Physica A 410 (2014) 287–301

Contents lists available at ScienceDirect

Physica A

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

Protein sequence complexity revisited. Relationship with fractal 3D structure, topological and kinetic parameters

E. Tejera [a], J. Nieto-Villar [b,c], I. Rebelo [a,*]

[a] Instituto de Biologia Molecular y Celular (IBMC)/Faculdade de Farmácia, Departamento de Bioquímica, Universidade do Porto, Portugal

[b] Dpto. de Química-Física, Fac. de Química, Universidad de La Habana, Cuba

[c] Cátedra de Sistemas Complejos "H. Poincaré", Universidad de La Habana, Cuba

HIGHLIGHTS

- Sequence complexity indexes differentiate native from random sequences.
- Proteins with increased complexity revealed a higher number of domains.
- Proteins with lower complexity revealed an increased folding/unfolding constant rate.
- Significant correlation was found between sequence complexity and fractal dimension.

ARTICLE INFO

Article history:
Received 12 December 2013
Received in revised form 9 April 2014
Available online 16 May 2014

Keywords:
Fractal dimension
Protein length
Folding rate
Networks topology
Recurrence analysis
Protein sequence

ABSTRACT

The study of protein sequence complexity is not a new area and several methodological approaches are available in order to describe or represent the protein sequence information. The present study explored the relationship between sequence complexity and structural fractal dimension, secondary structure information, number of domains and also kinetic parameters considering several methodologies. Our results indicate that some sequence complexity indexes are sensitive enough to differentiate native from random sequences, even when the differences are small. We also found that proteins with increased complexity present a higher number of domains, increased length and mean solvent accessibility. Moreover, proteins with lower complexity revealed an increased folding and unfolding constant rate. Interestingly, we found a significant correlation between protein sequence complexity and structural fractal dimension and a significant effect of the secondary structure classes.

1. Introduction

The study of complexity in protein sequences and structures had captivated the attention of many researches across years. On the other hand, fractality had mainly been explored in protein structure, probably because its studies in sequences raise several challenges that will be further discussed. Furthermore, due to the abundance of protein sequences compared to 3D structures, there is a wide scope of methodologies for describing and extracting the protein sequence information [1–5].

In a previous study, we discuss several aspects of fractality in protein structure, exposing several methods for fractal indexes calculation and their relation to structural, thermodynamic and kinetic variables [6]. It is our goal in this study to

* Correspondence to: Rua de Jorge Viterbo Ferreira no. 228, 4050-313 Porto, Portugal. Tel.: +351 220428557.
E-mail addresses: irebelo@ff.up.pt, mirenejesus@hotmail.com (I. Rebelo).

http://dx.doi.org/10.1016/j.physa.2014.05.019
0378-4371/© 2014 Elsevier B.V. All rights reserved.

