THE CHAOTIC STRUCTURE OF BACTERIAL VIRULENCE PROTEIN SEQUENCES

Sevdanur Genc[1], Murat Gok[2], Osman Hilmi Kocal[3]

[1]Institute of Science, Yalova University, Yalova, Turkey
sevdanurgenc@gmail.com
[2,3]Department of Computer Engineering, Yalova University, Yalova, Turkey
murat.gok@yalova.edu.tr, osman.kocal@yalova.edu.tr

ABSTRACT

Bacterial virulence proteins, which have been classified on structure of virulence, causes several diseases. For instance, Adhesins play an important role in the host cells. They are inserted DNA sequences for a variety of virulence properties. Several important methods conducted for the prediction of bacterial virulence proteins for finding new drugs or vaccines.

In this study, we propose a method for feature selection about classification of bacterial virulence protein. The features are constituted directly from the amino acid sequence of a given protein. Amino acids form proteins, which are critical to life, and have many important functions in living cells. They occurring with different physicochemical properties by a vector of 20 numerical values, and collected in AAIndex databases of known 544 indices.

For all that, this approach have two steps. Firstly, the amino acid sequence of a given protein analysed with Lyapunov Exponents that they have a chaotic structure in accordance with the chaos theory. After that, if the results show characterization over the complete distribution in the phase space from the point of deterministic system, it means related protein will show a chaotic structure.

Empirical results revealed that generated feature vectors give the best performance with chaotic structure of physicochemical features of amino acids with Adhesins and non-Adhesins data sets.

KEYWORDS

Bioinformatics, Virulence Protein Sequences, Attribute Encoding, Chaotic Structure, Classification.

1. INTRODUCTION

Proteins, which have vital importance for organisms, reacts in all biochemical reactions. Physicochemical properties of amino acids are the most important determinant to formation for three-dimensional structure of proteins and binding orders of amino acids. In addition, physicochemical properties determine functions of proteins and life cycles.

David C. Wyld et al. (Eds) : ITCS, CST, JSE, SIP, ARIA, DMS - 2015
pp. 147–155, 2015. © CS & IT-CSCP 2015 DOI : 10.5121/csit.2015.50115

Physicochemical properties have different 544 input data. We think that these properties shows a chaotic structure because they create a certain system and this certain system also affect themselves.

According to chaos, a deterministic system can behave in irregular. In other words a deterministic system can behave in an unexpected way. Chaos that depends on certain parameters usually appear in undetermined, complicated and nonlinear systems.

544 physicochemical properties should be analyzed in a scalar form of data in the phase space by regenerating. Equality of movements in phase space will show the results of the model to quantiles such as positioning attractor dimensions and Lyapunox exponents of system.

Bacterial virulence protein sequences have similar patterns. On these similarities are pretty difficult estimation for classification. In studies conducted to date, based on different strategies about various methods have been proposed to estimation of virulence proteins. To illustrate, the first studies and developed methods were based to search similarity such as BLAST [1] ve PSI-BLAST [2]. In more recent times, machine learning algorithm used for estimation. In the recent times, studied for estimation of bacterial virulence proteins about physicochemical properties.

Our aim in this study is that proving a chaotic structure of physicochemical properties of amino acids that constitute bacterial virulence proteins.

2. MATERIALS AND METHODS

2.1 Dataset

In this study, Adhesins dataset has been used on SPAAN. Dataset have 469 Adhesins and 703 non-Adhesins proteins. Adhesins protein sequences were downloaded from http://www.ncbi.nlm.nih.gov using the keyword 'Adhesin'. Non-adhesins, The rationale we used here was to collect sequences of enzymes and other proteins that function within the cell. They probably have remote possibility of functioning as adhesins and would differ in compositional characteristics (Nakashima and Nishikawa, 1994) [3].

Also, current version of AAIndex included in the dataset. The AAIndex contains 544 amino acid indices. For each the properties of 20 amino acids, input consists of reference information, a short description of the index, an accession number and the numerical values.

In additional to this study, algorithm of this application developed on MatLab. For all Lyapunov Exponents were calculated by Tisean Software.

2.2. Physicochemical Properties Of Amino Acids

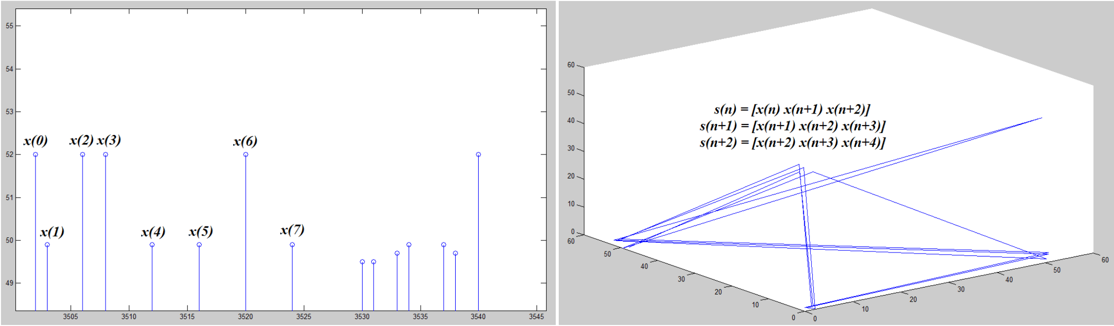
Amino acids which, determine the functions of proteins have different physicochemical properties such as hydrophobicity, polarity and molecular weight. These properties, termed the amino acid indices, can be represented with 20 numeric values of vectors.

2.3. Chaotic Type Features

In exploring the link between physicochemical properties and chaos, phase spaces of physicochemical properties are necessary. They are constructed from 1-D series, and by considering a high-dimensional vector, the dynamics of physicochemical properties production system can be unfold. The phase-space vector of physicochemical properties are reconstructed as follows;

$$\mathbf{s}(n) = [\ x(n)\ x(n+1)\ \ \ldots\ \ x(n+(D_E - 1))] \qquad (1)$$

where x(n) is the *n*th sample of the physicochemical properties, D_E is the embedding dimension of the phase space [4].



(a) (b)

Figure 1. (a) series of the signal for physicochemical properties and (b) reconstructed phase space of the signal for $D_E = 3$.

Figure 1. (a) shows a segment of the scalar physicochemical properties with its time-delay vector trajectory. Figure 1. (b) shows phase space for D_E.

Phase spaces make signal dynamics clearly observable, which is not easy to see in a series representation. Some similar patterns come into closer proximity in phase-space representation, though they are apart in series representation. After determining the appropriate embedding dimension D_E for series, chaotic-type features, such as Lyapunov Exponent, calculated for D_E -dimensional phase space [4].

2.4. Lyapunov Exponents

The Lyapunov Exponent is a quantitative measure for the divergence of nearby trajectories, the path that a signal vector follows through the phase space. The rate of divergence can be different for different orientations of the phase space. Thus, there is a whole spectrum of Lyapunov Exponents - the number of them is equal to the number of dimension of the phase space. A positive exponent means that the trajectories, which are initially close to each other, move apart over time (divergence). The magnitude of a positive exponent determines the rate as to how rapidly they move apart. Similarly, for negative exponents, the trajectories move closer to each other (convergence) [4].

The lyapunov exponent is calculated for each dimension for the phase space as

$$\lambda = \lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} \ln \frac{d(\mathbf{s}(n+1), \mathbf{s}(m+1))}{d(\mathbf{s}(n), \mathbf{s}(m))}. \qquad (2)$$

Here, **s**(n) is the reference point and **s**(m) is the nearest neighbor of **s**(n) on a nearby trajectory. d(**s**(n),**s**(m)) is the initial distance between the nearest neighbors. d(**s**(n + 1),**s**(m + 1)) is the distance between s(n + 1) and s(m + 1) which are the next pair of neighbors on their trajectories. It must be considered that the LE calculation algorithm finds a new nearest neighbor **s**(m) for each **s**(n), (n = 1,2,...,N). There are D_E Lyapunov exponents (i.e., {λ_1, λ_2,..., λ_{DE} }) in descending order. λ_1 is known as the largest LE and a positive largest LE is the indicator of chaos [4].

