Journal of Theoretical Biology 382 (2015) 320–327

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.
- 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
Available online 8 July 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/© 2015 Elsevier Ltd. All rights reserved.

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_i^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.

3. Results and discussion

3.1. Local vs. global entropy: a theoretical approach

The following section will be devoted to describe the theoretical approach proposed relating local and global entropies. For clarity, the complete list of symbols used and their respective description is provided in Table 1.

Defining the sequence entropy as follow:

$$S=-k\sum_{i=1}^{N^{\Omega}}P_i\ln(P_i)=-\frac{k}{N}\sum_{i=1}^{N^{\Omega}}f_i\ln(f_i)+k\ln N \tag{1}$$

We can rearrange the sum over N^{Ω} in Eq. (1) considering that the alphabet (all possible type of residues in a protein sequence) will be represented as: $N^{\Omega}=\{A,R,N,D,C,E,Q,G,H,I,L,K,M,F,P,S,T,W,Y,V\}$.

The alphabet inside the low-complexity region (N_n^{Ω}) will be a subset of N^{Ω} ($N_n^{\Omega}\subset N^{\Omega}$), we define the subgroup C as: $C\notin(N_n^{\Omega}\cap N^{\Omega})$, accounting for all the residues present outside the LCRs except those that are also present inside the LCRs. For example, if $N_n^{\Omega}=\{A,R,N\}$, then $C=\{D,C,E,Q,G,H,I,L,K,M,F,P,S,T,W,Y,V\}$

Following the previous notations we can rearrange Eq. (1) as follow:

$$S=-\frac{k}{N}\sum_{j=1}^{C}f_j^o\ln\left(f_j^o\right)-\frac{k}{N}\sum_{i=1}^{N_n^{\Omega}}(f_i^n+f_i^x)\ln(f_i^n+f_i^x)+k\ln N \tag{2}$$

where f_j^o corresponds to the frequency of the residue "j" outside the LCRs but this residue is not present in the LCRs. We divided the 20 residues in two groups: (a) those who are only present outside the low complexity regions, denoted with "j" and belonging to the group "C" and (b) those who are present in the LCRs (and that can be present or not outside the LCRs), denoted with "i" and belonging to the group N_n^{Ω}. According previous considerations, all residues in N_n^{Ω} have to be present inside LCRs. Therefore the term $f_i^n+f_i^x$ represents the frequency of the residue "i" in the inner (f_i^n) and the outer (f_i^x) portion of the LCRs. Obviously, if we consider that the residues inside LCRs are not the same than those outside the LCRs, then $f_i^x=0$. Considering, for example, a sequence AFRNTDDCEMMFFTTSWDMAVFY where MMFFTT corresponds to the low-complexity region then: $N_n^{\Omega}=\{M,F,T\}$, $N^{\Omega}=\{A,R,N,D,C,E,S,W,V,Y,M,F,T\}$ and $C=\{A,R,N,D,C,E,S,W,V,Y\}$, therefore, $f_A^o[0,1]=2$, $f_R^o=1,f_N^o=1,f_D^o=3,f_C^o=1,f_E^o=1,f_S^o=1,f_W^o=1,f_V^o=1,f_Y^o=1$ while inside the LCR we will have: $f_M^n+f_M^x=2+1$, $f_F^n+f_F^x=2+2$ and $f_T^n+f_T^x=2+1$.

In order to simplify the Eq. (2) we will consider

1. Outside the LCRs, the residues are randomly and uniformly distributed.
2. The LCRs are usually small fragments composed by a reduced number of residues with a high repetition rate. Based on these aspects we can consider that if a sequence is relatively long, then, for the same residue inside and outside the LCRs $f_i^x=mf_i^n$ where m is a constant. This means that for instance, if an alanine residue has five repetitions in the LCRs and if the sequence is long, the number of alanine repetitions outside the LCRs will be proportional to five. Considering m as a constant imply that this value is independent of the residue type inside the LCR. (This assumption will be further discussed.)

Based on the previous considerations we can rewrite f_j^o as

$$f_j^o=\frac{N-\sum_i^{N_n^{\Omega}}f_i^x}{C}=\frac{N_o^{ext}}{C} \tag{3}$$

where N_o^{ext} is the number of residues outside the LCRs excluding those also present inside the LCRs. Considering Eq. (3) and replacing in Eq. (2)

$$\begin{aligned}S=&-\frac{kN_o^{ext}}{N}\ln\left(\frac{N_o^{ext}}{C}\right)-\frac{k}{N}\sum_{i=1}^{N_n^{\Omega}}f_i^n(1+m)\ln\left(f_i^n(1+m)\right)+k\ln N\\&=-\frac{kN_o^{ext}}{N}\ln\left(\frac{N_o^{ext}}{C}\right)-\frac{k}{N}(1+m)\sum_{i=1}^{N_n^{\Omega}}f_i^n\ln\left[f_i^n(1+m)\right]+k\ln N\\&=-\frac{kN_o^{ext}}{N}\ln\left(\frac{N_o^{ext}}{C}\right)-\frac{k}{N}(1+m)\sum_{i=1}^{N_n^{\Omega}}f_i^n\ln f_i^n\\&-\frac{k}{N}(1+m)\ln(1+m)\sum_{i=1}^{N_n^{\Omega}}f_i^n+k\ln N\\S=&-\frac{kN_o^{ext}}{N}\ln\left(\frac{N_o^{ext}}{C}\right)-\frac{k}{N}(1+m)\sum_{i=1}^{N_n^{\Omega}}f_i^n\ln f_i^n\\&-\frac{k}{N}(1+m)\ln(1+m)N_n+k\ln N\end{aligned} \tag{4}$$

Consider that there are several LCRs and, for simplicity, all these regions are identical between each other with a length N_n and entropy S^n then

$$S^n=S^n=-\frac{k}{N_n}\sum_{i=1}^{N_n^{\Omega}}f_i^n\ln(f_i^n)+k\ln N_n \tag{5}$$

