Journal of Computational Intelligence in Bioinformatics
ISSN: 0973-385X Volume 3 Number 2 (2010) pp. 257–265

http://www.ripublication.com/jcib.htm

Mathematical Lensing of DNA Repeat Sequences

K. Meena[1], K. Menaka[2], T.V. Sundar[3*] and K. R. Subramanian[4]

[1]Vice Chancellor, Bharathidasan University, Tiruchirappalli – 620 024, India
[2]Lecturer, Department of I.T. & Applications, Shrimati Indira Gandhi College, Tiruchirappalli – 620 002, India
[3]Assistant Professor, Post Graduate and Research Department of Physics, National College, Tiruchirappalli – 620 001, India
**E-mail: sundar_vaidyanathan@yahoo.co.in*
[4]Professor, Department of Computer Applications, Shrimati Indira Gandhi College, Tiruchirapalli – 620 002, India

Abstract

The growth pace of tools for the comparative analysis of the intra genomic regions of DNA sequences is very small when compared with the output of the genomic sequence data. This necessitates a search for good and efficient techniques to rapidly scan, collate, analyze and retrieve sequences of interest and enable comparative analysis. Numerical and statistical approaches are promising candidates to face this challenge. The various methods available in matrix algebra, statistical procedures and numerical procedures can effectively be utilized for facing the above challenge. The novel technique of Mathematical Lensing, containing the combined power of matrix, statistical and numerical procedures, proposed and illustrated in this paper is one such attempt in the above direction. The results show compactness for comparative sequence studies and scope for potential applications.

Key Words*:* Matrix Algebra; Mathematical Lensing.

Introduction

Enormous DNA sequencing data have been accumulating on public DNA databases of DDBJ, GenBank and EMBL. The human genome contains approximately three billion base pairs of DNA. Within this there are between 30,000 and 70,000 genes, which together add up to less than five percent of the entire genome. Most of the remaining things are made up of several types of non-coding repeated elements. Most gene sequences are unique, found only once in the genome. In contrast, repetitive

DNA elements are found in multiple copies, in some cases thousands of copies. Unlike genes, most repetitive elements do not code for protein or RNA. Repetitive elements have been found in most other eukaryotic genomes but their functions are mainly unknown. Their presence and spread causes several inherited diseases, and they have been linked to major events in evolution. Hence, finding the reasons for such genetic disorders from rare/common repeat spread of repetitive data is one of the widely attempted problems in genetics.

The relationship, for instance, between related genes of a species and the repeat elements between the genome may serve as useful diagnostic indicators for diseases **[1].** Currently, more information is required with regard to what repeat elements are specific to what diseases and whether this information can be used to predict the disease onset or progression. The parameters, such as, the occurrence of GC content, AT/GC ratio, Dominant Eigen Values (DEV), Best Linear Fit of Data (BLFD) are evaluated by analyzing the sequence data in this work for a set of related repeated DNA sequences pertaining to the rice gene. Moreover, a comparative analysis among the different types of repeat pertained into the same rice gene elements has also been performed in order to elucidate what structural trends or patterns are present in the gene, due to these various DNA repetitive elements.

Methodology

For the analysis, some Tnr category repetitive elements, the transposable elements found in the genome of Oryza sativa (rice) are taken from the Repbase repository **[2]**. The Sequence characteristics of the repetitive elements were analyzed in the numerical domain by a combinatorial technique called ***Mathematical Lensing***, devised by the authors for the first time. The lensing technique is a combination of Matrix Algebra, Statistical Testing and Least Squares method. The inspiration for the nomenclature of this technique is in analogy with the functioning of an optical lens. An optical lens collects beam of light rays in parallel, when source(s) are at infinite distance (here biological sequences) on its one side, selects the portion of the light beam which is incident on its surface (here statistical testing) and allows it to traverse through it. The traversed light is bent by the other curved portion of the lens (here matrix processing), and results in convergence at a point called the focus of the lens. The sharpness and quality of image obtained (here straight line fit of data) obtainable at the focal point, is then used for further optical processing (here sequence analysis). The mathematical steps involved are elaborated below:

