Journal of Theoretical Biology ■ (■■■■) ■■■–■■■

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Unravelling the relationship between protein sequence and low-complexity regions entropies: Interactome implications

F Martins [a], R. Gonçalves [a], J Oliveira [a], M. Cruz-Monteagudo [c], J.M. Nieto-Villar [d], C. Paz-y-Miño [c], I. Rebelo [a,b,*], E. Tejera [c]

[a] Department of Biochemistry, Faculty of Pharmacy, University of Porto, Portugal
[b] UCIBIO@REQUIMTE, Portugal
[c] Instituto de Investigaciones Biomédicas, Universidad de las Américas, Quito, Ecuador
[d] Dpto. de Química-Física, Fac. de Química, Universidad de La Habana, Cuba. Cátedra de Sistemas Complejos "H. Poincaré", Universidad de La Habana, Cuba

HIGHLIGHTS

- An approximated theoretical model is proposed relating global and local entropy.
- Sequence entropy is related to size instead of the number of low-complexity regions.
- Residue propensity toward low-complexity regions relates with physicochemical properties.
- Low-complexity regions size instead of its number change increase in hubs proteins.
- Hubs proteins show an increment in sequence entropy.

ARTICLE INFO

Article history:
Received 31 March 2015
Received in revised form
12 June 2015
Accepted 28 June 2015

Keywords:
Low-complexity-regions
Protein sequence entropy
Interactome
Hubs

ABSTRACT

Low-complexity regions are sub-sequences of biased composition in a protein sequence. The influence of these regions over protein evolution, specific functions and highly interactive capacities is well known. Although protein sequence entropy has been largely studied, its relationship with low-complexity regions and the subsequent effects on protein function remains unclear. In this work we propose a theoretical and empirical model integrating the sequence entropy with local complexity parameters. Our results indicate that the protein sequence entropy is related with the protein length, the entropies inside and outside the low-complexity regions as well as their number and average size. We found a small but significant increment in the sequence entropy of hubs proteins. In agreement with our theoretical model, this increment is highly dependent of the balance between the increment of protein length and average size of the low-complexity regions. Finally, our models and proteins analysis provide evidence supporting that modifications in the average size is more relevant in hubs proteins than changes in the number of low-complexity regions.

1. Introduction

The low-complexity regions (LCRs) in protein sequence basically results of specific patterns in the primary structure characterized by a low diversity of amino acids or a high repetition of a given amino acid. However, the composition variability of these regions is high as well as their functional relationships (Haerty and Golding, 2010; Rado-Trilla and Alba, 2012; Simon and Hancock, 2009). LCRs, have been associated with intrinsically disordered regions (Karlin et al., 2002; Kumari et al., 2015; Luo et al., 2012; Toretsky and Wright, 2014), different rates of mutability (Mularoni et al., 2006) and with some preponderance in hubs proteins (Coletta et al., 2010; Cumberworth et al., 2013; Dosztanyi et al., 2006; Kumari et al., 2015). Moreover, the presence of LCRs tends to be higher in eukaryotic organisms (Karlin et al., 2002) and also plays an important role in several diseases (Haerty and Golding, 2010; Rado-Trilla and Alba, 2012; Simon and Hancock, 2009). The relevance of the LCRs in the properties of biological systems is clear, however, its structure and direct biological implications are still under intense study (Kumari et al., 2015; Toretsky and Wright, 2014).

* Correspondence to: Department of Biochemistry, Faculty of Pharmacy, University of Porto. Rua de Jorge Viterbo Ferreira n.º 228, 4050-313 Porto. Portugal.
E-mail address: irebelo@ff.up.pt (I. Rebelo).

http://dx.doi.org/10.1016/j.jtbi.2015.06.049
0022-5193/

We can consider the complexity in the LCRs as a local property of the entire protein sequence in opposite to the complexity calculated using the entire protein sequence. As previously mentioned, it is well known that LCRs are related with key protein structural and functional aspects. However, the complexity calculated with the entire sequence (usually calculated by Shannon entropy formulation) (Strait and Dewey, 1996) has being also widely related with protein structure and functional aspects. Previous works suggest that natural protein sequences can be differentiated from random sequences based on structural features (De Lucrezia et al., 2012; Munteanu et al., 2008b; Szoniec and Ogorzalek, 2013). Moreover, complexity evaluation of the entire sequence have been used in different classification tasks (Aguiar-Pulido et al., 2012; Giuliani et al., 2000; Munteanu et al., 2008a), also associated with secondary and tertiary structure information as well as kinetic properties (Concu et al., 2009; Gonzalez-Diaz et al., 2004; Gonzalez-Diaz et al., 2007; Liao et al., 2005; Tejera et al., 2014) as well as Shannon entropy prediction of drug-protein interaction networks (Prado-Prado et al., 2011) and other biological networks (Riera-Fernandez et al., 2012). Interestingly, no previous research on the relationship between entire sequence entropy and LCRs complexity has been found.