After calculating lyapunov exponents for all of the nearest neighbor pairs on different trajectories, the lyapunov exponent for the whole signal is calculated as the average of these lyapunov exponents, as the magnitude on d(s(n),s(m)) depends on the whole phase space only, and distortion does not change the global form of the phase space [4].

2.5. Tisean Software Project

In this study, developed of this project by Tisean. It is a software project for the analysis of time series with methods based on the chaos theory. Tisean Package is analysis of time series with methods based on the theory of nonlinear deterministic dynamical systems. Tisean software can be downloaded from http://www.mpipks-dresden.mpg.de/~tisean/. Also, we studied with Cygwin for Tisean software. Cygwin can be downloaded from https://www.cygwin.com/. In this problem, using Tisean Package with MATLAB by Cygwin on Windows 7 operation system.

3. EXPERIMENTAL SETUP

In this study, we advanced an algorithm that using Lyapunov exponents for the purpose of classification using physicochemical properties of residues. Through this algorithm, we will demonstrate chaotic structure of bacterial virulence protein sequences. We think that protein sequences show chaotic structure with the physicochemical properties of amino acids. Our hypothesis prove that, we developed an application which using Tisean Software Project with MatLab. Figure 3. shows a work flow diagram forming of physicochemical properties.

According to Figure 2, first step, we used the *AAIndex* for values of 20 amino acids. AAIndex has 544 physicochemical properties to each amino acids. Thus, we have created a matrix with size of [20 X 544]. Thereby, getting the length of each protein have created a matrix with size of [1 X Protein Length]. For example, show regard to protein that is hemolysin / hemagglutinin - like protein HecA (Erwinia chrysanthemi). It consists by 3848 amino acid. So, This sequence should be in size of 3848, and also should be create a matrix with size of [1 X 3848].

According to that matrix with size of [1 X 3848], can create a matrix with 544 physicochemical properties for each amino acids about that proteins. So, length of the just created a matrix should be in size of [544 X Protein Length]. Also respectively created 544 files within terms of protein length, line by line. And then, they saved and moved in one directory, called *AA2File*.

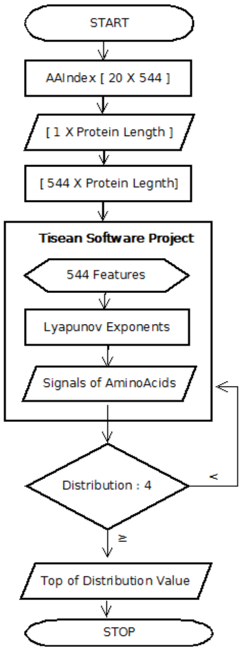


Figure 2. Work Flow Diagram forming of physicochemical properties

The AA2File folder includes each files of 544 properties, in addition to that will be calculated all of them with using Tisean Software Project by Matlab. And then, receive results of 544 properties from Tisean. The file extensions named as .LYAP. Also, they saved and moved in one directory, named as File2Lyap.

Whole of result are received by Tisean, because of that analyzed to convert at lyapunov exponent. Lyapunov Exponents have observable results at LYAP file; such as protein length, signals belong to each amino acids of a protein, average relative forecast erros, average absolute forecast errors, average neighborhood size, average number of neighbors and estimated KY-Dimension. Last of all, according to the signs of lyapunov exponents is given to the decision by Tisean.

Tabel 1. shows a lyapunov exponent results of the protein with 3848 length, as follows;

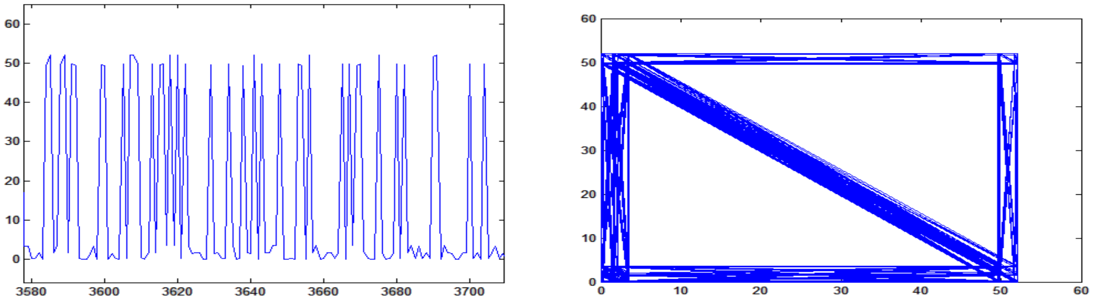
Table 1. The LEs results of the protein with 3848 length

