PRELIMINARY INVESTIGATION ON NONLINEAR DYNAMICAL MODELING OF THE BIOLOGICAL SEQUENCES

Mai S. Mabrouk[1], Ahmed K. Atwa[2], Heba M. Afify[2], Manal Abdel-Wahed, Nahed H. Solouma[3], and Yasser M. Kadah[2]
[1]Department of Biomedical Engineering, Misr University for Science and Technology, 6 October, Egypt
[2]Institute of Biomedical Engineering, Cairo University, Giza, Egypt
[3]L aser in Engineering, NILES, Cairo University
E-mail: Msm_eng@k-space.org

Abstract- **In this paper, we investigate the chaotic behavior of the biological sequences among the different species. Throughout this work, we have characterized the biological sequences according to their moment invariant, correlation dimension, and largest Lyapunov exponent estimates. We have applied our model to a number of human and mouse genomes encoded into a set of integers (time series) using a plain table mapping scheme. Our results indicate that the nonlinear dynamical characteristics have yielded significant differences between the sequences of the different species. That is, we have been able to classify the different genome sequences according to their chaotic parameters estimates. On the other hand, through our investigation we have found that the use of the chaotic modeling of the biological sequences could open new frontiers in the sequence similarity search techniques.**

Keywords **- Chaos theory, Lyapunov exponents, correlation dimension, bioinformatics, genome modeling**

I. INTRODUCTION

The goal of biological sequence alignment is to identify regions of similarity (often interpreted as homology) between two or more sequences, and associate these regions with one another to enable further comparisons. Existing algorithms for sequence alignment and validation are adequate for many problems as even if a complete solution for sequence alignment were available, mathematical or statistical optimality and biological optimality are not equivalent, due to the inevitable violations of implicit or explicit evolutionary models [7],[8].

These challenges have led to the development of new methods for similarity detection; in this work, the genome sequence was encoded to a nonlinear dynamical time series (signal) for feature extraction by different techniques. Such techniques work by transforming the mostly qualitative diagnostic criteria into a more objective quantitative signal feature classification problem. Classical techniques have been used to address this problem such as the similarity detection using the autocorrelation function [1], using frequency domain features [2], time frequency analysis [3], and wavelet transform [4], [5]. Other techniques used adaptive filtering [6], sequential hypothesis testing [7],[8], as well as morphological features. Even though fairly good results have been obtained using such techniques, they seem to provide only a limited amount of information about the signal because they ignore the underlying nonlinear signal dynamics.

In the last two decades, there has been an increasing interest in applying techniques from the domains of nonlinear analysis and chaos theory in studying biological systems [9]. In the field of chaotic dynamical system theory, several features can be used to describe system dynamics including moment invariants, correlation dimension (D2) and Lyapunov exponents. In this work, these features have been used to explain different genome sequences encoded to its signal behavior by several studies [12]. In this paper, we address the problem of characterizing the nonlinear dynamics of our sequence. The implementation details to automatically compute three important chaotic system parameters namely, the moment invariants, correlation dimension and largest Lyapunov exponent, are discussed using the Open TSTool MATLAB package. The proposed implementations were used to compute these features for a twenty independent sequence encoded time series signals belonging to two different genomes: the human and mouse genome, downloaded from the Matrix Science - Help - Sequence Database Setup - IPI [10]. The results are studied to detect statistically significant differences among different genome types. Finally, statistical classification techniques are used such as K-means clustering to assess the possibility of similarity detection and classification using such parameters.