Integrating Eq. (5) in Eq. (4) we get

$$\begin{aligned}S=&-\frac{kN_o^{ext}}{N}\ln\left(\frac{N_o^{ext}}{C}\right)-\frac{k}{N}(1+m)\left[-\frac{N_n(S^n-k\ln N_n)}{k}\right]\\&-\frac{k}{N}(1+m)\ln(1+m)N_n+k\ln N\\S=&-\frac{kN_o^{ext}}{N}\ln\left(\frac{N_o^{ext}}{C}\right)+\frac{N_nS^n}{N}(1+m)-\frac{N_nk\ln N_n}{N}(1+m)\\&-\frac{kN_n}{N}(1+m)\ln(1+m)+k\ln N\end{aligned} \tag{6}$$

We will consider that multiples LCRs in a sequence are identical between each other and therefore could be considered as the N_n^N repetition of one of these regions with length N_n^S, therefore

we can notice that if $k=1$

$$\begin{aligned}S=&-\frac{N_o^{ext}}{N}\ln\left(\frac{N_o^{ext}}{C}\right)+\frac{N_nS^n}{N}(1+m)-\frac{N_n\ln N_n}{N}(1+m)\\&-\frac{N_n}{N}(1+m)\ln(1+m)+\ln N\end{aligned} \tag{7}$$

Thermodynamically, the value of k has several implications and is very hard to explain it in terms of sequence entropy. The common approach is to set k to 1 for simplicity, which have been followed by other authors in terms of sequence entropy (Strait and Dewey, 1996; Szoniec and Ogorzalek, 2013). This is probably related to an intrinsic difficulty to explain it in thermodynamic terms, in addition to be more closely related with the original Shannon information equation. Of course, if $C=\{\}$, which means that the LCRs and the regions outside the LCRs share exclusively the same type of residues, the first term in Eq. (7) will be absent and therefore

$$\begin{aligned}S=&\frac{N_nS^n}{N}(1+m)-\frac{N_n\ln N_n}{N}(1+m)\\&-\frac{N_n}{N}(1+m)\ln(1+m)+\ln N\end{aligned} \tag{7A}$$

Eq. (7A) together with Eq. (7) will be used for simulation purposes. Rewriting: $f_j^o=\frac{N_o^{ext}}{C}=\frac{N-m\sum_i^{N_n^{\Omega}}f_i^n}{C}=\frac{N-mN_n}{C}$, Eq. (8) can be expressed as

$$S=\left(\frac{mN_n}{N}-1\right)\ln\left(\frac{N-mN_n}{C}\right)+\frac{N_nS^n}{N}(1+m)$$

$$-\frac{N_n \ln N_n}{N}(1+m)-\frac{1}{N}(1+m)\ln(1+m)N_n+\ \ln N \qquad (8)$$

In Eq. (7) the terms $-\frac{N_o^{ext}}{N}\ln\left(\frac{N_o^{ext}}{C}\right)+\frac{N_n S^n}{N}(1+m)$ will be, respectively, related with $<S_{OLCR}>$ and $<S_{LCR}>$ calculated from real sequences. This means that they are related with entropy inside and outside the LCRs. The other two terms $-\frac{N_n\ \ln\ N_n}{N}(1+m)-\frac{1}{N}(1+m)\ln(1+m)N_n$ can be related with the length (or average length) of the LCRs ($<N_{LCR}>$), with a negative contribution to the global entropy. Of course, these terms will also depend of the number of LCRs ($LCR_{\#}$) if we consider that not all LCRs are equal between each other, which should be the case in real proteins sequences. Finally, the protein length is positively related with the protein sequence entropy. The theoretical approach leads us to consider that the variation in protein sequence entropy will depend not only of local entropies but also the length and size of the LCRs. However, several approximations were done in the model and therefore, the validity of this conclusion needs to be corroborated by analyzing the real sequence calculations.

3.2. Local vs. global entropy: analysis of real protein sequences

The significant association between the protein length (N) and the number of LCRs ($LCR_{\#}$) but not with the average number of residues in the LCRs ($<N_{LCR}>$) is apparent from Fig. 1 (Left). On the other hand, the increase in the mean entropy in the LCRs ($<S_{LCR}>$) can be associates with an increment in the average number of residues in the LCRs ($<N_{LCR}>$) instead of the number of LCRs ($LCR_{\#}$) (see Fig. 1 Right). As suggested by Eq. (7), such multiple relationships clearly suggest that a more complex model is required in order to include complementary variables.

The relationship between protein length and the number of LCRs (and consequently other related properties) is highly relevant. It implies that all studies intending to explore the variation in the number of LCRs with respect to sequence evolution, protein–protein interactions or even its relationships with structural features will require the simultaneous consideration of the protein length effect.

The use of multivariate models including the protein size and other local sequences parameters is required to obtain a representative description of the global entropy/complexity, which is in better agreement with the model proposed in Eq. (7). Actually, the first model obtained, with an R^2=0.289 (Eq. (9)) shows a negative correlation between $<S_{LCR}>$ and S, while $<S_{OLCR}>$ is positively correlated with S; a misleading result corrected when additional parameters are considered with a multivariable approach.

$$S=-0.116\cdot\langle S_{LCR}\rangle+0.385\cdot\langle S_{OLCR}\rangle+0.275\cdot\ \ln N+3.420 \qquad (9)$$

Including the average LCRs size and the number of LCRs in addition to the protein length the general model (Eq. (9A)) is quite explicative of the previous results and also of the general variance (R^2=0,49):

$$S=0.019\cdot\langle S_{LCR}\rangle+0.046\cdot\langle S_{OLCR}\rangle-0.007{*}\cdot$$
$$\langle N_{LCR}\rangle-0.015\cdot LCR_{\#}+0.111\cdot\ \ln N+3.328 \qquad (9A)$$

This model shows, as deduced from Eq. (7), that $<S_{LCR}>$ and $<S_{OLCR}>$ actually have a positive influence on the protein S while $<N_{LCR}>$ and $LCR_{\#}$ have a negative correlation. This empirical equation (Eq. (9A)) besides explaining 70% of the variance is quite similar to the equation obtained exclusively considering a theoretic approach (Eq. (7)). This empirical model together with the theoretical one already proposed (Eq. (7)) indicates that modifications in proteins sequence entropy will depend of several sequence characteristic. Therefore, when exploring entropic modifications (i.e. interactome) we will need to analyze all its components but first we need to fulfill some of the considerations used to obtain Eq. (7) and their implications/deviations in the analysis of natural sequences.