Generation of Transition proportion matrix

The concept of transition matrix of a data sequence is described in the book by J. C. Davis **[3]**. The concept was applied to study DNA sequences to study the nature of transitions from one kind of base to another to look out for information by Yu & Chen **[4]**. In their approach, for a given DNA sequence $\mathbf{S} = s_1 s_2 \ldots s_N$, a 4×4 matrix $\mathbf{A} = (\mathbf{t_{ij}})$, where t_{ij} means the number of times a given kind of base being succeeded by another in the sequence, is constructed. **A** is called the transition frequency matrix of **S** and it stands as a concise way of expressing the incidence of one kind of base following

another, irrespective of the length of the sequence. Then the tendency for one kind of bases to succeed another was represented by converting the frequency matrix to decimal fractions or percentages. Next, a matrix $\mathbf{P} = (P_{ij})$ is constructed by dividing each element by the grand total of all entries in **A**. Such a matrix represents the relative frequency of all the possible types of transitions, and is known as the transition proportion matrix of **S**. For example, for a sequence **S** = ATGCATGCA, say, the transition frequency matrix **A** will look like:

	A	G	C	T
A	0	0	0	2
G	0	0	2	0
C	2	0	0	0
T	0	2	0	0

For the above example, the transition proportion matrix **P** will become

	A	G	C	T
A	0	0	0	0.25
G	0	0	0.25	0
C	0.25	0	0	0
T	0	0.25	0	0

The unique advantage of using transition proportion matrix is that one can get normalized data sets(i.e. the grand total of all entries in **P** always add up to one) of the various sequences under comparison, irrespective of the lengths of the individual sequences. Then the real eigen values of the transition proportion matrix **P** of the DNA sequence are computed and tabulated for all the test categories. It is natural that such a parameter is relevant to the system's complexity. The eigen values of a system are a measure of the complexity of it and the corresponding eigen vectors corresponding to an eigen value have spatial significance. The eigen value spectrum obtained is subjected to the statistical procedure of ANalysis Of VAriance (ANOVA) and the F-ratio obtained may be used to decide the consensus nature of the spread of the eigen values and hence that of the sequence repeats also.

Correlation Analysis of the Sequence Data

The Statistical parameter of Correlation coefficient ***r*** or Pearson product-moment correlation coefficient is a measure of the linear relationship between two attributes of data. The value of ***r*** can range from -1 to +1 and is independent of the units of measurement. A value of ***r*** near 0 indicates little correlation between attributes; a

value near +1 (positive correlation) or -1 (negative correlation) indicates a high level of correlation.

When two attributes have a positive correlation coefficient, an increase in the value of one attribute indicates a likely increase in the value of the second attribute. A correlation coefficient of less than 0 indicates a negative correlation. That is, when one attribute shows an increase in value, the other attribute tends to show a decrease.

In the suggested lensing procedure, for the two sequences under study, say **S1** and **S2**, the elements of the corresponding transition proportion matrices **P1** and **P2**, having 16 transition types each, form the data sets for analysis. The sixteen elemental values denote the transition proportion percentages between the base pairs Adenine (A), Guanine(G), Cytosine (C) and Thymine(T) i.e. A→A, A→G, A→C, A→T, G→A, G→G, G→C, G→T, C→A, C→G, C→C, C→T, T→A, T→G, T→C and T→T type transitions of the respective sequences. Moreover they are encoded with the integral values of 1 to 16 for use in linear fit process. Then Correlation coefficient ***r*** is computed using the formula

$$\mathbf{r = (\sum x\, y) / (\sqrt{(\sum x^2)}\, \sqrt{(\sum y^2)})}$$

where **x** and **y** are the deviations of the elements of data sets **S1** and **S2**, measured from the respective mean values of the data sets. Highly correlated sequences are filtered out using the values of **r** and are subjected to numerical characterization. The upper bound in correlation occurs for perfectly correlated sequences and Sakoda coefficient is a parameter for measuring the degree of association between the data sets. It is determined using the relation

$$\mathbf{S = r \sqrt{(n/(n-1))}}$$

Where **n** is the number of elements in the data sets. In this problem as **n** is16 for any length of sequences, **S** can be obtained by multiplying **r** with 1.033.