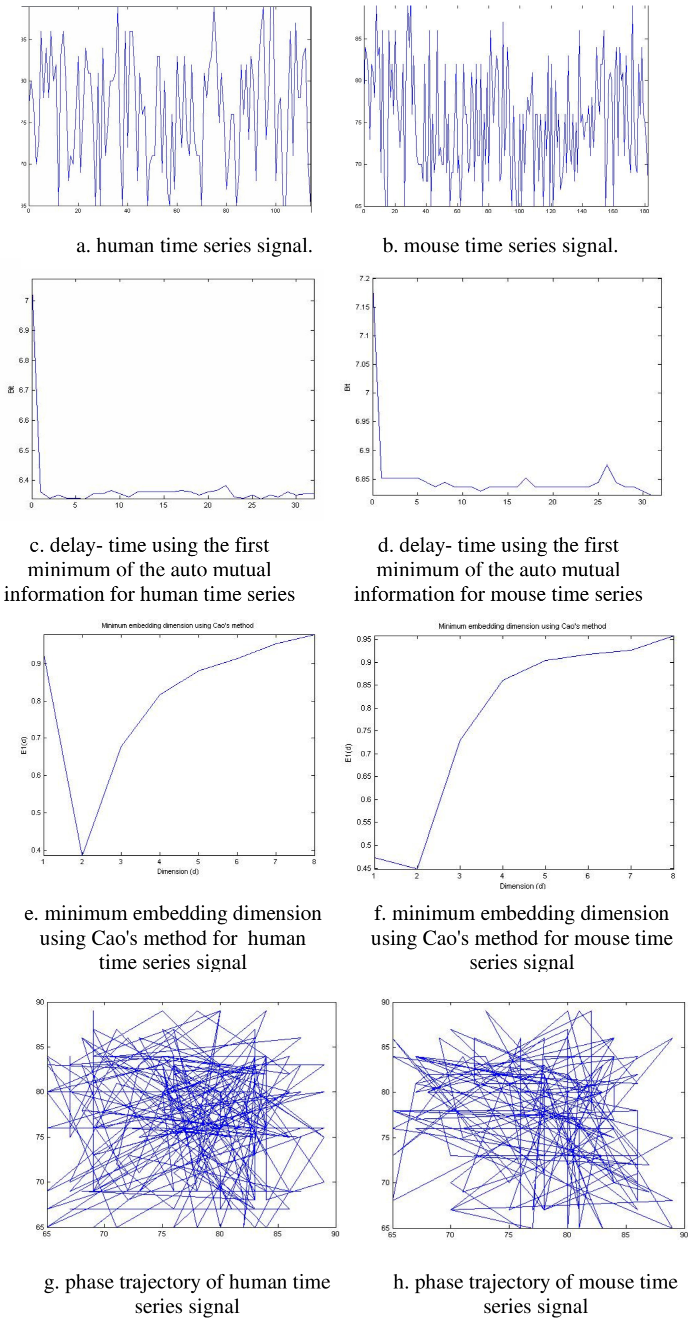
II. METHODOLOGY

A. Phase Space Trajectory Reconstruction.

In this section we briefly demonstrate basic steps for chaotic time series analysis. We start first by encoding the different genome sequences into a time series signal as shown in figure (1.a, b). A good choice for a delay time is yielded by using the first minimum of the auto mutual information function as shown in figure (1.c, d). The first minimum of the auto mutual information can be found at four. Now we need to know the minimal embedding dimension for both human and mouse time series signals. We use Cao's method with a delay time of four, a maximal dimension of eight, three nearest neighbors and reference point depending on the length of each signal. There is a kink in the graph shown in figure (1.e, f) produced by Cao's method at three. So we need a time delay reconstruction of human and mouse time series signals with embedding dimension 3 and delay 4. Finally we plotted the phase space trajectory for both human and mouse time series signals as shown in figure (1.g, h). The step following obtaining the phase trajectory of both human and mouse time series signal is the step of feature extraction. This can be done by applying the following three methods:

1. Moment invariants.
2. Correlation dimension.
3. Lyapunov exponent.

1) Moment Invariants: The mathematical description of a dynamical system consists of two parts: the *state* which is a snapshot of the process at a given instant in time and the *dynamics* which is the set of rules by which the states evolve over time. To study the dynamics of our system, we first need to reconstruct the state space trajectory.



a. human time series signal.

b. mouse time series signal.

c. delay- time using the first minimum of the auto mutual information for human time series

d. delay- time using the first minimum of the auto mutual information for mouse time series

e. minimum embedding dimension using Cao's method for human time series signal

f. minimum embedding dimension using Cao's method for mouse time series signal

g. phase trajectory of human time series signal

h. phase trajectory of mouse time series signal

Figure (1), basic steps of analyzing a chaotic time series system.