3.3. The m-parameter in real sequence analysis and amino acids propensities

Our theoretical approach consider m as constant implying that all residues have the same frequency inside and outside the LCRs, which is not true for real sequences (Fig. 2 Left). We can notice that the m-parameter have a wide variation and consequently could affect the results of the theoretical model. In order to study this effect we calculated the intra-sequence variation in the

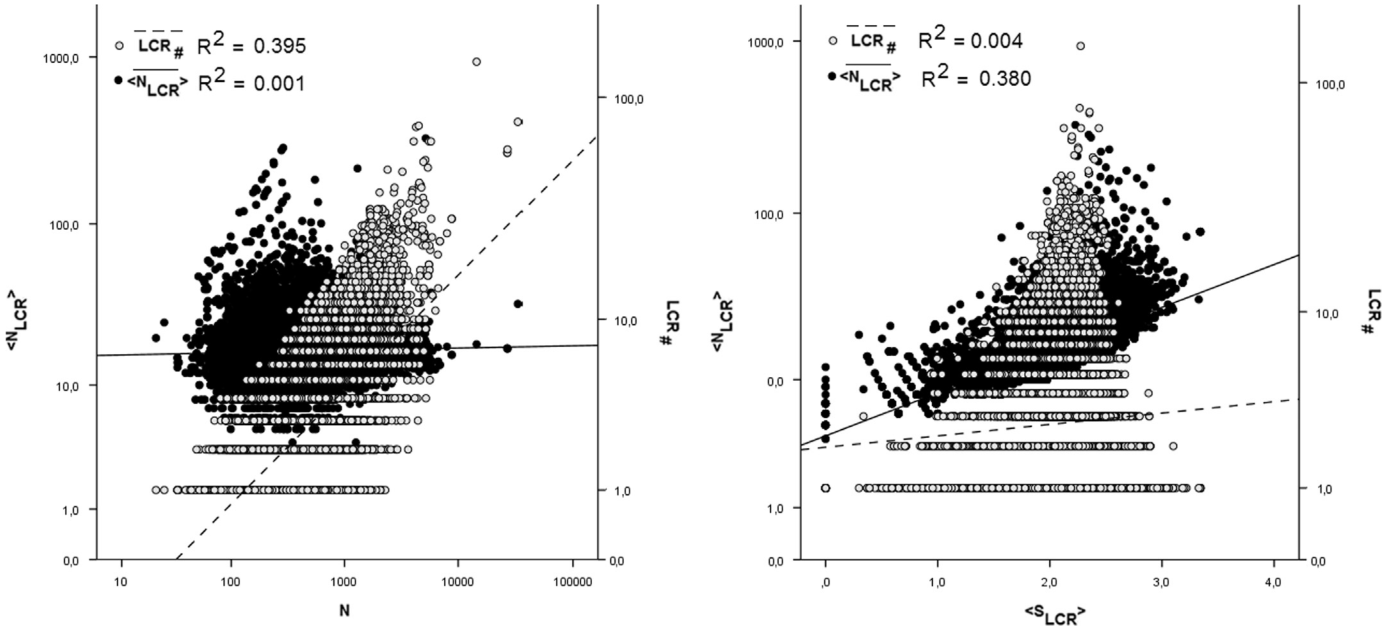

Fig. 1. (Left) Scatterplot of N versus $<N_{LCR}>$ and $LCR_{\#}$. No significant correlation is observed between N and $<N_{LCR}>$. (Right) Scatterplot of $<S_{LCR}>$ versus $<N_{LCR}>$ and $LCR_{\#}$. No significant correlation is observed between $<S_{LCR}>$ and $LCR_{\#}$. All axes are presented in logarithmic scale.

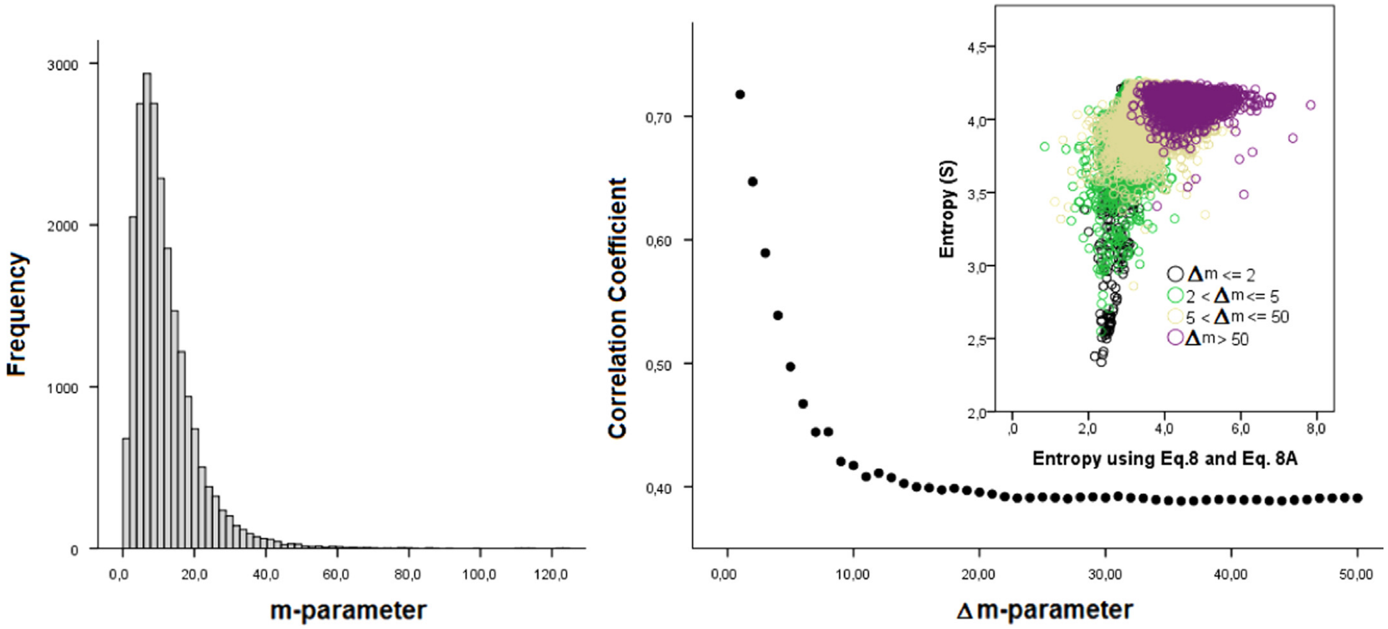

Fig. 2. (Left) Frequency distribution of the m-parameter in the protein sequence dataset. (Right) Variation in the correlation coefficient and the intra-sequence variation of the m-parameter. The correlation coefficient was obtained between the sequence entropy (S) and the entropy calculated (Eqs. (7) and (7A)) using the minimum m-value obtained for each sequence. From the colored graph it is apparent that increasing values of Δm increase the deviation between calculated and real entropy.

m-parameter (Δm) as the absolute difference between the minimal and maximal m-value in a sequence.

The minimum m-value is used in Eqs. (7) and (7A) for the calculation of the theoretical entropy (used in Fig. 2 Right) since using higher values of m (even the mean or median) the contribution of these two terms $-\frac{N_n \ln N_n}{N}(1+m)-\frac{N_n}{N}(1+m)\ln(1+m)$ can be too high to lead to negative values of entropy. On the other hand, our theoretical model considers constant the m-parameter. So, increments in Δm indicate a deviation from this approximation and therefore, a decrement in the correlation coefficient between the entropy calculated from the native sequences and those calculated with our theoretical model (see Fig. 2 Right). However, we can notice a high level of stability in the range of $0 \leq \Delta m \leq |50|$. We can also notice that even with lower values of Δm the calculated entropy is consistently smaller than the real entropy of the sequence and therefore we need to consider this difference when using the model in further sections.

In Eq. (7) the m-parameter represents the ratio between the frequency of a residue inside and outside the LCRs. We can formulate the m-parameter (without invariance consideration) as

$$m_i=\frac{f_i^x}{f_i^n}=\frac{f_i-f_i^n}{f_i^n}=\frac{f_i}{f_i^n}-1$$

where the term $\frac{f_i}{f_i^n}$ is clearly inversely related to the residue propensity toward LCRs, but not equal. This means, the residue propensity toward LCRs will be $T_i=\frac{f_i^n}{f_i}$, however, when a residue is not present in the LCRs, $T_i=0$ but m_i cannot be computed. Similarly, if the residue is equally distributed inside and outside the LCRs then, $T_i=0.5$ and $m_i=1$ $(f_i=2f_i^n)$ but if the residue only exist in the LCRs $T_i=1$ and $m_i=0$ $(f_i=f_i^n)$. Therefore, the value of m for a residue "i" can be related to their propensity to be present in the LCRs but, properly speaking, m is not a propensity index toward LCRs. In Table 1 are shown the mean m-value calculated as $m^1=\frac{1}{NS}\sum\frac{f_i^x}{f_i^n}$ and $m^2=\frac{\sum f_i^x}{\sum f_i^n}$ and the propensity index expressed as $100\frac{\sum f_i^n}{\sum f_i}$. In all cases the sums runs over the entire number of sequences (NS).

Table 2
Median m-parameter and propensity index toward LCRs by residues type.

Residue name	Residue code	Mean m-value (m^1)	Mean m-value^2 (m^2)	LCR propensity	Disorder propensity[a]
Proline	P	8.92	3.44	18.44	1
Alanine	A	9.83	4.73	15.12	0.45
Glycine	G	10.82	4.49	15.10	0.437
Serine	S	10.22	4.82	15.02	0.713
Glutamic acid	E	10.64	4.97	13.58	0.781
Leucine	L	11.09	6.44	12.16	0.195
Arginine	R	11.49	5.64	11.33	0.394
Glutamine	Q	12.29	6.12	9.73	0.665
Lysine	K	12.47	6.38	9.70	0.588
Threonine	T	13.66	7.56	8.54	0.401
Valine	V	14.68	9.23	7.39	0.263
Aspartic acid	D	14.22	8.09	7.29	0.407
Histidine	H	12.11	7.37	5.37	0.259
Isoleucine	I	15.38	10.68	5.10	0.09
Cysteine	C	11.03	7.04	5.01	0
Phenylalanine	F	14.12	9.68	4.92	0.117
Asparagine	N	15.34	11.14	4.40	0.285
Methionine	M	11.31	8.43	4.16	0.291
Tyrosine	Y	13.06	8.70	3.96	0.113
Tryptophan	W	8.03	7.01	3.49	0.004

[a] The disorder propensity as presented in Theillet et al. (2013) and it is calculated based on the fractional difference in the amino acid compositions between the intrinsically disordered and ordered proteins obtained by renormalizing these values to lie between 0 and 1.

It is well known that LCRs composition is enriched primarily of polar and acidic amino acids like E, S, G and A (Haerty and Golding, 2010; Huntley and Clark, 2007). Moreover, the P and L (in lesser degree) repetitions in intrinsic disordered regions of eukaryote organism have been very well described and are also known to be related with important features of the interactome (Hancock and Simon, 2005; Karlin et al., 2002; Lise and Jones, 2005; Theillet et al., 2013). This previous knowledge agrees with our propensity indexes values (see Table 2). In fact, a very strong linear correlation was found between the LCR propensity and disordered propensity indexes (R^2=0.637). The two ways of calculating the mean m-values are very similar (R^2=0.659) with a major difference with respect to the tryptophan residue. Tryptophan (W) is a rare amino acid, poorly represented in the protein sequence space. Additionally, it is the residue with fewer propensities to be present in both LCRs and disordered structures and consequently cataloged as an "order promoter" by some authors (Lise and Jones, 2005; Lobanov et al., 2010; Theillet et al., 2013; Weathers et al., 2007). The first way to calculate the m-value (m^1) is highly representative of local patterns being the average made over the m-value calculated for each residue in a sequence. However, the second approach (m^2) will present a more global pattern because the sum is not over the individual m-values while the ratio is between the total distribution of the residue inside and outside the LCRs.

In fact we can notice that using m^2 the W residue do not present the lowest value even when it is not also the highest as should be expected considering the propensity of the contribution to intrinsic disordered structures. We test the correlation between m^1 and the residues frequency ($R^2 = 0.027$) rejecting the idea that this atypical distribution of W is dependent of the low residue frequency but a particular behavior of the W distribution inside and outside the LCRs. In any case, this means that when the residue W is present inside the LCRs, the ratio between inside and outside will be moved toward a major contribution inside the LCRs when compared with other residues type. A more rigorous analysis should be made in order to clarify the environment of the W residue inside the LCRs leading to this specific distribution.

Table 3
Indexes comparison across hubs and non-hubs protein groups.

	Non-hub proteins ($n = 6362$)	Hub proteins ($n = 1649$)	p-Value
$\boldsymbol{S}$	4.057 (2.339–4.262)	4.067 (2.965–4.246)	0.006
ln$\boldsymbol{N}$	6.223 (3.22–9.08)	6.366 (4.39–8.69)	0.000
$<\boldsymbol{S_{LCR}}>$	2.124 (0–3.201)	2.129 (0–2.942)	0.622
$<\boldsymbol{S_{OLCR}}>$	3.469 (0–4.228)	3.487 (1.284–4.213)	0.197
$\boldsymbol{LCR_{\#}}$	3.84 (1–64)	4.3 (1–41)	0.000
$<\boldsymbol{N_{LCR}}>$	17.611 (4–326.222)	18.479 (5–154)	0.000

Note: The values are presented as: mean (min–max) for each group.

Table 4
Influence of each variable in the hubs/non-hubs classification using logistic models.

	EXP(B)	p-Value
$\boldsymbol{S}$	2.014	0.017
$<\boldsymbol{S_{LCR}}>$	0.870	0.140
$<\boldsymbol{S_{OLCR}}>$	1.004	0.953
$\boldsymbol{LCR_{\#}}$	1.001	0.853
$<\boldsymbol{N_{LCR}}>$	1.019	0.000
ln$\boldsymbol{N}$	1.282	0.000

3.4. Relationship between entropy and the interactome

Most indexes revealed a significant difference when compared (using ANOVA test) between hubs and non-hubs protein groups (see Table 3). As expected, $LCR_{\#}$ and $<N_{LCR}>$ are significantly increased in Hub proteins which is consistent with previous findings on the role of LCRs in the interactome (Coletta et al., 2010; Cumberworth et al., 2013; Dosztanyi et al., 2006; Persi and Horn, 2013). However, according to our theoretical (Eq. (7)) and empirical models (Eq. (9A)), the variation in the sequence entropy is related with multiple sequence features and therefore the ANOVA results (see Table 3) although useful, could not fully represent the complex nature of the sequence entropy. It is necessary to evaluate the simultaneous influence of all variables proposed in our model with respect to hubs and non-hubs proteins.

We can notice in Table 3 that the sequence entropy increase for Hubs proteins. However, $LCR_{\#}$ and $<N_{LCR}>$ also increase, indicating that the protein size effect should be higher to maintain a positive increment of entropy following our previous equations. On the other hand, we already discussed that $LCR_{\#}$ and also S have a dependent relationship with protein length and therefore, combining all this variables, some of them could loss the statistical significance in a hubs/non-hubs comparison.

The results presented in Table 4, obtained from a logistic regression model clearly indicate that when all variables are considered, the local entropies and the number of LCRs ($LCR_{\#}$) are not statistically significant. Moreover, the analysis reveals that the protein length and the average size of the LCRs have the major contributions in hubs/non-hubs classification.

As discussed above, previous studies report a higher number of LCRs in hub proteins. However, our results indicate that this observation could not come only from the increment of the protein size but also from the variation in the average size of the LCRs. In fact, the mean size of the LCRs is more relevant to hubs studies than the number of LCRs (Fig. 3). Consequently, if no multivariate model is considered, the results will be highly dependent of the heterogeneity of the protein sample used in the study.

We can notice that even when the sequence entropy increase in hubs proteins, it is not a high increment. In order to understand this modification we used the Eq. (8) to study the variation of the

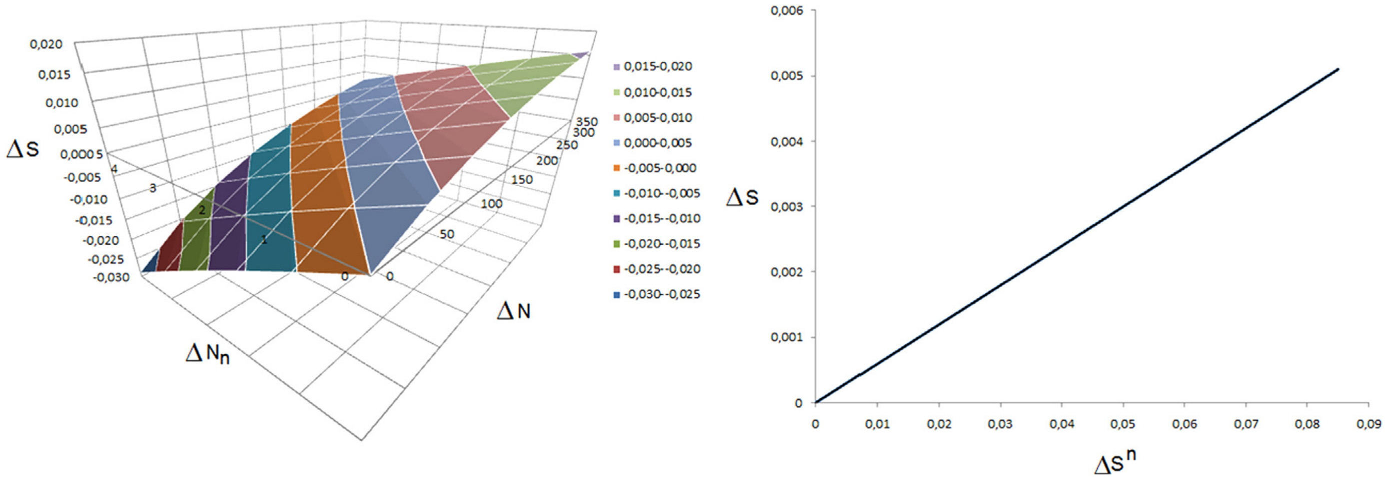

Fig. 3. Calculated entropy variation (Eq. (8)) with respect to the variations in N and N_n (Left). Effects of the S^n variation in the calculated sequence entropy increment (Right). The values $m = 0$ and $C = 10$ were used in all cases.

entropy under different considerations using similar values to those presented in Table 3. We consider in our simulations $m=0$, for simplicity and $C=10$ because it is approximately the mean value in our data. An increment in the protein length around 100 residues and an increment of 1 residue in the average size of the LCRs (simulated by N_n in our model), similarly to the values in Table 3, will produce and increment in S around 0.005 (Fig. 3 Left). It is a similar order of variation presented in Table 3 for the entropy change between hubs and non-hubs. Moreover, even when the modifications in N_n will reduce considerable the entropy, a simultaneous increment of protein length will compensate this by increasing the entropy value (Fig. 3 Left). In our case of hubs proteins, the balance between these two contrary effects are favoured to the protein length increment but also, this suggest that small changes in the average size of the LCRs could have a drastic effect in the protein interaction capabilities. Our results also indicate that modification of the entropy inside the LCRs will produce the smallest increment in the sequence entropy (Fig. 3 Right). This is consistent with no significant variations between hubs and non-hubs, considering the values obtained in Table 3.

This simultaneous effect of protein length and average size of the LCRs can be explained considering that hubs can be composed of single or multiples domains proteins. Proteins with multiples domains could interact with several proteins and will have a larger size that those with singles domains (Bertolazzi et al., 2013; Tejera et al., 2014). On the other hand, other authors have been found that single-domain hubs have a larger fraction of disorder than multi-domain hubs (Cumberworth et al., 2013; Patil and Nakamura, 2006). Therefore we should expect that hubs with single domains will have an increased average size of the LCRs. However, since most of hubs are multi-domains proteins (Bertolazzi et al., 2013; Patil et al., 2010a; Patil et al., 2010b; Schuster-Bockler and Bateman, 2007) and there is also an independent relationship between domains number and sequence entropy (Tejera et al., 2014), we should expect a dominant influence of the protein length instead of the average size of the LCRs. The same explanation can be provided considering also that the number of residue in a domain (known as domain coverage (Bertolazzi et al., 2013)) actually increase in hubs proteins (Bertolazzi et al., 2013; Xia et al., 2008), which also could support the dominant effect of the protein size.

Considering the above discussed results, several topics will need further study, i.e., the implications and applications of the obtained empirical and theoretical equations in protein sequence evolution or the analysis of the protein–protein interaction networks of different organisms. Both, the scope and the amount of information needed to study these topics necessarily require a separated work which should be addressed in future studies.

4. Conclusions

Our results indicate that the entropy of a protein sequence is related with the entropies inside and outside of the low-complexity regions. However, this relationship can be better described when the protein size and the mean size of the low-complexity regions are included in multivariate models, in addition to the number of low-complexity regions. By the first time is presented a theoretical and an empirical model supporting the relationship between these variables.

Additionally, a residue propensity index toward low-complexity regions presence was proposed and related with propensity toward structural disordered regions. We also analyzed the influence of these indexes in our theoretical model.

Contrary to previous works, our results indicates that is the average size of the low-complexity regions rather than its number what is really increased in hub proteins and that also the protein size play an important role in the interactome. We also found that the sequence entropy increase in hubs proteins and that this increment is small and related to the balance between the increment in the protein length and the increment in the average size of the low-complexity regions, which happen in hubs compared to non-hubs proteins.

Acknowledgments

This study was partially supported by Fundação para a Ciência e a Tecnologia (FCT), Grant SFRH/BPD/72391/2010.

References

Aguiar-Pulido, V., Munteanu, C.R., Seoane, J.A., Fernandez-Blanco, E., Perez-Montoto, L.G., Gonzalez-Diaz, H., Dorado, J., 2012. Naive Bayes QSDR classification based on spiral-graph Shannon entropies for protein biomarkers in human colon cancer. Mol. Biosyst. 8, 1716–1722. http://dx.doi.org/10.1039/c2mb25039j.

Alba, M.M., Laskowski, R.A., Hancock, J.M., 2002. Detecting cryptically simple protein sequences using the SIMPLE algorithm. Bioinformatics 18, 672–678.

Bertolazzi, P., Bock, M.E., Guerra, C., 2013. On the functional and structural characterization of hubs in protein-protein interaction networks. Biotechnol. Adv. 31, 274–286. http://dx.doi.org/10.1016/j.biotechadv.2012.12.002.

