Role of Low-Complexity Sequences in the Formation of Novel Protein Coding Sequences

Macarena Toll-Riera,1 Núria Radó-Trilla,1 Florian Martys,1 and M. Mar Albà*1,2

1Evolutionary Genomics Group, Research Programme in Biomedical Informatics, Universitat Pompeu Fabra (UPF)—Institute Municipal d’Investigació Mèdica (IMIM), Barcelona, Spain
2Catalan Institution for Research and Advanced Studies (ICREA), Barcelona, Spain
*Corresponding author: E-mail: malba@imim.es.
Associate editor: Helen Piontkivska

Abstract

Low-complexity sequences are extremely abundant in eukaryotic proteins for reasons that remain unclear. One hypothesis is that they contribute to the formation of novel coding sequences, facilitating the generation of novel protein functions. Here, we test this hypothesis by examining the content of low-complexity sequences in proteins of different age. We show that recently emerged proteins contain more low-complexity sequences than older proteins and that these sequences often form functional domains. These data are consistent with the idea that low-complexity sequences may play a key role in the emergence of novel genes.

Key words: low-complexity sequence, gene age, protein domain evolution.

Low-complexity regions (LCRs) are amino acid sequences that contain repeats of single amino acids or short amino acid motifs. They are extremely abundant in eukaryotic proteins (Green and Wang 1994; Golding 1999; Marcotte et al. 1999). In fact, the majority of proteins from a wide range of eukaryotic species show a significant tendency toward being more repetitive than expected given their amino acid composition (Alba, Tompa, et al. 2007). Many LCRs are highly unstable due to the action of replication slippage and recombination (Ellegren 2004), and the uncontrolled expansion of short sequence motifs causes several human diseases, including Huntington’s disease and other neurodegenerative disorders (Gatchel and Zoghbi 2005), as well as a number of developmental diseases (Brown and Brown 2004). The abundance of LCRs seems paradoxical given their high pathogenic potential. One hypothesis to explain their persistence is that they increase phenotypic variation within populations, facilitating adaptation (Kashi and King 2006). While many LCRs are of unknown function, there are examples of LCRs playing various functional roles, including the modulation of protein–protein interactions (Xiao and Jeang 1998), protein–nucleic acid interactions (Shen et al. 2004), and protein subcellular localization (Salichs et al. 2009). Expansion or contraction of LCRs can therefore potentially impact protein function.

An alternative hypothesis to explain the abundance of LCRs is that they facilitate the formation of novel coding sequences (Green and Wang 1994). Analysis of human family genotypes has shown that, when the repeats are short, they are more likely to expand than to contract (Xu et al. 2000), which favors the extension of initially short “seed” repeats into longer repeats. Accumulation of subsequent mutations may lead to the emergence of new useful protein functions. A more radical idea is that repetitive sequences are important for the generation of completely novel coding sequences. In the early 80s, Ohno and Epplen proposed that the first protein encoding sequences were probably highly repetitive, as expansion of repeated tracts was more likely to yield long polypeptide chains with no interrupting codons than when sequences had a random amino acid composition (Ohno and Epplen 1983; Ohno 1984). Inspired by this idea, we decided to test if recently emerged genes contain more LCRs than older genes. Although there have been some observations that point in this direction (Nishizawa et al. 1999; Alba and Castresana 2005), the question had not been yet examined in detail. To learn about the contribution of LCRs to protein function, we also quantified how many LCRs were located in already described protein domains and how many were located in regions not corresponding to domains.

To study the correlation of gene age and LCR content, we obtained three groups of human proteins that arose at different periods “Mammalian” (−300 to 100 Ma), “Vertebrate” (−500 to 300 Ma) and “Old” (−500 Ma). For proteins containing hits to Pfam protein domains (Finn et al. 2008), the classification was based on the phylogenetic distribution of such domains, as determined by domain-specific hidden Markov model searches in different eukaryotic proteomes (see Materials and Methods). For proteins not containing hits to Pfam protein domains, the classification was based on database sequence similarity searches using BlastP (Altschul et al. 1997). We searched for LCRs in all classified proteins using the SEG algorithm, which identifies regions of biased composition enriched in one or a few amino acids (Wootton and Federhen 1996). The majority of proteins from each of the age classes contained at least one LCR (83.1% Old, 83.7% Vertebrate, and 87% Mammalian), confirming the strong pervasiveness of these sequences. However, in younger proteins, LCRs occupied a significantly larger fraction of the sequence than in older
proteins (fig. 1). On average, the LCR content of Mammalian proteins was double the LCR content of proteins classified as Old (the values for all proteins can be found in supplementary file 1, Supplementary Material online). This relationship was maintained in proteins containing known domains and in proteins lacking them (supplementary table 1, Supplementary Material online), and consistent results were obtained using a different algorithm to measure sequence repetitiveness, SIMPLE (Alba, Laskowski, et al. 2002) (supplementary table 2, Supplementary Material online).

We next examined whether the composition of LCRs varied depending on the age of the protein. We did not find any significant differences except for cysteine, which was more abundant in the LCRs of older proteins (supplementary figure 1, Supplementary Material online). Therefore, the differences in low-complexity sequence content between proteins of different age were essentially quantitative, not qualitative. The overrepresented amino acids in LCRs were similar to those forming the bulk of human amino acid tandem repeats (Karlin et al. 2002), i.e., proline, alanine, serine, glycine, leucine, and glutamic acid. Several authors have reported a positive relationship between coding sequence GC content and repeat abundance (Nakachi et al. 1997; Alba and Guigo 2004). This prompted us to test whether the higher LCR abundance in novel genes could be explained by differences in GC content. We also found that coding sequences with a higher GC content tended to be more enriched in LCRs, but this did not alter the reported relationship between gene age and LCR content (supplementary figure 2, Supplementary Material online). Many LCRs are predicted to correspond to disordered structures (Simon and Hancock 2009). We compared the PDB structures matching proteins with an LCR content >25% (26 cases) and proteins with no LCR content at all (1,873 cases), focusing on the ratio between the interatomic N-terminus to C-terminus distance (d, in angstroms) and the number of residues (N) as a measure of structural compactness. The average d/N in LCR-containing proteins was 0.78 compared with 0.39 for proteins with no LCRs (significantly different at P < 10^-5, supplementary figure 3, Supplementary Material online), denoting that the former tend to be more elongated.