The most common method to do this is using delay time embedding theorem to create a larger dimensional geometric object by embedding into a larger m-dimensional embedding space [1]. The embedding dimension m must be large enough for delay time embedding to work. When a suitable m value is used, the orbits of the system do not cross each other. The dimension m in which false neighbors disappear is the smallest dimension that can be used for the given data. The data is ready now for feature extraction by the moment invariants. These invariants are constructed using the generalized fundamental theorem of moment invariants (GFTMI), which was formulated [1]. In 1962, Hu [2] presented the fundamental theorem, of moment invariants (FTMI) for recognition of two dimensional images, subjected to general linear transformation. Only in 1991, after 21 years of publication [3], the CFTMI was formulated by another author [4]. Features obtained by moment invariants are simple calculated features that do not change under translation, scaling or rotation. The following equations calculate the seven features extracted from the ten human and mouse time series signals.

$$m_{P1\ldots Pn} = \int_{-\infty}^{\infty}\ldots\int_{-\infty}^{\infty} x_1^{P1}\ldots x_n^{Pn} p(x)\,dx1\ldots dx_n \qquad (1)$$

The central moments:

$$\mu_{P1\ldots Pn} = \int_{-\infty}^{\infty}\ldots\int_{-\infty}^{\infty} (x_1 - \bar{x}_1)^{P1}.\,(x_n - \bar{x}_n)^{Pn}\, p\,(x)\,dx1.dx_n \qquad (2)$$

$$\text{Where } \bar{x}_1 = \frac{m_{1\ldots0}}{m_{0..0}}, \ldots\ldots \bar{x}_n = \frac{m_{0\ldots1}}{m_{0\ldots0}} \qquad (3)$$

The seven features for moment invariants

$$\varphi_1 = \frac{1}{\mu^{n+2}} \begin{vmatrix} \mu_{2\ldots0} & \cdots & \mu_{1\ldots1} \\ \cdots & \cdots & \cdots \\ \mu_{1\ldots1} & \cdots & \mu_{0\ldots2} \end{vmatrix} \qquad (4)$$

$$\varphi_2 = \frac{1}{\mu^4}(\mu_{20}\mu_{02} - \mu^2_{11}). \qquad (5)$$

$$\varphi_3 = \frac{1}{\mu^{10}}((\mu_{30}\mu_{03} - \mu_{21}\mu_{12})^2 - 4(\mu_{30}\mu_{12} - \mu^2_{21})(\mu_{21}\mu_{03} - \mu^2_{12})). \qquad (6)$$

$$\varphi_4 = \frac{1}{\mu^6}(\mu_{40}\mu_{04} - 4\mu_{31}\mu_{13} + 3\mu^2_{22}). \qquad (7)$$

$$\varphi_5 = \frac{1}{\mu^9}(\mu_{40}\mu_{22}\mu_{04} + 2\mu_{31}\mu_{22}\mu_{13} - \mu_{40}\mu^2_{13} - \mu^2_{31}\mu_{04} - \mu^3_{22}). \qquad (8)$$

$$\varphi_7 = \frac{1}{\mu^5}(\mu_{200}\mu_{020}\mu_{002} + 2\mu_{110}\mu_{101}\mu_{011} - \mu_{200}\mu^2_{011} - \mu^2_{110}\mu_{002} - \mu^2_{101}\mu_{020}). \qquad (9)$$

$$\varphi_8 = \frac{1}{\mu^7}(\mu^2_{20}\mu_{04} - 4\mu_{20}\mu_{11}\mu_{13} + 2\mu_{20}\mu_{02}\mu_{22} + 4\mu^2_{11}\mu_{22} - 4\mu_{11}\mu_{02}\mu_{31} + \mu^2_{02}\mu_{40}). \qquad (10)$$

2) *Correlation Dimension Estimation*: The correlation dimension provides a straightforward way to measure the spatial organization and hence the predictability (finite dimensionality) and