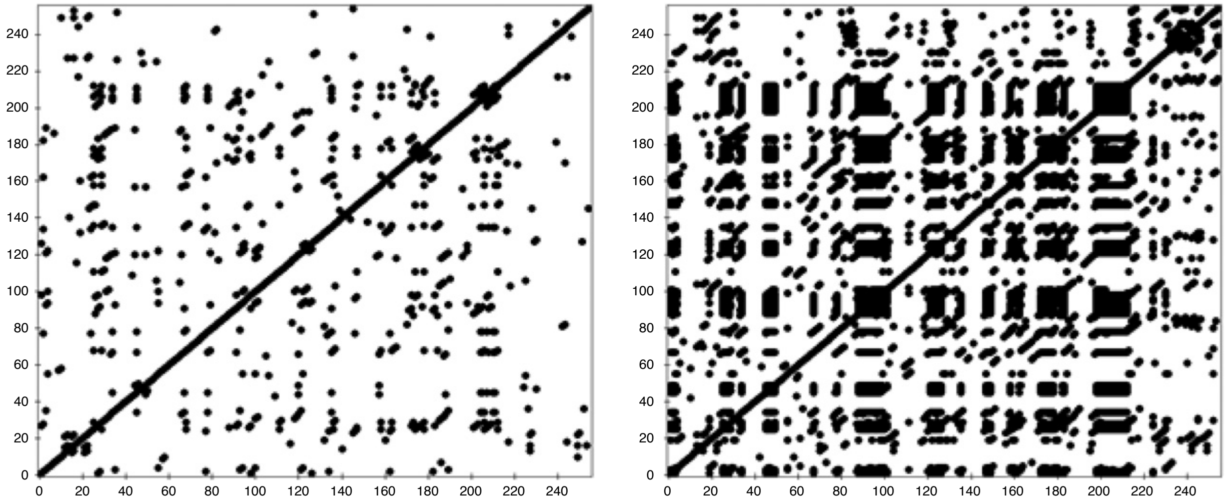

Fig. 1. Recurrence plot of protein 1A7J considering $m = 4$, $\tau = 1$ and $\varepsilon = 10\%$ of the maximum distance. The time series was obtained by (left) direct SA substitution and (right) applying Eq. (1).

perform a similar approach, but considering only the primary structure of proteins. We applied several methodologies for sequence fractality and complexity calculation and demonstrated that some of these indexes are related to the protein folding rate, the protein classes, the secondary structures and other topological parameters. We also identified some correlations between the primary structure and tertiary structure complexity.

2. Theoretical background in protein sequences

There are several approaches to perform a complexity analysis or a general mathematical description of protein sequences. In general, these methods can be divided into two major branches: (a) local analysis (i.e., low-complexity regions [7]) and (b) a global analysis (i.e., sequence entropy and star graph analysis [8,9]). In this study we are focused on the global analysis and specifically the methods that: (1) transform the protein sequence into a numeric sequence by replacing the residue with a numeric value and (2) direct quantification of the sequence based on residues distribution and arrangement.

The first group of methods are connected with the idea of time series (TS) analysis [4,5]. This means that amino acid (AA) sequence is expressed as some propensity index, i.e., AA solvent accessibility, AA hydrophobicity [1,2,5,10] leading to a "time" variation of the selected property. Therefore, we are not analyzing protein sequence itself, but its translation in terms of some AA property that will define the TS correlation patterns. In this study, we consider the AA median solvent accessibility propensity (SA) [11], because the well-known effect of this magnitude in the structural process likes folding, stability and even functionality. In several previous studies, the TS was obtained by direct substitution of propensity value, but we will follow an alternative approach. As SA_i the SA value for the residue at position i, a global property (P_i) is defined as:

$$P_i = \sum_{i,j\neq 1}^{N} SA_i SA_j \mathrm{e}^{-(j-i)^2}. \qquad (1)$$

The final time series is then $\{P_1, P_2, P_3, \ldots, P_N\}$. In all cases, the time series data were normalized (mean $(P) = 0$ and std $(P) = 1$).

2.1. Protein sequence indexes by recurrence quantification analysis

Use of recurrence quantification analysis (RQA) of protein sequences is by far the most common method [3,10,12–14]. The RQA is inspired by the recurrence plot and leads to a family of descriptors in order to quantify the recurrence plot. In general, if P is the protein sequence already translated by some AA property and: $x(i) = (P(i), P(i+\tau), \ldots, P(i+\tau(m-1)))$ where τ is the time delay and m is the embedding dimension, then the recurrence plot (Fig. 1) is defined as [12]: $R(i,j) = \theta(\varepsilon - \|x(i) - x(j)\|)$.

As we can see, $R(i,j)$ is a binary matrix which representation in a two-dimensional space is actually the recurrence plot. The common indexes used in protein sequence analysis are: recurrence rate $\mathrm{RR} = \frac{1}{N^2}\sum_{i,j=1}^{N} R(i,j)$, determinism $\mathrm{DET} = \frac{\sum_{l=l_{\min}}^{N} l\cdot P(l)}{\sum_{i,j=1}^{N} R(i,j)}$, laminarity $\mathrm{LAM} = \frac{\sum_{v=v_{\min}}^{N} v\cdot P(v)}{\sum_{v=1}^{N} v\cdot P(v)}$, average diagonal line $L = \frac{\sum_{l=l_{\min}}^{N} l\cdot P(l)}{\sum_{l=l_{\min}}^{N} P(l)}$, tramping time $\mathrm{TT} = \frac{\sum_{v=v_{\min}}^{N} v\cdot P(v)}{\sum_{v=v_{\min}}^{N} P(v)}$