Numerical Characterization of Gene Sequences Based on Chaos Principle

Sunita, Amrita Priyam, Parikshit Munda

Abstract— In this paper new approach is being proposed to numerically represent a gene. This can be further used to build a model that can search a gene more accurately and in less time than present methods. The method consists of three parts: primary sequence was reduced to a few of binary sequences, based on the classifications of the four nucleic acid bases. Then, by using a encoding rule, binary sequences were converted into DNA signal, which were used as input vector to calculate six embedded phase space fractal dimension(PSFD) as new invariants for the DNA primary sequences. Using these invariants, similarities among the primary sequences for one gene belonging to 10 different species is being computed.

Index Terms— chaos, fractals, genes, phase space,dna

I. INTRODUCTION

With the completion of Human Genome Project and fast increase of many complete genomes of prokaryotes and eukaryotes, there is sudden increase in number of genes and hence, a huge database of genes is present. Hence, it is becoming a difficult and time consuming task to search a gene of interest from this database. Also, finding sequence similarity to a gene is a important factor in finding the clue regarding functionality of a gene.

There are number of methods present to search for homologous sequences like Dynamic Programming method (Needleman-Wunsch, Smith-Watermann) and heuristics method like FASTA and BLAST. The dynamic programming method is guaranteed to find an optimal alignment given a particular scoring function; however, identifying a good scoring function is often an empirical rather than a theoretical matter. Although dynamic programming is extensible to more than two sequences, it is prohibitively slow for extremely long sequences. [1] Blast, although is fast and most widely used, but it may not give optimal alignment. Also, with increase in size of database, its search time is increasing considerably.

To calculate Motivated by these facts, we designed a method to numerically represent a gene, that can be further used to build a model that can search a gene more accurately and in less time than present methods. So, keeping these things in mind, we first mapped the string representation of gene into binary one using three rules[2], a)purine and pyrimidine classification of the four bases in DNA sequence, b) amino group and keto group; c)the weak hydrogen bond and the strong hydrogen bond. We give three kinds of mapping from four bases {A,C,G,T} in DNA sequence to the number set {0, 1}.

Manuscript Received May, 2013.

Sunita, Department of Computer Science and Engineering, C.I.T Tatisilwai Ranchi, India.

Dr. Amrita Priyam, Department of Computer Science and Engineering, B.I.T Ranchi, India.

Parikshit Munda, Department of Computer Science and Engineering, C.I.T Tatisilway, Ranchi, India.

Using each of three mappings, we convert a DNA sequence into three binary sequences, then calculated distance between two consecutive one's as input vector six embedded phase space fractal dimension(PSFD)[3] as new invariants for the DNA primary sequences. Using these invariants, we compare the similarities among the primary sequences for one gene belonging to 10 different species.

II. CHAOS THEORY AND DNA

A. Introduction

Chaos ,an Ancient Greek term, mainly refers to the state lacking order or predictability. So, chaos theory is mainly related to the study of behaviour of dynamical systems, which are extremely sensitive to initial conditions. A small change in the initial condition may yield very widely diverging outcomes, that make long-term prediction of the chaotic systems impossible.[4] In case of biology, we can see chaos in many systems, like medical study of epilepsy, in study of ECG, etc.

B. Chaos within DNA

It is an open question whether DNA sequence follows a chaos or not. In case of DNA sequence, it is biologically proven that nucleotides are not added in DNA in a random manner, but in a deterministic manner. So, presence of deterministic chaos property in DNA sequence cannot be ruled out, and this property can be used to characterise the nature of DNA sequence. Therefore, surrogate and predictability analysis was done, but still it did not give enough evidence, so that we can definitely say that DNA has chaos feature or not.

III. FRACTAL THEORY

A. Introduction

Fractal can be defined as a rough or a fragmented geometric shape that can be divided in many subparts, and subpart is approximately a compressed size copy of the whole. The strange attractor is one of the example of a fractal. Fractals are fundamental to chaos theory. Fractals allow scientists to study behavior of the systems in a way that is much more easy to understand in comparison to numbers. A particular statistical formula can be used to describe the fractal geometry which behaves chaotic in nature[5]

B. Phase Space Fractal Dimension

Fractal dimension can be defined as a measure of the complexity of self-similar object. We can also say that it measures 'how many points are there, which lie on a given set'. But here in case of DNA signal, fractal dimension is not actual a fractal dimension but it is a phase space fractal dimension, which may or may not be a integer value[6]. The nearest integer value tells the number of variables and the embedded dimensions that are

needed to represent the relevant system. Phase space is a space where all the states of a system can be represented. The phase space for a random signal gives some random graph with fractional dimension, Fd as Fd = ∞ (infinite).