Coletta, A., Pinney, J.W., Solis, D.Y., Marsh, J., Pettifer, S.R., Attwood, T.K., 2010. Low-complexity regions within protein sequences have position-dependent roles. BMC Syst. Biol. 4, 43. http://dx.doi.org/10.1186/1752-0509-4-43.

Concu, R., Dea-Ayuela, M.A., Perez-Montoto, L.G., Prado-Prado, F.J., Uriarte, E., Bolas-Fernandez, F., Podda, G., Pazos, A., Munteanu, C.R., Ubeira, F.M., Gonzalez-Diaz, H., 2009. 3D entropy and moments prediction of enzyme classes and experimental-theoretic study of peptide fingerprints in Leishmania parasites. Biochim. Biophys. Acta 1794, 1784–1794. http://dx.doi.org/10.1016/j.bbapap.2009.08.020.

Cumberworth, A., Lamour, G., Babu, M.M., Gsponer, J., 2013. Promiscuity as a functional trait: intrinsically disordered regions as central players of interactomes. Biochem. J. 454, 361–369. http://dx.doi.org/10.1042/bj20130545.

De Lucrezia, D., Slanzi, D., Poli, I., Polticelli, F., Minervini, G., 2012. Do natural proteins differ from random sequences polypeptides? Natural vs. random proteins classification using an evolutionary neural network. PLoS One 7, e36634. http://dx.doi.org/10.1371/journal.pone.0036634.

Dosztanyi, Z., Chen, J., Dunker, A.K., Simon, I., Tompa, P., 2006. Disorder and sequence repeats in hub proteins and their implications for network evolution. J. Proteome Res. 5, 2985–2995. http://dx.doi.org/10.1021/pr060171o.

Giuliani, A., Benigni, R., Sirabella, P., Zbilut, J.P., Colosimo, A., 2000. Nonlinear methods in the analysis of protein sequences: a case study in rubredoxins. Biophys. J. 78, 136–149. http://dx.doi.org/10.1016/s0006-3495(00)76580-5.

Gonzalez-Diaz, H., Molina, R., Uriarte, E., 2004. Markov entropy backbone electrostatic descriptors for predicting proteins biological activity. Bioorg. Med. Chem. Lett. 14, 4691–4695. http://dx.doi.org/10.1016/j.bmcl.2004.06.100.

Gonzalez-Diaz, H., Saiz-Urra, L., Molina, R., Santana, L., Uriarte, E., 2007. A model for the recognition of protein kinases based on the entropy of 3D van der waals interactions. J. Proteome Res. 6, 904–908. http://dx.doi.org/10.1021/pr060493s.

Haerty, W., Golding, G.B., 2010. Low-complexity sequences and single amino acid repeats: not just “junk” peptide sequencesGenome 53, 753–762. http://dx.doi.org/10.1139/g10-063.

Hancock, J.M., Simon, M., 2005. Simple sequence repeats in proteins and their significance for network evolutionGene 345, 113–118. http://dx.doi.org/10.1016/j.gene.2004.11.023.

Huntley, M.A., Clark, A.G., 2007. Evolutionary analysis of amino acid repeats across the genomes of 12 Drosophila speciesMol. Biol. Evol. 24, 2598–2609. http://dx.doi.org/10.1093/molbev/msm129.

Karlin, S., Brocchieri, L., Bergman, A., Mrazek, J., Gentles, A.J., 2002. Amino acid runs in eukaryotic proteomes and disease associations. Proc. Natl. Acad. Sci. U S A 99, 333–338. http://dx.doi.org/10.1073/pnas.012608599.

Kumari, B., Kumar, R., Kumar, M., 2015. Low complexity and disordered regions of proteins have different structural and amino acid preferences. Mol. Biosyst. 11, 585–594. http://dx.doi.org/10.1039/c4mb00425f.

Li, X., Kahveci, T., 2006. A Novel algorithm for identifying low-complexity regions in a protein sequence. Bioinformatics 22, 2980–2987. http://dx.doi.org/10.1093/bioinformatics/btl495.

Liao, H., Yeh, W., Chiang, D., Jernigan, R.L., Lustig, B., 2005. Protein sequence entropy is closely related to packing density and hydrophobicity. Protein Eng. Des. Sel. 18, 59–64. http://dx.doi.org/10.1093/protein/gzi009.

Lise, S., Jones, D.T., 2005. Sequence patterns associated with disordered regions in proteins. Proteins 58, 144–150. http://dx.doi.org/10.1002/prot.20279.

Lobanov, M.Y., Furletova, E.I., Bogatyreva, N.S., Roytberg, M.A., Galzitskaya, O.V., 2010. Library of disordered patterns in 3D protein structures. PLoS Comput. Biol. 6, e1000958. http://dx.doi.org/10.1371/journal.pcbi.1000958.

Luo, H., Lin, K., David, A., Nijveen, H., Leunissen, J.A., 2012. ProRepeat: an integrated repository for studying amino acid tandem repeats in proteins. Nucleic Acids Res. 40, D394–D399. http://dx.doi.org/10.1093/nar/gkr1019.

Mularoni, L., Guigo, R., Alba, M.M., 2006. Mutation patterns of amino acid tandem repeats in the human proteome. Genome Biol. 7, R33. http://dx.doi.org/10.1186/gb-2006-7-4-r33.

Munteanu, C.R., Gonzalez-Diaz, H., Magalhaes, A.L., 2008a. Enzymes/non-enzymes classification model complexity based on composition, sequence, 3D and topological indices. J. Theor. Biol. 254, 476–482. http://dx.doi.org/10.1016/j.jtbi.2008.06.003.

Munteanu, C.R., Gonzalez-Diaz, H., Borges, F., de Magalhaes, A.L., 2008b. Natural/random protein classification models based on star network topological indices. J. Theor. Biol. 254, 775–783. http://dx.doi.org/10.1016/j.jtbi.2008.07.018.

Patil, A., Nakamura, H., 2006. Disordered domains and high surface charge confer hubs with the ability to interact with multiple proteins in interaction networks. FEBS Lett. 580, 2041–2045. http://dx.doi.org/10.1016/j.febslet.2006.03.003.

Patil, A., Kinoshita, K., Nakamura, H., 2010a. Domain distribution and intrinsic disorder in hubs in the human protein-protein interaction network. Protein Sci. 19, 1461–1468. http://dx.doi.org/10.1002/pro.425.

Patil, A., Kinoshita, K., Nakamura, H., 2010b. Hub promiscuity in protein–protein interaction networks. Int. J. Mol. Sci. 11, 1930–1943. http://dx.doi.org/10.3390/ijms11041930.

Peri, S., Navarro, J.D., Amanchy, R., Kristiansen, T.Z., Jonnalagadda, C.K., Surendranath, V., Niranjan, V., Muthusamy, B., Gandhi, T.K., Gronborg, M., Ibarrola, N., Deshpande, N., Shanker, K., Shivashankar, H.N., Rashmi, B.P., Ramya, M.A., Zhao, Z., Chandrika, K.N., Padma, N., Harsha, H.C., Yatish, A.J., Kavitha, M.P., Menezes, M., Choudhury, D.R., Suresh, S., Ghosh, N., Saravana, R., Chandran, S., Krishna, S., Joy, M., Anand, S.K., Madavan, V., Joseph, A., Wong, G.W., Schiemann, W.P., Constantinescu, S.N., Huang, L., Khosravi-Far, R., Steen, H., Tewari, M., Ghaffari, S., Blobe, G.C., Dang, C.V., Garcia, J.G., Pevsner, J., Jensen, O.N., Roepstorff, P., Deshpande, K.S., Chinnaiyan, A.M., Hamosh, A., Chakravarti, A., Pandey, A., 2003. Development of human protein reference database as an initial platform for approaching systems biology in humans. Genome Res. 13, 2363–2371. http://dx.doi.org/10.1101/gr.1680803.

Persi, E., Horn, D., 2013. Systematic analysis of compositional order of proteins reveals new characteristics of biological functions and a universal correlate of macroevolution. PLoS Comput. Biol. 9, e1003346. http://dx.doi.org/10.1371/journal.pcbi.1003346.

Prado-Prado, F., Garcia-Mera, X., Abeijon, P., Alonso, N., Caamano, O., Yanez, M., Garate, T., Mezo, M., Gonzalez-Warleta, M., Muino, L., Ubeira, F.M., Gonzalez-Diaz, H., 2011. Using entropy of drug and protein graphs to predict FDA drug-target network: theoretic–experimental study of MAO inhibitors and hemoglobin peptides from *Fasciola hepatica*. Eur. J. Med. Chem. 46, 1074–1094. http://dx.doi.org/10.1016/j.ejmech.2011.01.023.

Rado-Trilla, N., Alba, M., 2012. Dissecting the role of low-complexity regions in the evolution of vertebrate proteins. BMC Evol. Biol. 12, 155. http://dx.doi.org/10.1186/1471–2148-12–155.

Riera-Fernandez, P., Munteanu, C.R., Escobar, M., Prado-Prado, F., Martin-Romalde, R., Pereira, D., Villalba, K., Duardo-Sanchez, A., Gonzalez-Diaz, H., 2012. New Markov-Shannon Entropy models to assess connectivity quality in complex networks: from molecular to cellular pathway, parasite–host, neural, industry, and legal–social networks. J. Theor. Biol. 293, 174–188. http://dx.doi.org/10.1016/j.jtbi.2011.10.016.

Schuster-Bockler, B., Bateman, A., 2007. Reuse of structural domain-domain interactions in protein networks. BMC Bioinform. 8, 259. http://dx.doi.org/10.1186/1471–2105-8–259.

Simon, M., Hancock, J.M., 2009. Tandem and cryptic amino acid repeats accumulate in disordered regions of proteins. Genome Biol. 10, R59. http://dx.doi.org/10.1186/gb-2009-10-6-r59.

Strait, B.J., Dewey, T.G., 1996. The Shannon information entropy of protein sequences. Biophys. J. 71, 148–155. http://dx.doi.org/10.1016/s0006-3495(96)79210-x.

Szoniec, G., Ogorzalek, M.J., 2013. Entropy of never born protein sequences. Springerplus 2, 200. http://dx.doi.org/10.1186/2193-1801-2–200.

Tejera, E., Nieto-Villar, J., Rebelo, I., 2014. Protein sequence complexity revisited. Relationship with fractal 3D structure, topological and kinetic parameters. Phys. A: Stat. Mech. Appl. 410, 287–301. http://dx.doi.org/10.1016/j.physa.2014.05.019.

Theillet, F.-X., Kalmar, L., Tompa, P., Han, K.-H., Selenko, P., Dunker, A.K., Daughdrill, G.W., Uversky, V.N., 2013. The alphabet of intrinsic disorder. Intrinsically Disord. Proteins 1, e24360. http://dx.doi.org/10.4161/idp.24360.

Toretsky, J.A., Wright, P.E., 2014. Assemblages: functional units formed by cellular phase separation. J. Cell Biol. 206, 579–588. http://dx.doi.org/10.1083/jcb.201404124.

Weathers, E.A., Paulaitis, M.E., Woolf, T.B., Hoh, J.H., 2007. Insights into protein structure and function from disorder-complexity space. Proteins 66, 16–28. http://dx.doi.org/10.1002/prot.21055.

Wootton, J.C., Federhen, S., 1993. Statistics of local complexity in amino acid sequences and sequence databases. Comput. Chem. 17, 149–163. http://dx.doi.org/10.1016/0097–8485(93)85006-X.

Xia, K., Fu, Z., Hou, L., Han, J.D., 2008. Impacts of protein-protein interaction domains on organism and network complexity. Genome Res. 18, 1500–1508. http://dx.doi.org/10.1101/gr.068130.107.