In the present work we propose a theoretical model correlating the complexity in the LCRs and the entropy of the entire sequence. We explored the model in real protein sequences as well as the implications of this relationship for the protein interactome and protein randomness.

2. Materials and methods

2.1. Proteins sequence dataset and protein–protein interaction network

The list of all human proteins was downloaded from Human Protein Resource Database (HPRD, release 9) (Peri et al., 2003). This database is widely used in protein–protein network interaction studies and it is also well annotated in terms of protein sequence information. From a total of 30,046 proteins sequences contained in the database, only 22,108 sequences remained after removing those repeated and/or without low-complexity-regions. The HPRD database was also used to extract information from protein–protein network interactions ($n=9.673$ proteins). Labeling a protein as a hub depends of different considerations (Patil and Nakamura, 2006), however, the frequent approach is to define a cutoff value in terms of connectivity to classify hubs and non-hubs proteins in a network (Bertolazzi et al., 2013; Cumberworth et al., 2013; Patil and Nakamura, 2006). In this work, the classification of hubs and non-hubs was done as proposed by Dosztanyi et al. (2006) where a protein is considered as a hub if it is connected with 10 or more proteins.

2.2. Identification of low complexity region and entropies calculations

Among the multiple algorithms available to identify LCRs in protein sequences (Alba et al., 2002; Li and Kahveci, 2006; Wootton and Federhen, 1993) the SEG algorithm (Wootton and Federhen, 1993) is the most frequently used. So, the SEG algorithm was selected for the identification of LCRs.

Briefly, the SEG program divides the sequence into contrasting segments of low-complexity and high-complexity. Locally optimized low-complexity segments are determined with defined levels of stringency, according to formal definitions of local compositional complexity. The SEG algorithm automatically determines the segment length and the number of segments in the protein sequence in two different stages: (1) identification of low complexity segments according to the stringency and resolution of the search and (2) local optimization. For this, SEG implements a mobile window with local computations of the Shannon entropy followed by an optimization of the window size, finally identifying the LCRs.

Considering the protein sequence and the already defined LCRs we defined three entropy indexes using Shannon entropy formalism (Strait and Dewey, 1996): (1) the mean entropy in the LCRs ($<S_{LCR}>$) calculated as the mean entropy in each LCRs. (2) The mean entropy outside the LCRs ($<S_{OLCR}>$) similarly calculated as the mean entropy in each region different that those forming LCRs. (3) The entire sequence entropy (S) calculated using the entire sequence. The approach used for entropy calculation in the entire sequence allows us to perform a global analysis on protein sequence complexity and has been used in previous studies (De Lucrezia et al., 2012; Szoniec and Ogorzalek, 2013; Tejera et al., 2014).

Additionally we also considered the following indexes in our analysis: total number of residues in the LCRs (N_{LCR}), the number of LCRs ($LCR_{\#}$), the average number of residues in the LCRs ($<N_{LCR}>$, defined as $N_{LCR}/LCR_{\#}$) and the length of the protein sequence (N).

Table 1
List of symbols used in all further equations.

Symbol	Description
S	Shannon entropy of the entire sequence.
$<S_{LCR}>$	Mean entropy in the LCRs, calculated as the mean entropy over each LCRs.
$<S_{OLCR}>$	Mean entropy outside the LCRs, calculated as the mean entropy in each region different that those forming LCRs.
$LCR_{\#}$	The number of LCRs.
N	Protein length (total number of residues in the sequence).
N_{LCR}	The total number of residues in LCRs.
$<N_{LCR}>$	The average number of residues in the LCRs (defined as $N_{LCR}/LCR_{\#}$).
f_i	Frequency of the residue "i" in the entire sequence.
f_i^n	Frequency of the residue "i" inside the LCRs. The index "i" indicate a residue present in the LCRs and that can also be or not outside the LCRs.
f_i^x	Frequency of the residue "i" outside the LCRs. The index "i" indicate a residue present in the LCRs and that can also be or not outside the LCRs.
f_j^o	Frequency of the residue "j" outside the LCRs. The index "j" indicate a residue which is only present outside the LCRs.
N_o^{ext}	The number of residues outside the LCRs excluding those also present inside the LCRs.
N^{Ω}	The group of different residues in the entire sequence.
N_n^{Ω}	The group of different residues present in the LCRs.
C	The group of different residues present exclusively outside the LCRs.
S^n	Entropy of a single LCR. Considering that all LCRs are identical then the mean entropy of the LCRs will be equally S^n.
N^n	Number of residues in a single LCR.

Please cite this article as: Martins, F, et al., Unravelling the relationship between protein sequence and low-complexity regions entropies: Interactome implications. J. Theor. Biol. (2015), http://dx.doi.org/10.1016/j.jtbi.2015.06.049