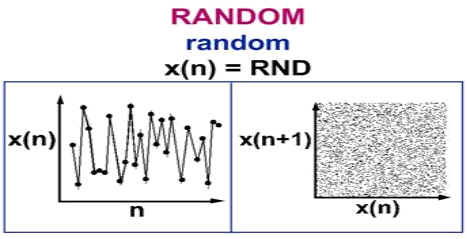

Figure 3.1: Phase Plot for random signal [7]

Phase space for a chaotic signal gives some deterministic type of graph and fractal dimension,Fd is Fd =1

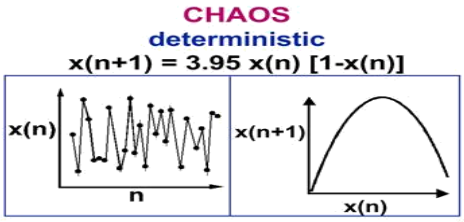

Figure 3.2: Phase Plot for Chaotic Signal [7]

It is an open question whether DNA sequence follows a chaos or not. In case of DNA sequence, it is biologically proven that nucleotides are not added in DNA in a random manner, but in a deterministic manner. So, presence of deterministic chaos property in DNA sequence cannot be ruled out, and this property can be used to characterise the nature of DNA sequence. Therefore, surrogate and predictability analysis was done, but still it did not give enough evidence, so that we can definitely say that DNA has chaos feature or not.

IV. MATERIALS AND METHODS

All gene sequences were taken from KEGG[8] database . 10 Gene sequences of adenosine deaminase(ADA) from 10 different organisms were taken along with their BLAST scores.

A. *Conversion of Neucleotide Sequence into Binary String*

As we know gene sequence is composed of 4 nucleotide bases commonly known as A, C, G, and T.Acording to their Chemical properties we grouped these bases into three pairs: purine R = {A, G} and pyrimidine Y = {C, T}, or amino group M = {A,C} and keto group K = {G, T} and according to the strength of the hydrogen bond, *i.e*., weak H–bonds W = {A, T} and strong H–bonds S = {G, C}.

Let, $S=n_1\ n_2\ n_3\ n_4\ n_5$..... be a gene sequence. Using above criteria, we can transform a primary nucleotide sequence into three Binary(1,0) sequences by three homomorphic maps $Ø_i$, i=1,2,3,4,5, $Ø_i(S)= Ø_i(n_1)Ø_i(n_2)Ø_i(n_3)$……., as follows[2].

$Ø_i$ = 1 if $n_1 \in R$
0 if $n_1 \in Y$
$Ø_i$ = 1 if $n_1 \in M$
0 if $n_1 \in K$
$Ø_i$ = 1 if $n_1 \in W$
0 if $n_1 \in S$

In this way, we converted each gene sequence into corresponding three binary strings.

B. *Conversion of Binary String into DNA Signal*

Encoding rule used for conversion of Binary String into DNA signal is distance between consecutive 1's present in Binary string. So, we get three signal for each gene sequence.

e.g. S = A T T G C C T G C C
v1= 1 0 0 1 0 0 0 1 0 0
v2= 1 0 0 0 1 1 0 0 1 1
v3= 1 1 1 0 0 0 1 0 0 0

s1= 1 3 4
s2= 1 4 1 3 1
s3= 1 1 1 4

C. *Compution of Phase Space Fractal Dimension (PSFD)*

Phase Space of a signal is defined as a space where all states of signal can be represented. Let a DNA signal be represented by X = (x_1 ,x_2, x3 ,…… x_n), then phase space for the signal in two dimensional embedding space contains co-ordinates of the points [(x_1 ,x_2), (x_2 ,x_3), (x_3 ,x_4),…….(x_{n-1} ,x_n)].

Similarly, in 3-dimensional embedding space, phase space of signal contains co-ordinates of points [(x_1 ,x_2, x_3), (x_2 ,x_3, x_4), (x_3 ,x_4, x_5),…….(x_{n-2},x_{n-1} ,x_n)], and so on. A method was developed by Grassberger and Procaccia [6] by using correlation which gives approximate fractal dimension value.

PSFD calculated by correlation integral CI, is given by

$$C_m(r)=(1/N^2)\sum_{i,j=1;i\neq j}^{N} Z(r-|X_I-X_j|)$$

Where

$$Z(x)=\begin{cases}1; \text{if } r-|X_i-X_j|>0\\ 0; \text{if } r-|X_i-X_j|\leq 0\end{cases}$$

Where Cm(r) is CI for an embedding dimension, m; Z(X) is Heaviside function.