The Lyapunov Exponents Results	
Length of Protein (Amino Acid) :	3848
Signals belong to each amino acids of a protein :	
9.044008e-001 4.018379e-001 2.226827e-001 1.310360e-001 4.904960e-002 -1.972889e-002 -1.013593e-001 -1.959663e-001 -3.338790e-001 -7.586726e-001	
Average relative forecast errors :	1.157800e+000
Average absolute forecast errors :	2.145047e+001
Average Neighborhood Size :	2.398085e+001
Average num. of neighbors :	3.000000e+001
Estimated KY-Dimension :	10.000000

Finally, LEs results show evidence of the better distribution by LYAP file, because of that the chaos theory analyzed on 2-D or 3-D in phase space. Our results demonstrate that negative lyapunov exponents and positive lyapunov exponents can be determined as quickly and easily, simply by analyzing data in Tisean as Table 1. In such a case that, the first few positive and negative exponents are enough to characterize the complete distribution. So, need to know for sure that the chaotic system should to have at least one positive lyapunov exponents. According to that this protein has five positive and five negative lyapunov exponents as Table1 shown. So, there are more protein like this sample on Adhesins and non-Adhesins Datasets.

4. Experimental Results

In this hypothesis, will explain with an experiment that fit for purpose with our algorithm. First of all, shows physicochemical properties on phase space, after then calculate results of lyapunov exponents. At the beginning that create a *x* vector by matrix size of [1 X 3848]. Figure 3. (a) shows a line plot of the data values in *x*.



(a) (b)

Figure 3. (a) Creates a line plot of the data values in *x* and (b) shows a 2-D line plot of data.

*n*th sample of the physicochemical properties are calculated on 2-D phase space by MATLAB. Thus, Figure 3. (b) shows a 2-D line plot of data, *Y* axis values and *X* axis values. And finally, *n*th sample of the physicochemical properties are calculated on 3-D phase space by MATLAB. Thus, Figure 4. shows *X, Y* and *Z* are vectors, plots one or more lines in three-dimensional space through the points which coordinates are the elements of *X, Y* and *Z.*

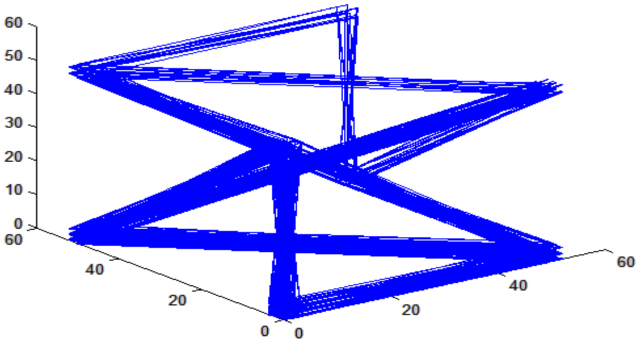


Figure 4. Creates a 3-D line plot of the data values in *X,Y and Z.*

In examined to Lyapunov Exponents signs, according to chaos has been shown an alteration on the phase space, it may be appear or not. So, Lyapunov Exponent signs were vary chaotic structure at phase space. Also, according to length of phase space that LEs show reconciliation and rapprochement at all dimensions of LEs signs as Figure 4. shown. It has a real distribution on phase space.

For this study, used Adhesins Dataset. Thus, give an example of the results, consider a protein that is hemolysin / hemagglutinin - like protein HecA. Top of distributions' values have results between greater or equal to 4. For instance, these values generally have four positive and six negative signs, five positive and five negative signs, six positive and four negative signs etc.

If analyze the Table 2., show the best results from 544 properties on AAIndex Table via Tisean - Lyapunov Exponent Calculator.