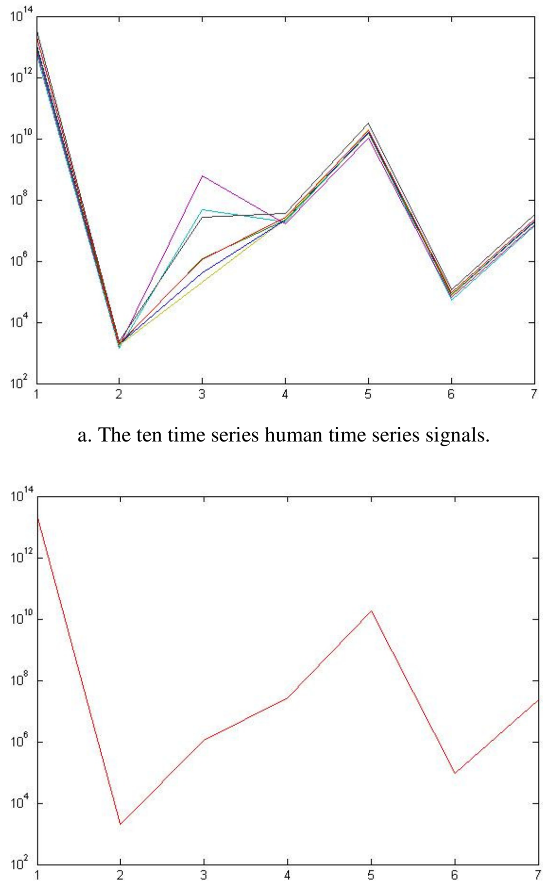
dimensionality of a given signal (time series). That is, the measure of correlation dimension provides a way to determine whether the signal (time series) has fractional dimension i.e. chaotic attractors. However, in order to estimate the correlation dimension it is required to reconstruct the state space trajectory of the time series. This can be accomplished using the delay time embedding theorem through the creation of larger dimensional geometric object by embedding into a larger m-dimensional embedding space. The embedding dimension m should be large enough for delay time embedding to work. When a suitable value for m is used, the orbits of the system don't cross each other [1]. However, we have selected the first minimum of the mutual information function as a suitable value for the embedding time lag. The embedding dimension has been estimated using the Cao's method [6]. Nevertheless, we have computed the correlation dimension using Taken's estimator provided with the TSTOOL add-on toolbox for MATLAB.

3) *Lyapunov exponent*: The notion of Lyapunov exponents is a generalization of the idea of the eigenvalues as a measure of stability of a fixed point (characteristic exponent) as it provides a measure of stability of a periodic orbit. That is, Lyapunov spectrum (exponents) characterizes the behavior (contraction or expansion) of the trajectories close to a fixed point. Therefore, these exponents provide a mean to measure the sensitivity to perturbed initial conditions. For a system to undergo chaotic dynamics, it must have at least one positive Lyapunov exponent. The largest Lyapunov exponent (lambda1), nevertheless, may be regarded as an estimator to the dominant chaotic behavior of the system [1]. However, in this work we have used the TSTOOL largest lyapunov estimation algorithm. This algorithm is similar to Wolf's algorithm and provides an efficient estimation of the largest lyapunov exponent through the calculation of the scaling (rate of increase) of the prediction error (separation of nearby trajectories) versus the prediction time.

III. Results

A. Moment Invariants

We have applied moment invariants feature extraction method to a twenty human and mouse sequences encoded time series signals for feature extraction, these signals are plotted in figure (2). Figure (2.a), shows the ten human time series signals and figure (2.b), shows the other ten time series mouse signals. It is clear from figure (2.b) that the ten mouse signals are very similar but not identical. As described in the methodology section after calculating the seven features extracted by moment invariants, the mean of both human and mouse features is taken and plotted versus each other as shown in figure (3). It is clear from this figure, that the third feature is the most discriminate feature.



a. The ten time series human time series signals.

b. The ten time series mouse time series signals.

Figure (2), the twenty human and mouse time series signals

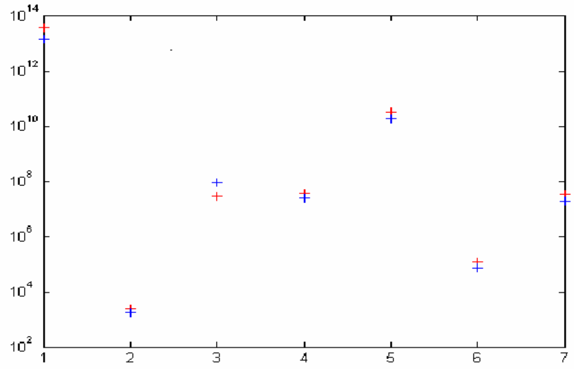


Figure (3), a comparison between μ_h and μ_m.

B. Correlation Dimension

We have applied the Taken's estimator to the encoded sequences in order to obtain their correlation dimension estimates. The following table renders the fractional (chaotic) correlation dimension estimates of a set of human genes and mouse genes obtained from [10].

Table (1), Correlation Dimension (D2) estimates of the translated gene sequences dataset.

Human D2	Mouse D2
4.7579	3.7796
4.8145	4.2047
3.2511	6.6643
4.7954	4.2047
4.8337	4.5172
3.25	2.6328
10.7250	3.5917
7.4067	3.2439
15.0129	5.1459
15.7937	10.4523

We have used these estimates in order to classify the encoded sequences into their respective genomes using the K-means clustering algorithm. The classification accuracy using the correlation dimension estimates is depicted in table 3.

C. Largest Lyapunov Exponent

The largest Lyapunov exponent (LLE) has been estimated manually from the scaling (linear increase) of the prediction error versus the prediction time using the TStool algorithm. The following table depicts the LLE estimates of a set of human genes and mouse genes.

Table (2), Largest Lyapunov Exponent (labda1) estimates of the translated gene sequences dataset.

Human LLE	Mouse LLE
0.000778 (D2=4.7579)	0.003 (D2=3.7796)
0.0003347 (D2=4.8145)	0.001 (D2=6.6643)
0.0001666 (D2=3.2511)	0.00146 (D2=6.66)
0.000137 (D2=3.25)	0.0010398 (D2=10.4523)
0.000428 (D2= 4.7954)	0.00078 (D2=3.2439)
0.0001544 (D2=4.8337)	0.000493 (D2=3.5917)
0.00045 (D2=10.7250)	0.00033 (D2=4.5172)
0.000406 (D2=7.4067)	0.0005 (D2=3.59)
0.000262 (D2= 15.7937)	0.00140 (D2=6.6)
0.000137 (D2=15.0129)	0.0030 (D2=3.77)

Furthermore, we have applied the K-means clustering algorithm to classify the sequences into their respective genomes. The classification accuracy using the LLE is depicted in table 3.

Table 35), accuracy of the proposed feature extraction methods to K- means Clustering classifier.

	Moment invariants	Correlation dimension	Lyapunov exponent
Human	80%	40%	100%
Mouse	100%	80%	60%

IV. Discussion and Conclusion

In this work, we have characterized the biological sequences based on their nonlinear dynamical behavior. That is, we have established a nonlinear dynamical model consists of moment invariant, correlation dimension (D2), largest Lyapunov exponent (lambda1) estimates of plain integer mapping encoded sequences. The pattern of this model's parameters has varied considerably between the different genomes. Furthermore, we have used the K-means clustering algorithm in order to classify the different sequences into their respective genomes.

Experiments were performed on a dataset obtained from [10] to evaluate the reliability of the proposed nonlinear dynamical model. The proposed model has yielded reasonable classification accuracy between the human sequences and the mouse sequences. Nonetheless, due to the existence of similarities between some of the human sequences and the mouse sequences, our model has yielded low classification accuracy in some cases. Therefore, it is required to use longer sequences (more than 300 bases) in order to enhance the performance of our proposed model for the similar sequences.

In conclusion, throughout this work we have found that the natural nonlinear dynamics that the biological sequences undergo differ between the different species. Therefore, it is rather encouraging to distinguish between the different species according to the nonlinear dynamical characteristics of their respective translated gene sequences.

References

[1] M. I. Owis, A. H. Abou-Zied, A. M. Youssef, and Y. M. Kadah, "Study of features based on nonlinear dynamical modeling in ECG arrhythmia detection and classification," *IEEE. Trans. Biomedical Engineering.* vol. 79, pp. 733-736, July 2002.

[2] M. T. Rosenstein, J. J. Collins, and C. J. De Luca, "A practical method for calculating largest Lyapunov exponent from small datasets," *Physica D* **65**:117-134, 1993.

[3] M. Small and K. Judd, "Detecting nonlinearity in experimental data," *Intl. Journal. Bifurcation and chaos.* vol. 8, No.6 pp. 1231-1244, 1998.

