Supplementary data.

Role of low-complexity sequences in the formation of novel protein coding sequences.

Toll-Riera, M., Radó-Trilla, N., Martys, F., Albà, M.M.

TABLE 1

	Proteins with known domains				Proteins with no known domains			
	N	Average	Median	SD	N	Average	Median	SD
Old	11,039	9.14	6.83	9	1,816	9.41	7.31	8.9
Vertebrate	473	12.89	9.93	13.36	851	13.35	10.77	11.71
Mammalian	62	18.61	13.55	16.77	358	16.62	13.82	13.31
All	*11,580*	*9.34*	*6.94*	*9.33*	*3,204*	*11.76*	*8.58*	*12.04*

Table 1. Percentage of the protein occupied by low-complexity sequences in proteins of different age. Low-complexity sequences were identified with the SEG program (see Methods in main manuscript file). Differences between all pairs of groups were highly significant (Mann-Whitney-Wilcoxon test, $p<10^{-3}$).

TABLE 2

	Proteins with domains				Proteins with no domains			
	N	Aver	Median	SD	N	Aver	Median	SD
Old	11,039	1.913	1.788	0.518	1,816	1.979	1.875	0.508
Vertebrate	473	2.073	1.91	0.679	851	2.139	2.004	0.596
Mammalian	62	2.604	2.122	1.408	358	2.283	2.121	0.710
All	*11,580*	*1.923*	*1.794*	*0.537*	*3,204*	*2.060*	*1.928*	*0.571*

Table 2. Simplicity values in proteins of different age. Simplicity is a measure of the local repetitiveness of an amino acid and it was calculated with the SIMPLE program (see Methods in main manuscript text). Differences between Old and Vertebrate and Old and Mammalian were highly significant (Mann-Whitney-Wilcoxon test, $p<10^{-3}$).

FIGURE 1

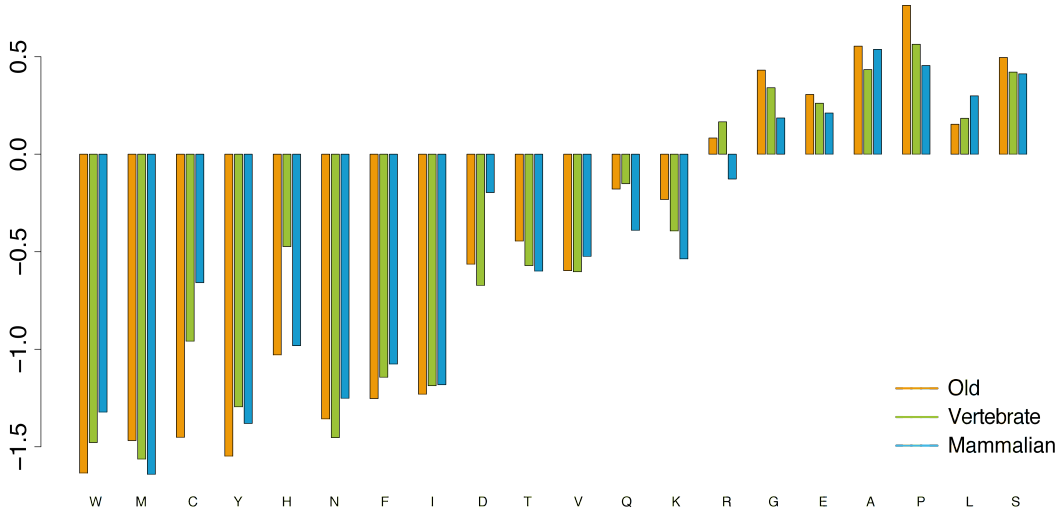

Figure 1. Over- and under-represented amino acids in low-complexity regions (LCRs) of human proteins of different age. The log2 ratio of observed versus expected amino acid frequency is shown. The observed frequency was the frequency of the amino acid within LCRs. The expected frequency was obtained from the whole protein sequence dataset. Only cysteine showed significantly different frequency depending on the age group (Test of Equal or Given Proportions; $p<10^{-5}$).

