Evolutionary forces on different flavors of intrinsic disorder in the human proteome

Sergio Forcelloni[a], Andrea Giansanti[a,b]

[a] Physics Department of ‘La Sapienza’, Rome, 00185, Italy.
[b] Istituto Nazionale di Fisica Nucleare, INFN, Rome, 00185, Italy.
e-mail: Sergio.forcelloni@uniroma1.it

Keywords: intrinsically disordered proteins, codon usage bias, natural selection, mutational bias.

The codon usage bias (CUB) is the well-known phenomenon of an unequal use of synonymous codons in coding DNA [1]. These patterns reflect the action of weak selection working at the molecular level and allows quantifying both individual and combined effects of natural selection and mutational bias on genes. The prevailing hypothesis to explain the origin of CUB is the *selection-mutation-drift theory* [2], according to which it results from a balance between the natural selection and mutational bias. Natural selection can influence CUB of specific genes or even specific codon positions that require fine control (e.g., at the level of translation and co-translational protein folding). On the contrary, mutational bias tends to accumulate mutations asymmetrically on the whole genome of an organism. Although natural selection has been widely documented in prokaryotes and unicellular eukaryotes [3][4], it is not yet clear whether it does not exist or it is too weak to be detectable in human genomes [5], where sequence nucleotide patterns are embedded into large regions of low or high GC content (the so-called *isochores*) [6]. Many studies have provided evidence for natural selection on CUB in highly expressed and housekeeping genes in human [5][7]. However, it remains not well explored whether evolutionary pressures act differently on human genes depending on the structural properties of the encoded proteins and this represents one of the aims of the present research. In line with a previous study, we separated the human proteome in three broad variants of proteins characterized by different structural and functional properties: i) ordered proteins (ORDPs), ii) mostly ordered proteins with long intrinsically disordered protein regions (IDPRs), and iii) intrinsically disordered proteins (IDPs). ORDPs are expected to be more under control by natural selection than IDPs because one or few mutations (even synonymous) in the genes can result in a protein that no longer folds correctly. On the contrary, IDPs are generally thought to evolve more rapidly than well-structured proteins, due to the lack of structural constraints [8][9]. Using different genetic tools, we find compelling evidence that IDPs are the variant of human proteins on which the main drivers of evolutionary change (i.e., mutational bias and natural selection) act more effectively (Fig. 1, 2, 3), corroborating their hypothesized important role for evolutionary adaptability and protein evolvability. At the same time, we speculate that IDPs have a high tolerance to mutations (both neutral and adaptive) but also a selective propensity to preserve their structural disorder, i.e. flexibility and conformational dynamics under physiological conditions. Additionally, we provide new insights about the role of IDPs in human cancer. Indeed, IDPs are characterized by an abundance of CpG sites in the sequences (Fig. 4), implying a higher susceptibility to methylation resulting in C-T transition mutations (involved in several human cancers [10]). Our results provide new insight towards general laws of protein evolution identifying the intrinsically disordered proteins as reservoirs for evolutionary innovations. Quantifying selective and mutational forces acting on human genes could be useful for future studies concerning protein de-novo design, synthetic biology, biotechnology applications, and the identification of proteins that are relevant in human genetic diseases.

[1] G. Hanson and J. Coller, Nat. Rev. Mol. Cell. Biol. 19 (2018) 20-30.
[2] M. Bulmer, Genetics 149 (1991) 897-907.
[3] T. Ikemura, J. Mol. Biol. 146 (1981) 1–21.
[4] T. Ikemura, J. Mol. Biol. 158 (1982) 573–597.
[5] L. Ma et al., Biology Direct (2014) 9:17.
[6] G. Bernardi et al., Science 228 (1985) 953-958.
[7] A. O. Urrutia and L. D. Hurst, Genome Res. 13 (2003) 2260–2264.
[8] C. J. Brown et al., Mol. Biol. Evol. 27 (2011) 609–621.
[9] A. Afanasyeva et al., Genome Res. 28 (2018) 975–982.
[10] G. P. Pfeifer, Int. J Mol. Sci. 19 (2018) 1166.

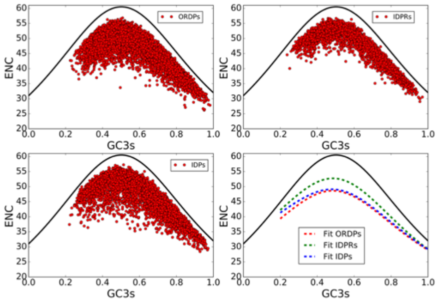

Figure 1. ENC-plots of ORDPs, IDPRs, and IDP. The solid black lines in all panels are plots of Wright’s theoretical curve corresponding to the case of no selective pressure. All distributions lie below the theoretical curve, indicating the action of natural selection on all the variants. Looking at the best-fit curves (bottom-right panel), IDPs and ORDPs appear to be subjected on average to the same balance between mutational bias and natural selection.

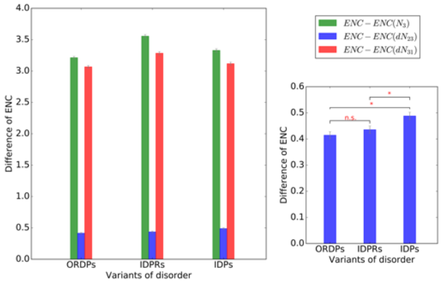

Figure 2. Residual codon bias not explained by (di)nucleotide content in 3 (N3), 2-3 (dN23), 3-1 (dN31) codon positions. ENC is the *effective number of codons*. IDPs are characterized by the highest extent of unexplainable bias that we would identify with some selective mechanisms that differentiate them from the rest of human proteome (bottom-right panel).

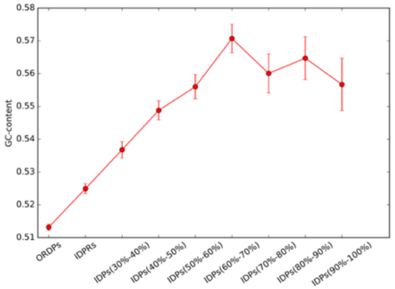

Figure 3. GC-content for ORDPs and IDPRs and for different percentiles of disordered residues in IDPs. Proteins with a high percentage of disorder (IDPs) are preferentially encoded by genes with higher GC content, indicating a strong impact of natural selection and mutational bias in controlling the long-time evolution of IDPs.

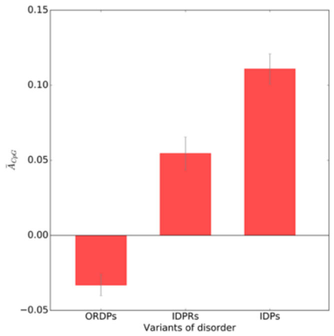

Figure 4. Average abundances of CpG-sites in human genes encoding for ORDPs, IDPRs, and IDPs. IDPs, having the highest fraction of CpG sites, are potentially the variant mostly subjected to CpG methylation (that is, high susceptibility to C-T transition mutation).