Protein domains are defined as independent sequence modules associated with a specific function and shared by different proteins. In general, repetitive regions are not expected to be abundant in protein domains, as their high mutability may disrupt protein structure and function. The exception is when repetitive sequences are directly involved in function. For proteins containing already described domains in our dataset, we examined which proportion of the domain region(s) and which proportion of the nondomain region(s) was occupied by LCRs. The analysis clearly showed that younger proteins contained comparatively more LCRs in protein domains than older proteins (table 1). Whereas in Old proteins, LCRs were strongly depleted from domains when compared with nondomain regions (ratio nondomain to domain 3.4), in Vertebrate proteins, the depletion was lower (ratio 2.14), and in Mammalian proteins, the underrepresentation of LCRs in domains was only marginal (ratio 1.3). There is evidence that recently evolved genes are subject to lower selective constraints than older ones (Alba and Castreres 2005; Cai and Petrov 2010) and thus are expected to tolerate LCRs better. However, the fact that LCRs are also very abundant in regions classified as protein domains in Pfam strongly suggest that they often contribute to novel protein functions.

One excellent example of a recently evolved repetitive protein family is caseins, the most abundant proteins in milk. Caseins evolved in the lineage leading to mammals and can be subdivided into Ca-sensitive and Ca-insensitive caseins. The former contain phosphorylated serine clusters that are used to bind calcium. Both types of caseins associate through LCR regions rich in proline and glutamine forming micelles, and their high calcium content is essential for the healthy growth of bone and teeth in neonates. In fact, caseins are evolutionary related to vertebrate tooth proteins, which also utilize phosphorylated serine residues for the binding of calcium (Kawasaki et al. 2011). Another interesting example is the case of proteins forming the cornified cell envelope of keratinocytes, including cornifins and involucrin. These proteins contain short motif repeats in which glutamic acid residues act as the substrate of
a transglutaminase cross-linking reaction, which results in the formation of a thick layer of cross-linked proteins (Eckert and Green 1986).

In summary, we have shown that recently formed proteins contain more LCRs than older proteins, which is consistent with the idea that they contribute to the formation of novel genes. In addition, LCRs are well represented in the domains of novel proteins, indicating that they are often functionally relevant.

Materials and Methods

We obtained 14,784 human proteins corresponding to 1:1 orthologous human and mouse genes from Ensembl (Hubbard et al. 2009). When there was more than one single protein per gene, that corresponding to the longest transcript was extracted. We used hmmpfam (Eddy 1998) to identify known protein domains. This program, which is part of the HMMER package, searches a sequence against a library of hidden Markov models obtained from domain alignments. We employed the library Pfam_ls (version 23), and an E-value cutoff of $10^{-5}$. We identified 11,574 human proteins that contain at least one protein domain. Next, the age of a protein was defined as the age of the oldest domain it contained, following the results from domain searches in 15 additional proteomes, as follows. Mammalian protein domains were defined as those only present in humans and in at least one other mammalian species (out of mouse, rat, or cow) but absent from the rest of species. Similarly, Vertebrate protein domains were found in at least one mammalian and one nonmammalian vertebrate (chicken, frog, fugu, or zebra fish) but absent from nonvertebrate species. Finally, Old protein domains were found in at least one mammalian, one nonmammalian vertebrate, and one nonvertebrate species (mosquito, fly, worm, vane tunicate, budding yeast, fission yeast, rice, and thale cress). Proteins without domains were dated using BlastP sequence similarity searches against the proteomes mentioned above (Altschul et al. 1997). We considered that a hit with an E-value of $10^{-4}$ was indicative of the presence of a homolog in the proteome. PDB structures were mapped into human sequences using BlastP, and only those hits with at least 99% identity were kept. Interatomic distances between amino acid alpha carbons (Cα) were calculated using a Python module. We identified LCRs in proteins with SEG using default parameters (Wootton and Federhen 1996). This yielded low-complexity sequences with an average length of 18.26 amino acids. Parameters in SIMPLE (Alba, Laskowski, et al. 2002) were a score of 1 for repeats of size 1–4 amino acids and a window size of 12. We examined amino acid compositional biases of LCRs by calculating the log2 ratio of observed versus expected amino acid frequencies. Here, observed corresponded to the relative frequency of the amino acid in the LCRs and expected the relative frequency of the amino acid in all proteins classified in that age group. We measured the GC content of coding sequences encoding proteins of different age as the percentage of G + C over the four nucleotide types. We developed specific Perl scripts to perform most tasks.

Supplementary Material

Supplementary file 1, tables 1 and 2, and figures 1–3 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank Jaime Bonet for providing a script to calculate interatomic distances from PDB files and for his advice on PDB files. We also thank Steve Laurie for English revision. We received financial support from Ministerio de Educación (FPU to M.T.-R.), Fundación Javier Lamas (N.R.-T.), Ministerio de Innovación y Tecnología (BIO2009-08160), and Institución Catalana de Recerca i Estudis Avançats (M.M.A.).

References