If |Xi – Xj| is less than r, then it is 1 otherwise 0, as r is kept between 1 and 0. Cm(r) increases with increase in r as

$$Cm(r) = \pm r^D$$

Taking log on both sides,

$$Log\ Cm(r) = D*log(r) + constant$$

One can calculate Cm(r) having a dimension m for increasing value of r. Correlation dimension D is estimated using regression for embedded dimension, by calculating slope of graph line of log(r) and log[Cm(r)]. Value of D will be closer to true value if value of m is increased.

For initial checking of tendency of each data towards the chaos process, phase-space pattern of each of the data-type was studied for even embedding dimension. Finally, 6th embedding dimension was heuristically selected for calculating PSFD of each of the data-type. A set of 6 fractal dimension were obtained for each vector and hence, total 18 fractal dimensions were calculated for each sequence. Below is the table containing

18 fractal dimension values for each sequence:

		1	2	3	4	5	6
Sequence 1	v1	0.2958	0.436	0.5732	0.6919	0.819	0.9385
	v2	0.3197	0.4774	0.6368	0.7996	0.9588	1.1115
	v3	0.4282	0.6198	0.7982	0.9608	1.1435	1.3531
Sequence 2	v1	0.295	0.4343	0.572	0.691	0.8181	0.9399
	v2	0.3211	0.4784	0.6351	0.7942	0.9446	1.0879
	v3	0.4283	0.6202	0.8053	0.9841	1.173	1.3718
Sequence 3	v1	0.2973	0.4374	0.5795	0.7026	0.8323	0.9561
	v2	0.328	0.489	0.649	0.8061	0.9606	1.114
	v3	0.4326	0.6221	0.8075	0.9685	1.1261	1.2915
Sequence 4	v1	0.2985	0.4386	0.5828	0.7222	0.8518	0.983
	v2	0.3214	0.4797	0.6328	0.776	0.9117	1.0577
	v3	0.4354	0.6236	0.8022	0.9574	1.1114	1.3061
Sequence 5	v1	0.2817	0.4144	0.5527	0.6821	0.8098	0.9401
	v2	0.305	0.4698	0.6333	0.7923	0.9621	1.124
	v3	0.45	0.6561	0.8494	1.0244	1.21	1.4121
Sequence 6	v1	0.281	0.4094	0.5453	0.6691	0.7912	0.9222
	v2	0.333	0.502	0.66	0.8144	0.9629	1.1162
	v3	0.4496	0.6429	0.8151	0.9775	1.1551	1.3155
Sequence 7	v1	0.278	0.4069	0.5302	0.6375	0.7506	0.8595
	v2	0.3178	0.4701	0.6206	0.7661	0.9012	1.0339
	v3	0.4187	0.6054	0.7894	0.9837	1.1923	1.4552
Sequence 8	v1	0.2857	0.4191	0.5461	0.6486	0.7437	0.8411
	v2	0.3122	0.4618	0.6067	0.7473	0.8778	1.0152
	v3	0.3781	0.5417	0.7017	0.8612	0.9972	1.1247
Sequence 9	v1	0.2672	0.3888	0.5179	0.6232	0.7437	0.8705
	v2	0.3187	0.4907	0.6553	0.8146	0.9523	1.1374
	v3	0.4193	0.584	0.7182	0.8531	0.9505	1.0251
Sequence 10	v1	0.237	0.3559	0.4812	0.6026	0.7218	0.8312
	v2	0.3464	0.5198	0.6905	0.8379	0.9812	1.1116
	v3	0.2955	0.4692	0.6438	0.8115	0.9961	1.1798

TABLE 1:- Total 18 (6 each for a distance vector) embedded PSFD for each gene sequences.

	Nsimil_ score	**NW_ score**	**BLAST_ score**
Seq 1	100	2184	2165
Seq 2	99.644	2124	2070
Seq 3	99.491	2039	1935
Seq 4	99.198	1517	1116
Seq 5	99.072	1396	1053
Seq 6	99.378	1471	1041
Seq 7	98.571	1162	720
Seq 8	97.261	1062	569
Seq 9	96.854	374	293
Seq10	96.802	239	182

Table 2: Scores calculated using different methods for 10 gene sequences taken from KEGG database.