FIGURE 2

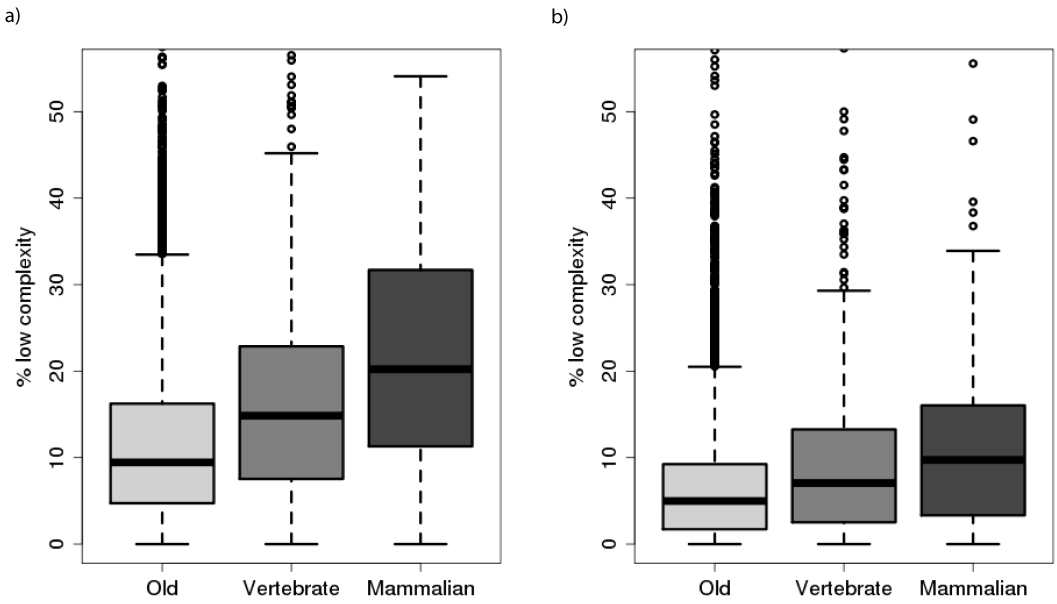

Figure 2. Effect of coding sequence GC content of the low-complexity sequence content. Box-plots of the percentage of the protein composed of low-complexity regions (LCRs), for proteins of different age. For each age group, we divided the coding sequences in two halves, the first one corresponding to high GC content and the second one to low GC content. a) Proteins encoded by coding sequences with high GC content (median values OLD 60%, VERTEBRATE 62%, MAMMALIAN 63%). b) Proteins encoded by coding sequences with low GC content (median values OLD 46%, VERTEBRATE 47%, MAMMALIAN 47%). The proteins in a) show more elevated LCR content than proteins in b). However, in both cases younger proteins show higher LCR content than older proteins.

FIGURE 3

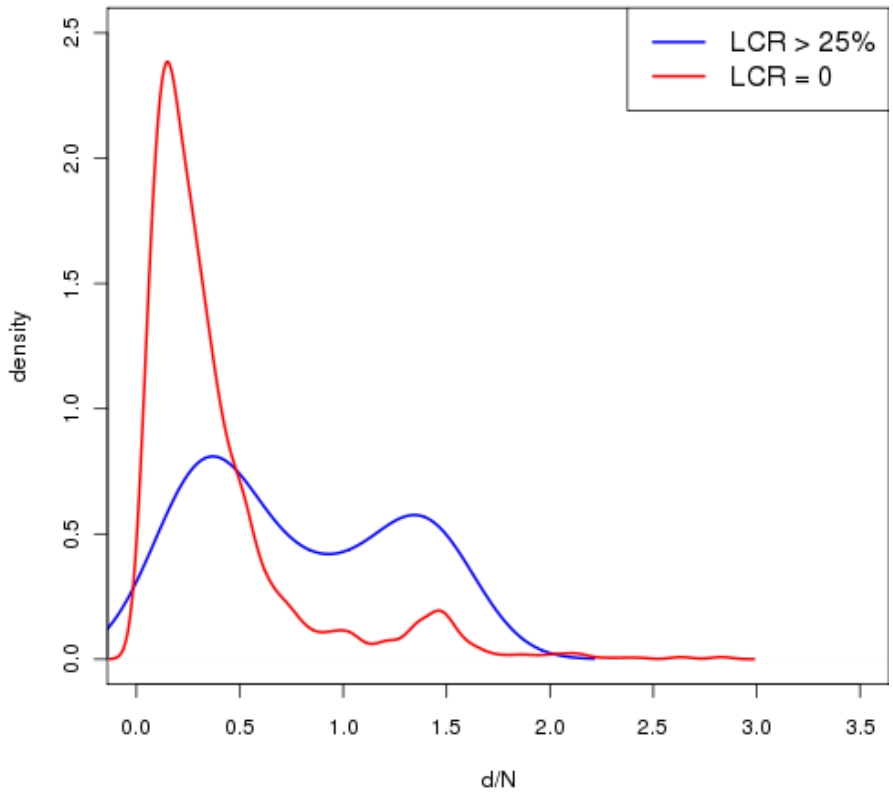

Figure 3. Protein structures containing low-complexity regions (LCRs) are less compact than protein structures with no LCRs. d: distance between the N-terminal and the C-terminal residues in the PDB entry (in angstroms). N: number of residues in the PDB entry. Higher d/N values indicate that the protein is more extended (less globular). LCR>25% corresponds to 26 protein structures corresponding to sequences with more than 25% LCR content (d/N average 0.78, median 0.64, standard deviation 0.48). LCR=0 corresponds to 1,873 protein structures with no LCR content at all (d/N average 0.39, median 0.26, standard deviation 0.4). The d/N distribution was significantly different between the two groups by a Mann-Whitney-Wilcoxon test ($p<10^{-5}$).