Non-random distribution of homo-repeats in eukaryotic and bacterial proteomes and their impact on biological functions

Oxana Galzitskaya, Michail Lobanov

Eukaryotic and bacterial proteomes contain proteins bearing simple amino acid motifs including homo-repeats consisting of a single multiply repeated an amino acid. Functional importance of many homo-repeats remain unclear. Understanding the amino acid tandem repeat function in different proteomes is one of the important functions of molecular biology. It turned out that homo-repeats play important roles in some biological process and may play a more important role in human diseases than it was previously recognized. It was shown that proteins containing alanine repeats of ten and more residues were able to aggregate [1]. It should be stressed that expansion of homo-repeats is a molecular basis for at least 18 human neurological diseases. It has been found several proteins associated with poly-A (alanine) developmental diseases: synpolydactyly type II (HOXD13), blepharophimosis (FOXL2), oculopharyngeal muscular dystrophy (PABPN1), infantile spasm syndrome (ARX), and holoprosencephaly (ZIC2) [2]. Expansion of poly-Q are implicated in several neurodegenerative diseases, including Huntington's disease and several spinocerebellar ataxia's. It should be noted that the length of the poly-Q repeat is critical to pathogenesis. We describe the aggregating properties of proteins such as aggregation values for each amino acid residue along the protein chain using our method FoldAmyloid. The length of homo-repeat that can affect on aggregation properties of protein chain has been found for each amino acid and compared with the random proteome. It has been found that the longer homo-repeats occur in a protein the stronger aggregation ability we observe for protein sequence. In 122 bacterial and eukaryotic proteomes, we observed that the number of proteins containing homo-repeats is significantly larger than expected from theoretical estimates. Our calculations indicate that the minimal homo-repeat length that is statistically significant varies with amino acid type and proteome. In an attempt to characterize proteins harbouring long homo-repeats, we found that those containing A, D, E, G, H, P, Q, R, S and T are enriched in structural disorder and have a large number of protein- and RNA-interactions. For L, S, A, G and P homo-repeats, we observed a tight link with human diseases. Moreover, S, E, P, A, Q, D and T homo-repeats are significantly enriched in neuronal proteins associated with autism and other disorders [3].

1. X.Fan, P.Dion, J.Laganiere, B.Brais, G.A.Rouleau (2001) Oligomerization of polyalanine expanded PABPN1 facilitates nuclear protein aggregation that is associated with cell death, Hum. Mol. Genet., 10:2341–2351.

2. L.Mularoni, A.Ledda, M.Toll-Riera, M.M.Alba` (2010) Natural selection drives the accumulation of amino acid tandem repeats in human proteins, Genome Res., 20:745–754.

3. M.Y.Lobanov, P.Klus, I.V.Sokolovsky, G.G.Tartaglia, O.V.Galzitskaya (2016) Non-random distribution of homo-repeats: links with biological functions and human diseases. Sci Rep., 6:26941.