D. Calculation of Similarity using normalized RMSD

Using 18 PSFD values of each sequence, similarity was calculated by Normalised RMSD using formula

$$\text{RMSD} = \sqrt{\left(\sum_{i=1}^{18}\left((x_i - y_i)2/(x_i 2 + y_i 2)\right)\right)/18}$$

Similarity(nsimil_score) = 100*(1-RMSD)

nsimil_score	100	99.644	99.491	99.198	99.072	99.378	98.571	97.261	96.854	96.802
NW_score_m	2962	2927	2881	2572	2451	2534	2257	2185	13	1393
NW_score_n	2184	2124	2039	1517	1396	1471	1162	1062	-374	239
Blast_score	2165	2070	1935	1116	1053	1041	720	569	293	182

Where,

nsimil_score = similarity score calculated by normalized RMSD

NW_score_m = Needleman-wunsch score calculated using MATLAB

NW_score_n = Needleman-wunsch score calculated using NCBI

Blast_score = Score given by BLAST algorithm

V. RESULT AND DISCUSSION

Correlation coefficient between our score and Needleman_wunsch score is 0.92047, and correlation coefficient between BLAST score and Needleman_wunsch score is 0.907238.

So, by comparing both correlation coefficient we can say that our score is more accurate than BLAST score. Hence this numerical representation can be used to incorporate in any similarity search engine to get more accurate and fast results.

VI. APPLICATION: DESIGN OF SEARCH ENGINE

As current search engine relies on string representation of data, these methods are very slow. After incorporating heuristics in these methods like in BLAST and FASTA, search time have been reduced to a greater extent. But, with the increase in size of database day by day, they are taking a considerable amount of time.

So, to reduce the amount of search time, we have to represent the data in numerical form and then these data can be clustered to partition the search space and search time can be reduced considerably.

Suppose we have 1 billion(N) sequences in our database. Time complexity for Dynamic Programming method (Needleman-wunsch method) would be N^2 (1,000,000,000 X 1,000,000,000), a huge amount of time. Again, time complexity for heuristics method like BLAST would be N (1,000,000,000), a considerable amount of time. So, if we can cluster the data in 10 cluster than we would be having 100million sequences in each cluster. Again, if we will further cluster the 100 million sequences in 10 clusters we would be having, 10 clusters of 10 million data. Again, in the same way if we will cluster the 10million data cluster into 10 clusters we would be having 10 clusters of 1 million each. And, if carry on like this we would be having a final cluster of 10 data only.

So, total computation required to search a sequence would be 10 for first 10 clusters, again 10 for 10 clusters in next level and so on. So, total computation would be 10+10+10+10+10+10+10+10+10+10 i.e. 100 computations (far less than even BLAST).

So, reducing search time to this extent, we have to partition the data, and for partitioning of data, numerical representation of data is required. Yet now, no efficient numerical representation is present for nucleotide data. So, our work is actually done to keep this thing in mind.

This concept can be incorporated in present search engines to considerably reduce the search time without compromising with efficiency.

V. CONCLUSION

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

This pilot study gives promise for representing genes sequences into numerical form so that we can considerably reduce the search time and space. We have tried to apply our method to a group of similar genes, and it shows that are method is much more accurate than present heuristic methods like BLAST and FASTA and will be more much more faster than Needleman-Wunsch and Smith-Waterman algorithm. This method will be more successful in finding the very similar gene sequences that many times missed by these heuristic methods. The good thing about this method is numerical representation and this numerical representation can be further used to partition search space and hence to reduce search time.

REFERENCES

[1] Eddy, S. R., What is dynamic programming? Nature Biotechnology, 22, 909-910 (2004)

[2] He P, Wang J, Numerical Characterisation of DNA Primary Sequence, Internet Electronic Journal of Molecular Design 2002, 1, 668-674

[3] Lahiri T, Kumar U, Mishra H, Subrata S, Roy A D, Analysis of ECG signal by chaos principle to help automatic diagnosis of myocardial infarction, Journal of Scientific & Industrial Research, 68, 886-870 (2009).

[4] Stephen H. Kellert, In the Wake of Chaos: Unpredictable Order in Dynamical Systems, University of ChicagoPress, 1993,p32, ISBN0-226-42976-8.

[5] http://en.wikipedia.org/wiki/Fractal.

[6] Yiming WEI, Ying FAN and Weixuan XU, "Nonlinear Dynamic Analysis for Inundated Area of Flood Disaster in China", Institute of Policy and Management, Chinese Academy of Sciences, P.O.Box 8712, Beijing 100080, and P.R. China, pp.692-696.

[7] Schuster, H.G. "Deterministic Chaos: An Introduction", Physik-Verla GmbH, 1984.

[8] http://www.genome.jp/kegg/genes.html

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication