation from local duplications. This model is corroborated by amino acid sequence similarity patterns among neighboring repeats from various repeat protein genes. The distribution of one- and two-repeat exons indicates that intron-facilitated repeat motif duplication, in which the start and end points of duplication are located in consecutive intronic regions, significantly exceeds intron-independent duplication. These results suggest that introns have contributed to the greater abundance of repeat protein genes in eukaryotic *versus* prokaryotic organisms, a conclusion that is supported by taxonomic analysis.

*Corresponding author

Keywords: repeat protein; introns; exons; geneformation; local duplication

Introduction

A central challenge in evolutionary biology is to deduce gene formation mechanisms that can account for the striking variations among protein tertiary structures. One plausible mechanism involves assembling novel genes from short gene fragments encoding diverse structures. The ''exon theory of genes''[1,2] is an example of such a modular model for protein evolution. According to this theory, genes encoding diverse protein structures were constructed from short coding modules bracketed by introns. One attractive feature of this model is that exon modules can easily be combined through large intronic regions without having to precisely join coding segments.[3] A crucial test of any modular model of protein construction, such as the exon theory of genes, involves decomposing modern protein structures into their original modules, thus deriving a protein structure basis set.[4] Although intron positions have been correlated to linker regions between compact structural units in globular proteins,[5] identifying ancestral genetic modules proposed from the exon theory of genes in modern protein structures remains a challenge, owing in part to the structural complexity of globular proteins.

In contrast, repeat proteins have much simpler structures that predispose them to a modular description. Repeat proteins contain tandem arrays of repeat motifs that are typically 20 to 60 residues in length. In some cases, repeat motif structures are packed together closely (for example, ankyrin repeats (ANKs),[6] leucine-rich repeats (LRRs),[7] and pumilio repeats (PUMs)[8]), and the structural stabilities of neighboring repeats are highly interdependent.[9,10] In other cases, repeat motifs are structurally autonomous (for example, sushi repeats (SUs),[11] epidermal growth factor repeats (EGFs),[12] and zinc finger repeats (ZFs)[13]). Genes encoding arrays of closely packed and structurally autonomous repeat motifs are abundant in nature, particularly in eukaryotic organisms.[14]

Despite their abundance, little is known about the formation mechanisms of repeat protein genes. Sequencing reports of individual repeat protein genes have shown regular intron spacing in some (but not all) of these genes, providing anecdotal support for intron-mediated gene formation on a case-by-case basis.[15–17] Here, we analyze intron and exon characteristics from a wide variety of repeat

Abbreviations used: LRR, leucine rich repeat; ANK, ankyrin; PUM, pumilio; SU, sushi; EGF, epidermal growth factor; LDL, low-density lipoprotein receptor; ZF, zinc finger; HMM, hidden Markov model.

E-mail address of the corresponding author: barrick@jhu.edu

protein genes to investigate their formation mechanisms in a statistically significant way. We find a distribution of intron positions within eukaryotic repeat protein genes that is consistent with substantial intron-facilitated gene formation. Although our results do not address the role of introns in early protein evolution,[18] our analysis suggests that spliceosomal introns may have facilitated the expansion of repeat protein genes in eukaryotes.

Results and Discussion

Many repeat motifs show highly conserved intron positions and phases

In this study, we examined intron and exon characteristics associated with LRR (the typical and plant-specific subtypes, LRR_{typ+ps}, and the ribonuclease inhibitor-like subtype, LRR_{ri}), ANK, PUM, SU, EGF, low-density lipoprotein receptor (LDL), ZF (the C2H2 family), tetratrico peptide repeat (TPR), WD-40, armadillo (ARM), spectrin (SPEC), kelch (KEL), EF-hand (EF), and HEAT motifs. Intron positions were mapped to repeat motif sequences on genes in the ExInt database[19] by using hidden Markov models (HMMs).[20] The ExInt database contains the amino acid sequences from 230,000 genes and associated information about exon lengths, intron positions, and intron phases (intron boundaries situated before the first, second, and third base in a codon are referred to as phase 0, 1, and 2, respectively). If introns were inserted independently into full-length repeat protein genes, there should be no position or phase correlation between neighboring introns. If instead, repeat protein genes were constructed from locally duplicated intron/exon building blocks, introns should be positioned with regular spacing and should have a common phase.

Consistent with this model of repeat protein gene expansion from smaller intron/exon building blocks, we find that several repeat motifs (LRR_{typ+ps}, LRR_{ri}, ANK, PUM, SU, EGF, LDL, and ZF) have highly conserved intron positions (Figure 1). Moreover, introns clustered at conserved positions show a significant phase bias (represented by the histogram bar colors; Figure 1). Although intron positions for PUM genes are located at two adjacent sequence positions with approximately equal frequency (Figure 1(d)), within individual PUM genes these differing intron positions are found in consecutive runs (as opposed to being randomly distributed, data not shown), consistent with local duplication. In contrast, intron positions from WD-40, TPR, ARM, EF, SPEC, KEL, and HEAT motifs are evenly distributed throughout the consensus sequence, and show no obvious phase bias (data not shown). The lack of intron position or phase conservation in these repeat motifs suggests that their introns accumulated through random insertion in genes that were not formed from duplicated intron/exon building blocks, but were formed from intron-independent repeat motif duplication. Alternately, these intron-free repeat protein genes may have lost their introns through genomic reinsertion of processed repeat protein mRNAs.

Conserved introns typically bracket one or two complete repeat motifs

To further investigate the origins of repeat protein genes with conserved intron positions, we examined the distribution of exon lengths from repeat protein genes. These distributions show clear peaks corresponding to single repeat motifs (see asterisks; Figure 2). The exon length distributions from LRR_{typ+ps}, LRR_{ri}, ANK, and PUM genes show additional peaks corresponding to two complete repeat motifs (also denoted by asterisks; the PUM peaks at 48 and 65 residues result from incomplete repeats located at the ends of PUM arrays; see Supplementary Table 1).

Although the exon length distributions show clear peaks at integer-repeat lengths, off-peak exon lengths are also common, especially in ANK repeats and in the structurally autonomous repeats (SU, EGF, LDL, and ZF). We find that the integer-length and the off-peak exons are strongly segregated among repeat protein genes: individual genes are either composed solely of integer-length exons or solely of random-length exons. This is illustrated in Figure 2, where the black shading identifies all the exons from repeat protein genes that contain at least two integer-length exons. The black histogram bars map almost exclusively to integer exon lengths, whereas the remaining (unshaded) bars define a uniform distribution of exon lengths. These results divide repeat protein genes into two classes: those with random-length exons, which likely accumulated introns through random insertion in full length repeat arrays, and those composed exclusively of integer-length exons which formed from local duplication of intron/exon units. The structurally autonomous repeat protein genes that appear to be formed from local duplication (Figure 2(e)–(h); shaded bars) have a broader distribution of exon lengths than the closely packed repeat proteins, which is consistent with a greater tolerance of the former group to small insertion/deletion events due to their lack of inter-repeat coupling.

Exon lengths suggest specific gene formation mechanisms

The exon length distributions from LRR_{typ+ps}, LRR_{ri}, ANK, PUM, SU, EGF, LDL, and ZF genes suggest a common formation mechanism involving frequent intron-facilitated duplication, interspersed by occasional intron-independent exon duplication. Repeat protein genes formed exclusively by intron-facilitated duplication would contain all one-repeat exons (Figure 3(a)), as is seen for the SU repeats in the *SELE* gene; this pattern is easily visualized by