Journal of Theoretical Biology 391 (2016) 13–20

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

Natural vs. random protein sequences: Discovering combinatorics properties on amino acid words

Daniele Santoni [a], Giovanni Felici [a], Davide Vergni [b,*]

[a] Institute for System Analysis and Computer Science "Antonio Ruberti", National Research Council of Italy, Via dei Taurini 19, 00185 - Rome, Italy
[b] Institute for Computing Application "Mauro Picone", National Research Council of Italy, Via dei Taurini 19, 00185 - Rome, Italy

HIGHLIGHTS

- We developed various measures of associativity between pairs of amino acids in sequences.
- We study amino acid associativity distributions for natural and random sequences.
- We adopt a machine learning approach based on the association values between couples of amino acids to separate natural sequences from random ones.

ARTICLE INFO

Article history:
Received 2 February 2015
Received in revised form
29 July 2015
Accepted 23 November 2015
Available online 2 December 2015

Keywords:
Protein sequence
Random sequence
Combinatorics of words
Amino acid association

ABSTRACT

Casual mutations and natural selection have driven the evolution of protein amino acid sequences that we observe at present in nature. The question about which is the dominant force of proteins evolution is still lacking of an unambiguous answer. Casual mutations tend to randomize protein sequences while, in order to have the correct functionality, one expects that selection mechanisms impose rigid constraints on amino acid sequences. Moreover, one also has to consider that the space of all possible amino acid sequences is so astonishingly large that it could be reasonable to have a well tuned amino acid sequence indistinguishable from a random one.

In order to study the possibility to discriminate between random and natural amino acid sequences, we introduce different measures of association between pairs of amino acids in a sequence, and apply them to a dataset of 1047 natural protein sequences and 10,470 random sequences, carefully generated in order to preserve the relative length and amino acid distribution of the natural proteins. We analyze the multidimensional measures with machine learning techniques and show that, to a reasonable extent, natural protein sequences can be differentiated from random ones.

1. Introduction

Starting from the forties, scientists began to study the amino acid composition of proteins, trying to characterize tissues and species by their amino acid frequencies (Beach et al., 1943). Insulin was the first protein to be sequenced in 1951 and 1952 (chains of bovine insulin B and A, respectively), in the studies that led Sanger (Sanger, 1949) to the Nobel prize in chemistry (1958). In the following years, as new technologies were developed, many other sequences became available opening new frontiers in Biology and Chemistry. In fact, in the seventies a consistent number of sequences was made available, and several new issues related to protein sequences arose. Besides the simple analysis of amino acid frequencies (Smith, 1966), nowadays scientists try to infer functional and structural features of proteins from amino acid sequence patterns or combinatorics properties. Many works show that classification methods are able to separate proteins into different families (Ferràn and Ferrara, 1991; Orengo et al., 1993; Blekas et al., 2005; Exarchos et al., 2006; Kocsor et al., 2006). A specific task is the protein remote homology detection that refers to find proteins with similar structure, starting from sequence similarity (Bowie et al., 1991; Dong et al., 2006; Lingner and Meinicke, 2006; Rangwala and Karypis, 2005), suggesting that natural proteins posses a peculiar structure. In many cases a supervised learning approach has been applied to the classification of proteins in several contexts (e.g., Morgado et al., 2001; Peto et al., 2008; Verma and Melcher, 2012). Moreover, starting from the analysis of short range regularities (Simon, 1989, 1993) and

* Corresponding author. Tel.: +39 06 49270955; fax : +39 06 4404306.
E-mail address: davide.vergni@cnr.it (D. Vergni).

http://dx.doi.org/10.1016/j.jtbi.2015.11.022

passing through the mapping of the protein sequence onto the trajectory of a random walk (Pande et al., 1994), some authors present results confirming the presence of specific features in proteins sequences.

On the opposite direction, many authors claim that the primary structure of a protein is essentially indistinguishable from a random sequence, since the space of all the possible amino acid sequences is so extraordinary large that random elements in the evolutionary process play a major role, leaving to the natural selection just little adjustment in order to obtain the goal of functionality. Various studies about information content (Weiss et al., 2000) or correlation (Weiss and Herzel, 1998; Crooks et al., 2004) in protein sequences affirm that their complexity is essentially the same of a random one. However, a direct study about the possibility to separate between natural and random sequences has been addressed only in very few works (Munteanu et al., 2008; De Lucrezia et al., 2012). The authors of Munteanu et al. (2008), via the construction of a *star network* based on topological indexes (Munteanu et al., 2013) of proteins, can predict with an accuracy of 90.77% whether a protein in their dataset is natural or random using only the amino acid sequence. The second work (De Lucrezia et al., 2012) based on the analysis of structural features of the secondary and tertiary structures, observed on natural and predicted on random sequences, obtains via a neural network a very high rate of correct classification (94.36%) for their dataset.

The above works indicate an interesting research line that we try to further investigate in the present work, aiming at building an effective technique to distinguish natural and random sequences, based only on combinatorics properties of the primary sequence. The main contribution of this paper is to be found in the way the proteins are represented in a multidimensional vector space, based on different association measures between couples of amino acids. For any given measure, we represent a sequence (natural or random) by the matrix of all the possible association values between amino acids. All the sequences are then managed by a supervised learning method (specifically: Support Vector Machines - SVM - see Cortes and Vapnik, 1995; Cristianini and Shawe-Taylor, 2000; Dibike et al., 2001) in the multidimensional space where they are represented. The experimental results, conducted on different samples of sequences, show a good and balanced average correct recognition rate (best result among the different measures is approximately 80%). This is indeed a satisfactory result considering that we accurately generated random sequences with the same statistical properties of the natural ones, keeping in mind that any different strategy could make the separation between random and natural proteins artificially simple.

The paper layout is as follows. In Section 2 we introduce the concept of association between couples of amino acids in a sequence by different proximity functions (Section 2.1). Then we show how to apply *Z*-score to the proximity functions to build different association measures (Section 2.2). In Section 2.3 we introduce both natural and random datasets. In Section 2.4 we show how to represent a sequence (natural or random) by an association matrix made of the *Z*-scores of all the possible amino acids couples, to which we apply a supervised learning approach to tell natural proteins from random ones (Section 2.5). In Section 3, firstly, the association matrices of the sequences in the dataset are analyzed via standard statistical methods (Section 3.1); secondly, the results related to the machine learning approach applied to the association matrices are shown (Section 3.2); thirdly, we compare our method with those of Munteanu et al. (2008) and De Lucrezia et al. (2012) (Section 3.3). Finally, we conclude the paper discussing the peculiarity and the relevance of the present work and suggesting potential application of our method.

2. Materials and methods

Following the same approach developed in Santoni and Pourabbas (2015), we designed a method to evaluate the association degree between pairs of amino acids in a sequence. Two amino acids are likely to be associated if the joint distribution of their relative distances in the sequence is sufficiently different from a distribution occurring in the case of amino acids placed at random. However, for natural sequences of limited length such a joint distribution cannot be estimated with reasonable accuracy; therefore, we introduce different proximity functions able to characterize the distances between amino acids in the sequence. Then we describe protein data and introduce the adopted supervised learning approach.

2.1. Proximity between amino acids

Let A be the alphabet of the 20 amino acids, and v be a word (or a sequence) on the alphabet A, such that $v=\{a_1a_2\dots a_N\}$, where $a_i \in A\ \forall\ i=1,2,\dots,N$. Given a sequence v and two amino acids a and $b \in A$, let $s_a=\{x_1,x_2,\dots,x_n\}$ and $s_b=\{y_1,y_2,\dots,y_m\}$ be arrays of occurrence positions in v of a and b, respectively (where $n>1$, $m>1$). For instance, given the sequence $v=$ AHCNCDDCAWYAHADCE, $s_A=\{1,11,13\}$, $s_C=\{3,5,8,15\}$.

2.1.1. "Minimal" proximity function

Let us define $P_m(a,b)$ as the minimal proximity function between letters a and b in the word v as follows

$$P_m(a,b)=\frac{1}{n}\sum_{i=1}^{n} d_i=\frac{1}{n}\sum_{i=1}^{n}\min_{j=1,\dots,m}\left\{|x_i-y_j|\right\} \tag{1}$$

The meaning of Eq. (1) is easily explained: for each position x_i of the occurrences of a we identify the closest occurrence of b at position y_j, and compute the distance d_i between x_i and y_j; then, we average all d_i and obtain a measure of proximity between the considered letters – see panel a) of Fig. 1.

2.1.2. Standard deviation of the "minimal" proximity function

A potentially interesting variation of the above presented proximity function is based on the standard deviation of the d_i:

$$P_{sm}(a,b)=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(d_i-P_m(a,b))^2} \tag{2}$$

Obviously with this function we obtain information about the spreading of the minimal distance between couples of amino acids.

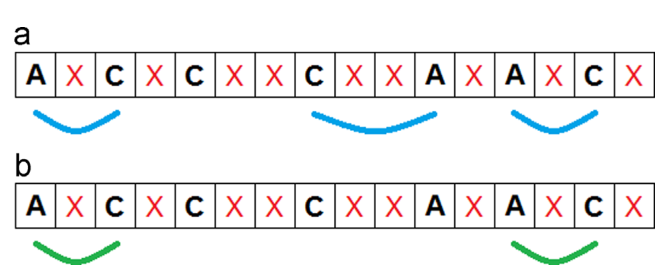

Fig. 1. Pictorial representation of $P_m(a,b)$ (panel a)) and $P_n(a,b)$ (panel b)). A small sequence is sketched emphasizing the amino acids *A* and *C*; *X* indicates any other amino acid. The proximity function $P_m(A,C)$ (from equation (Eq. (1)) in the sequence comes from computing the minimal distance between each occurrence of *A* and the closest occurrence of *C*. Therefore, one obtains: $P_m(A,C)=(2+3+2)/3$ (blue arcs in the panel a)). For the proximity function $P_n(A,C)$ (as described in equation (Eq. (3)) only couples with an *A* followed by a *C* without any of the *A* or *C* in the middle are considered: $P_n(A,C)=(2+2)/2$ (green arcs in panel b)).