[4] E. Ott, *Chaos in Dynamical Systems,* 1st ed., Cambdridge university press: Cambridge, 1994.

[5] T. Kapitaniak, *Chaos for Engineers,* 2nd ed., Springer: New York, April 2000.

[6] L. Cao, A. Mees, K. Judd, and G. Froyland, "Determining of the minimum embedding dimensions of input-output time series data," *Intl. Journal. Bifurcation and chaos.* vol. 8, pp. 1491-1504, 1997.

[7]Notredame,C.(2002) Recent progress in multiple sequence alignment: a survey. pharmacogenomics,3,131-144.

[8]Sullivan,J. and Swofford, D.L. (2001) Should we use model- based methods for phylogenetic inference substitution pattern are violated? *Syst. Biol.*, 50, 723-729.

[9]Guillén, S. G. Arredondo ,M. T., Martin G., and Corral J. M. F. (1989), Ventricular fibrillation detection by autocorrelation function peak analysis, *J. Electrocardiol.*, vol. suppl. 22, pp. 253–262.

[10]ftp://ftp.ebi.ac.uk/pub/databases/IPI/current

[11]Minami K, Nakajima H., and Toyoshima, T. (1999) Real-time discrimination of ventricular tachyarrhythmia with Fourier-transform neural network, IEEE *Trans. Biomed. Eng.*, vol. 46, pp. 179–185.

[12]Afonso V. X. and Tompkins W. J. (1995) Detecting ventricular fibrillation, selecting the appropriate time-frequency analysis tool for the application *IEEE Eng. Med. Biol. Mag.*, pp. 152–159.

[13]Khadra L., Al-Fahoum A. S., and Al-Nashash H. (1997) Detection of lifethreatening cardiac arrhythmias using thewavelet transformation," *Med. Biolo. Eng. Comput.*, vol. 35, pp. 626–632.

[14]Al-Fahoum A. S. and Hewitt, I. (1999) Combined wavelet transformation and radial basis neural networks for classifying life-threatening cardiac arrhythmias,
Med. Biolo. Eng. Comput., vol. 37, pp. 566–573.

[15]Thakor N. V. and Zhu Y. (1991) Applications of adaptive filtering to ECG analysis: Noise cancellation and arrhythmia detection, IEEE *Trans. Biomed. Eng.*, vol. 38, pp. 785–794.

[16]Thakor N. V., Zhu Y., and Pan K. (1990) Ventricular tachycardia and fibrillation detection by a sequential hypothesis testing algorithm," *IEEE Trans. Biomed. Eng.*, vol. 37, pp. 837–843.

[17]Thakor N. V., Natarajan A., and Tomaselli,G. F. (1994) Multiway sequential hypothesis testing for tachyarrhythmia discrimination," *IEEE Trans. Biomed. Eng.*, vol. 41, pp. 480–487.

[18]Abarbanel H. D. I, Brown R., Sidorowich J. J., and Tsimiring L. S.(1993),The analysis of observed chaotic data in physical systems, *Rev. Mod*
Phys., vol. 65, pp. 1331–1392.

[19]Casaleggio A. and Braiotta S. (1997) Estimation of Lyapunov exponents of ECG time series—The influence of parameters, *Chaos, Solitons Fractals*,
vol. 8, no. 10, pp. 1591–1599.

[20]Pritchard W. S. and Duke D.W (1995) Measuring chaos in the brain: A tutorial review of EEG dimension estimation, *Brain Cogn.*, vol. 27, no. 3, pp. 353–397.

[21]Mamistvalov A.G. (1974) On the Construction of Affine Invariants of n-Dimensional Patterns," *Bull. Acad. Science Georgian SSR,* vol. 76, no. 1, pp. 61-64.

[22]Hu M.K. (1962) Visual Pattern Recognition by Moment Invariants *IRE Trans. Information Theory,* vol. 8, pp. 179-187.

[23]Reiss T.H. (1991) the Revised Fundamental Theorem of Moment Invariants *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 13, no. 8, pp. 830-834.

[24]Alexander G. (1998) n-Dimensional Moment Invariants and Conceptual Mathematical Theory of Recognition n-Dimensional Solids IEEE Transactions on Pattern Analysis and Machine Intelligence , vol 20, No. 8.