UMPA, Ecole Normale Supérieure de Lyon
46, allée d'Italie, F69364 Lyon Cedex 07, France
Email: alexei.tsygvintsev@ens-lyon.fr

.

NATURAL VERSUS RANDOM PROTEINS: NOUVEL NEURAL NETWORK APPROACH BASED ON TIME SERIES ANALYSIS

ALEXEI TSYGVINTSEV

Abstract. We study the set of about 35000 primary structures of natural proteins of length more than 360 residues and the same size set generated via partial or total randomization. Associated to every sequence composed of 20 amino acids, a time series is formed from hydropathy values of the first 360 residues. To measure the absolute deviations of hydropathy index on different time scales, the 24-dimensional vector of total log-amplitudes is introduced. We describe then a configuration of the 1-hidden layer neural network which is trained to solve the binary classification problem of natural and random sequences. A satisfactory distinguishing accuracy random/natural of 88% is obtained.

1. Introduction

The term Never Born Proteins was originally introduced in [1] to describe a collection of randomly generated sequences composed of 20 amino acids which could posses some folding stability properties observed in the laboratory experiments. It seems nevertheless, that from the purely combinatoric point of view, there is no clear difference between primary structures of natural proteins observed in Nature and synthetic amino acids sequences randomly uniformly generated [9]. Based on the Shannon entropy [7], it was reported in [6] that natural protein sequences are random like Never Born Protein sequences i.e their corresponding entropies are asymptotically same. It should be noted that in the above and similar studies only primary structures, viewed as words composed of 20 amino acids, are considered. By contrast, many authors have stressed the importance of the quantitative features emerging from particular secondary or tertiary structures of proteins (experimentally known for natural and predicted for random ones) with the use of which more sharp distinction between random and natural can be achieved (see for example [8] where the evolutional network was designed for this purpose). At the same time,

Key words and phrases. neural networks, amino acid sequences, random sequences, never born proteins, primary structures, hydropathy.

a considerable number of controversial viewpoints on the question of random/natural proteins can be found in literature and apparently many conclusions are methodologically driven. A number of aspects of this problem require further investigation and clear precise definitions.

In this communication we address the problem of sorting out random/natural applying the artificial neural network approach based solely on primary structures and hydropathy values of individual amino acids. We study an initial data base of 35022 natural proteins of length $361 \leq N \leq 400$ residues taken from UniProtKB (http://www.uniprot.org) and the same size data set formed by uniformly randomised (at different degrees) sequences of the same length.

The particular size of $n = 360$ residues is chosen by considering a particularly high number of divisors of n which is 24.

We define in Section 2 the 24-dimensional vector, associated to every amino acid sequence S (natural or randomised) of the length n, whose entries are logarithms of the sum of local amplitudes computed for every partition of S. The corresponding time series is constructed in a natural way by adding the z-score values of hydropathy parameters [5] of amino acids along the sequence S starting from the first left residue. See [4] for some alternative applications of discrete time series representations of protein sequences.

In Section 3, a conventional in machine learning partition $80\,\% - 10\,\% - 10\,\%$ of the initial data set to training-validation-test sets is applied. We use the $24 - 24 - 1$ neural network trained to solve the binary classification problem random/natural for different length of randomized tails of natural primary structures. The trained neural network is capable to classify correctly $88\,\%$ of the testing set containing $50\,\%$ of random and $50\,\%$ of natural amino acids chains.

2. Total amplitudes of amino acid sequences

To analyse the primary structure of proteins, in order to use the time series tools, we need to transform a given sequence of amino acids, composed of 20 residues $\{A, C, D, \dots\}$. into the numerical form. We begin by considering the hydropathy values of amino acids, defined in [5] (Table 1) which are then normalised using z-score (Table 2).