Table 2. Gives the best results for protein HecA

Featues of AAIndex	**Positive**	**Negative**
14	4	6
96	4	6
114	4	6
173	4	6
185	4	6
247	4	6
252	4	6
400	***5***	***5***
488	4	6
496	4	6
537	4	6
538	4	6
539	4	6
543	4	6

If analyze the Table 3., shown Lyapunov Exponents, that include to five postive and five negative on phase space, according to chaos theory has distribution as chaotic structure.

Table 3. Lyapunov Exponents belong to each aminoacids for protein HecA

LEs belong to each aminoacids of protein HecA	
Positive	**Negative**
9.044008e-001	*-1.972889e-002*
4.018379e-001	*-1.013593e-001*
2.226827e-001	*-1.959663e-001*
1.310360e-001	*-3.338790e-001*
4.904960e-002	*-7.586726e-001*

Consequently, we have proved this hypothesis that physicochemical properties have distribution on phase space as chaotic structure. So, there are positive and negative signs from lyapunov exponents of phase space at this distribution, and in addition that belong to each amino acids of proteins on Adhesins dataset.

5. Results

In this paper, we analyzed to Adhesins and non-Adhesins datasets at chaotic structure at phase space. Shortly after, we determined the best physicochemical features on these dataset by result of Lyapunov Exponents. Therefore, if analyzed to other datasets on phase space, in our opinion that dataset will obtain different results about physico chemical features. In the future works, we will be using our developed method, also will be improving this algorithm for classification to unfolded protein regions.

Unfolded protein regions plays an important role in determining transcriptional and translational regulation of the protein, protein-protein, protein-DNA interactions and tertiary structure. To date, studies have been shown that unfolded regions associated with cancer, cardiovascular, diabetes, autoimmune diseases and neurodegenerative disorders.

As a result, will be explained to generate feature vectors that using with chaotic structure for prediction of unfolded protein regions and also we will be making comparison with similar studies in the literature about classification and recognition.

Acknowledgements

This work was performed with aim of master thesis under the auspices of the Department of Computer Engineering under The University of Yalova.

References

[1] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, and D.J. Lipman, “Basic Local Alignment Search Tool”, J. Molecular Biology, vol. 215, pp. 403-410, 1990

[2] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman, “Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs”, Nucleic Acids Research, vol. 25, pp. 3389-3402, 1997.

[3] Sachdeva, G., Kumar, K., Jain, P., and Ramachandran, S. "SPAAN: a software program for prediction of adhesins and adhesin-like proteins using neural networks", Bioinformatics, vol. 21, pp. 483-491, 2005.

[4] Kocal, O.H., Yuruklu, E., Avcibas, I. "Chaotic-Type Features for speech steganalysis", IEEE Transactions on Information Forensics and Security, vol. 3, pp. 651-661, 2008a

AUTHORS

Osman Hilmi Kocal was born in Istanbul, Turkey, in 1967. He received the B.Sc., M.Sc., and Ph.D. degrees in electronics and telecommunication engineering from the Technical University of Istanbul, Istanbul, in 1989, 1992, and 1998, respectively. He was a Research Assistant with the Technical University of Istanbul and Turkish Air Force Academy, Istanbul, respectively. Currently, he is an Assistant Professor with the Department of Computer Engineering, Yalova University, Yalova, Turkey. His research interests include chaotic signals and adaptive signal processing.

Murat Gok was born in Kirsehir, Turkey in 1976. He received the B.Sc. degree in electrical and computer education from the Marmara University, Istanbul, in 2000. The M.Sc. degree in electrical and computer education from the Mugla University, Mugla, in 2006. And the Ph.D. degree in electrical and computer education from the Sakarya University, Sakarya, in 2011. Currently, he is an Assistant Professor with the Department of Computer Engineering, Yalova University, Yalova, Turkey. His research interests include pattern recognition, machine learning algorithms, feature extraction and selection, bioinformatics, protein classification and decision support systems.

Sevdanur Genc was born in Istanbul, Turkey in 1983. She received the B.Sc. degree in computer engineering from the Suleyman Demirel University, Isparta, 2010. Also, She is currently pursuing the M.Sc. degree in computer engineering from the Yalova University, Yalova. Her research interests include cloud computing, parallel programming, bioinformatics, remote sensing, machine learning algorithms, computer vision and pattern recognition.