A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
1.8	2.5	-3.5	-3.5	2.8	-0.4	-3.2	4.5	-3.9	3.8	1.9	-3.5	-1.6	-3.5	-4.5	-0.8	-0.7	4.2	-0.9	-1.3

Table 1. Hydropathy values of 20 amino acids

A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
H_1	H_2	H_3	H_4	H_5	H_6	H_7	H_8	H_9	H_{10}	H_{11}	H_{12}	H_{13}	H_{14}	H_{15}	H_{16}	H_{17}	H_{18}	H_{19}	H_{20}
0.7	1.0	-1.0	-1.0	1.1	0.0	-0.9	1.6	-1.1	1.4,	0.8	-1.0	-0.3	-1.0	-1.3	-0.1	-0.0	1.5	-0.1	-0.2

TABLE 2. Z-score Hydropathy values

For the sake of simplicity, we provide the truncated values only while 16-digits precision is used in all numerical computations.

Let $S = (S_1, S_2, \ldots, S_{360})$, $S_i \in \Sigma$ be any sequence of amino acids where

$$\Sigma = \{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y\}\,. \tag{2.1}$$

The time series X associated to S is defined as follows

$$X = (x_1, x_2, \ldots, x_{360}), \quad x_1 = H_1, \quad x_i = x_{i-1} + H_i = \sum_{k=1}^{i} H_k, \quad i \geq 2\,, \tag{2.2}$$

by adding recursively the z-score hydropathy values along the amino acid chain starting from its first residue on the left. Figures 1-4 contain graphs of these series for 4 particular natural proteins. We observe that their structures and regularity can be quite different : from the sideways behaviour (Fig.1-2) to the trend-like one (Fig. 3-4). Fig. 5 illustrates the time series of a typical synthetic sequence whose all residues (excepting the very first one which is fixed to be "M") are chosen randomly.

We fix now $n = 360$ and define

$$D = \{1, 2, 3, 4, 5, 6, 8, 9, 10, 12, 15, 18, 20, 24, 30, 36, 40, 45, 60, 72, 90, 120, 180, 360\}\,, \tag{2.3}$$

the set of all 24 divisors of n.

For a particular divisor $d \in D$ one gets a partition of X

$$X = X_1^d \cup X_2^d \cup \cdots \cup X_{k_d}^d, \quad X_i^d = (x_{(i-1)d+1}, \ldots, x_{id}), \quad k_d = 360/d\,, \tag{2.4}$$

in k_d fragments of equal size d.

For $i \in \{1, \ldots, k_d\}$, the corresponding local i^{th} amplitude is defined according to

$$a_{id} = \max_{x_k \in X_i^d}(x_k) - \min_{x_k \in X_i^d}(x_k)\,. \tag{2.5}$$

The sequence of sums of local amplitudes

$$A_d = \sum_{i=1}^{k_d} a_{id}, \quad 1 \leq d \leq 24\,, \tag{2.6}$$

is mapped then to the 24-dimensional vector

$$a = [a_1, \ldots, a_{24}], \quad a_m = \log(A_m), \quad 1 \leq m \leq 24\,, \tag{2.7}$$

called the *total log-amplitude* of the time series X. Figures 6-7 contain graphical representations of a for two particular natural and random truncated primary sequences. The linear and correlated patterns can be clearly spotted. This is not really surprising, since, as was reported in [2], for many stochastic time series, some entries of total log-amplitude vectors (2.7) obey linear law with the slope given by the fractal dimension of the series in question.

3. Neural network description and solving the classification problem

For numerical simulations, the Java neural network framework Neuroph 2.96 was used.

First, we selected the data set DATA from UniProtKB (http://www.uniprot.org) of primary structures of natural proteins of the length $361 \leq N \leq 400$ which contains 35022 sequences (rows) in total. This set was systematically row shuffled before each new network training. Then, DATA is partitioned into 3 subsets according to

$$\mathrm{DATA} = \mathrm{DATA}^{\mathrm{tr}}_{80\,\%} + \mathrm{DATA}^{\mathrm{val}}_{10\,\%} + \mathrm{DATA}^{\mathrm{test}}_{10\,\%}\,. \tag{3.1}$$

For every of the 3 above subsets, its $\alpha\,\%$-randomised "clone" is created to form a parallel to (3.1) randomised partitioning

$$\mathrm{RAN}(\alpha) = \mathrm{RAN}(\alpha)^{\mathrm{tr}}_{80\,\%} + \mathrm{RAN}(\alpha)^{\mathrm{val}}_{10\,\%} + \mathrm{RAN}(\alpha)^{\mathrm{test}}_{10\,\%}\,, \tag{3.2}$$

where every row of the table $\mathrm{RAN}(\alpha)^i_j$ is obtained by uniform randomization of $\left[\frac{\alpha}{100}n\right]$, $n = 360$ amino acids of every row of the table DATA^i_j starting from the last residue on the right, while leaving others to be intact (with the very first residue on the left "M" always kept unchanged).

The sets of both types are then joined together to build a new data set

$$D = D_1 + D_2 + D_3\,, \tag{3.3}$$

with

$$\begin{aligned} D &= (\mathrm{DATA}, \mathrm{RAN}(\alpha)) & D_1 &= (\mathrm{DATA}^{\mathrm{tr}}_{80\,\%}, \mathrm{RAN}(\alpha)^{\mathrm{tr}}_{80\,\%}) \\ D_2 &= (\mathrm{DATA}^{\mathrm{val}}_{10\,\%}, \mathrm{RAN}(\alpha)^{\mathrm{val}}_{10\,\%}) & D_3 &= (\mathrm{DATA}^{\mathrm{test}}_{10\,\%}, \mathrm{RAN}(\alpha)^{\mathrm{test}}_{10\,\%})\,, \end{aligned} \tag{3.4}$$

which is used to train the $24 - 24 - 1$ neural network. The Sigmoid activation function was chosen and the z-score normalisation was applied to input vectors whith parameters (empirical average and variance) computed from the $\mathrm{DATA}^{\mathrm{tr}}_{80\,\%}$ set only. The desired values of the network's data set were fixed according to: "0" $\sim \alpha\,\%$-random input row and "1" $\sim$ natural primary structure row. Training was done using the back-propagation in a batch mode using the D_1 set. To define the stoping rule, one waits until the LMS

error computed on the validation D_2 set starts to increase while the LMS error evaluated on the training D_1 set is still decreasing.

4. Results and conclusion

The capacity of trained network was evaluated to classify sequences from the test set D_3, not used during the training, and which composition was 50/50 natural or random. The results are quite satisfactory and listed in Table 3.

α	100 %	50 %	33 %	19 %
A_D	88 %	77 %	71 %	65 %

Table 3. Distinguishing accuracy A_D of the neural network for different α's

Decreasing of A_D with lower values of α is expectable since $A_D \to 50\%$ as $\alpha \to 0$ with the difference random/natural disappearing. The maximum 88% of the accuracy can be explained by the fact that only limited part of protein's structure is coded by hydropathy values of amino acids. Our preliminary findings indicate that neural networks are able to uncover hidden distinguishing patterns in total log-amplitude vectors of natural and random amino acid sequences. Since these numerical features are intimately correlated to fractal and self-similarity characteristics of time series [2], it would be of interest to establish links between our results and other studies of the protein fractal structures [3].

Acknowledgments. The author is grateful to I. Gannaz end C. Combet for fruitful discussions and provided valuable assistance in accessing the protein data base.

References

[1] Chiarabelli C, Vrijbloed JW, De Lucrezia D, Thomas RM, Stano P, Polticelli F, Ottone T, Papa E, Luisi PL., *Investigation of de novo totally random biosequences, Part II: On the folding frequency in a totally random library of de novo proteins obtained by phage display*, Chem Biodivers. 2006 Aug; 3(8):840-59

[2] Mikhail M Dubovikov, Nikolai V Starchenko, *Econophysics and the fractal analysis of financial time series*, Physics-Uspekhi, Volume 54, Number 7, 2011

[3] Darja Kanduc, Giovanni Capone, Giuliano Losa, *The Fractal Dimension of Protein Information*, Advanced Studies in Biology, Vol. 2, 2010, no. 2, 53 - 62

[4] Gupta R, Mittal A, Singh K., *A Time-Series-Based Feature Extraction Approach for Prediction of Protein Structural Class*, EURASIP J Bioinform Syst Biol. 2008:235451. doi: 10.1155/2008/235451.

[5] Kyte J, Doolittle RF., *A simple Method for Displaying the Hydropathic Character of a Protein*, J Mol Biol. 1982 May 5;157(1):105-32

[6] Grzegorz Szoniec, Maciej J Ogorzalek, *Entropy of never born protein sequences*, Springerplus. 2013; 2(1): 200, doi: 10.1186/2193-1801-2-200

[7] Shannon CE, *A mathematical theory of communication*, The Bell System Technical Journal,Vol. 27, pp. 379–423, 623–656, July, October, 1948

[8] De Lucrezia D, Slanzi D, Poli I, Polticelli F, Minervini G, *Do Natural Proteins Differ from Random Sequences Polypeptides? Natural vs. Random Proteins Classification Using an Evolutionary Neural Network*, (2012) PLoS ONE 7(5): e36634. https://doi.org/10.1371/journal.pone.0036634

[9] Weiss O, Jiménez-Montaño MA, Herzel H., *Information content of protein sequences*, J Theor Biol. 2000 Oct 7; 206(3):379-86

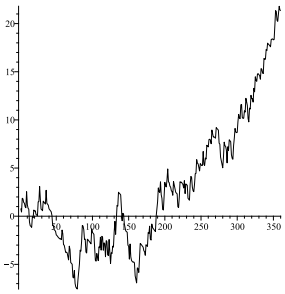

FIGURE 1. Time series formed by first 360 residues of of the Putative movement protein, Q91TW8

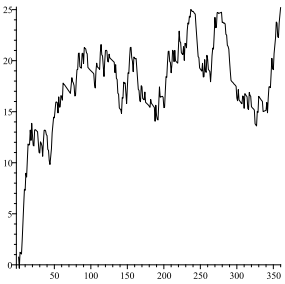

FIGURE 2. Probable GPI-anchored adhesin-like protein PGA32, Q5ADQ7

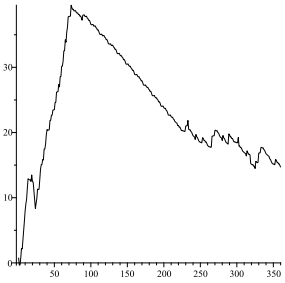

FIGURE 3. Prisilkin-39, C0J7L8

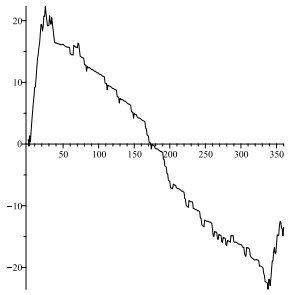

FIGURE 4. 29C-likeproteinDDB_G0287399, Q54KD5

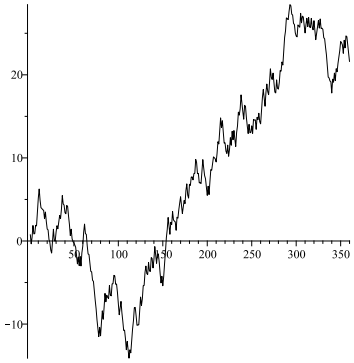

FIGURE 5. Uniformly random sequence of 360 residues

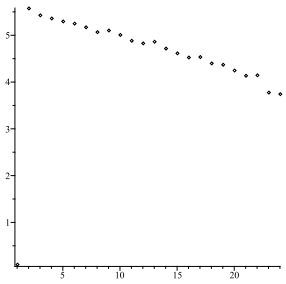

FIGURE 6. The total log-amplitude vector for random sequence from Fig.5

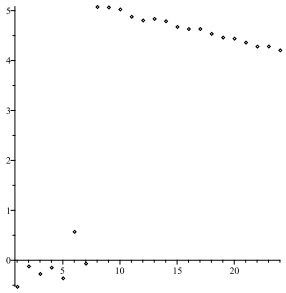

FIGURE 7. The total amplitude vector of 360-residues left tail of GDSLesterase, Q9SVU5