Linear Fit of Sequence Data and Slope Analysis

The filtered sequences by statistical analysis can be further characterized numerically by linear curve fitting. The line of best fit, of the form $\mathbf{Y = A*X + B}$, can be arrived by the method of Least Squares, as the method provides the technique for minimizing the error in the data sets. The constants **A** and **B** can be determined by solving the normal equations

$$\mathbf{nA + B\sum X = \sum Y}$$
$$\mathbf{A\sum X + B\sum X^2 = \sum XY}$$

The closeness of the slopes of the lines (i.e values of **A**) would serve as a measure of the consensus nature of the repeat sequences.

Results and Discussion

A transposon is a segment of DNA with capability for independently replicating itself and inserting the copy into a new position within the same or another chromosome or plasmid, i.e. it literally "jumps" around from one chromosome to another. They can be problematic, as they may disrupt the normal function of an important gene by their random insertion into the middle of that gene. However, they are found in almost all

organisms, and their presence in a genome indicates that genetic information is not fixed within the genome. For instance, the six transposable elements Tnr4, Tnr5, Tnr11, Tnr12, Tnr13 and RIRE9 were identified as insertion sequences in the terminal inverted repeat sequences (TIRs) of Tnr1 and were analyzed in detail for biological significance **[5]**.

The basic sequence details of the nine Tnr category repetitive sequences tested for illustration of this work are given in Table 1 and graphical representation of the nucleotide transition trends, are shown in figure 1. The trend analysis graphs are plotted for relative percentage variations in the transitions of base pair types versus nucleotide transition types. The trends show similarity in the richness of AT, CA and TA transitions and poorness of AG, GA, GC, CG and TC transitions. The trends are almost symmetrical for AA and TT type transitions while both convexity and concavity are observed for rest of the transition types in the trend lines.

Table 1: List of the general characteristics of some of the TNR category Repetitive sequences found in the genome of Oryza sativa.

S.No.	Repeat Sequence Id.	Total number of bases	GC content	AT / GC ratio
1	TNR2A	152	34.87	1.87
2	TNR2	157	37.58	1.66
3	TNR9	176	42.05	1.38
4	TNR1	237	22.78	3.39
5	TNR8-1	418	28.95	2.45
6	TNR8-21	428	23.83	3.20
7	TNR12	528	28.98	2.45
8	TNR11	811	29.47	2.39
9	TNR3	1536	41.67	1.4

The Guanine-Cytosine (GC) content is an important attribute in genomic studies, as it is used to scan the basic makeup of the genome, as well as understanding coding sequence evolution. There is a vast variety and specificity associated which can be correlated to many important evolutionary trends. The implications have been used by biologists to generate phylogeny trees and to explain certain principles that evolutionary pressures have exerted on these genomic sequences. The genome of the rice *indica* subspecies was analyzed by Yu et al. **[6]**. They found a new kind of fine-scale GC heterogeneity. Their study in detail about GC frequencies, in a collection of rice full-length cDNAs, aligning them to the genome, led to the finding that the genes were richer in GC at the 5’ end than at the 3’ end. The trend was not only observed in the coding sequence but also in introns. As a consequence of these GC gradients, codon and amino acid usage are also affected, showing 5’ to 3’ gradients. Here, in the annotated repeat sequence data analyzed one can see GC content fluctuations ranging

between 22.78, for TNR1 and 42.05, for TNR9. For all the sequences, the AT/GC ratio was greater than unity, with a minimum of 1.38 for TNR9 and a maximum of 3.39 for TNR1.

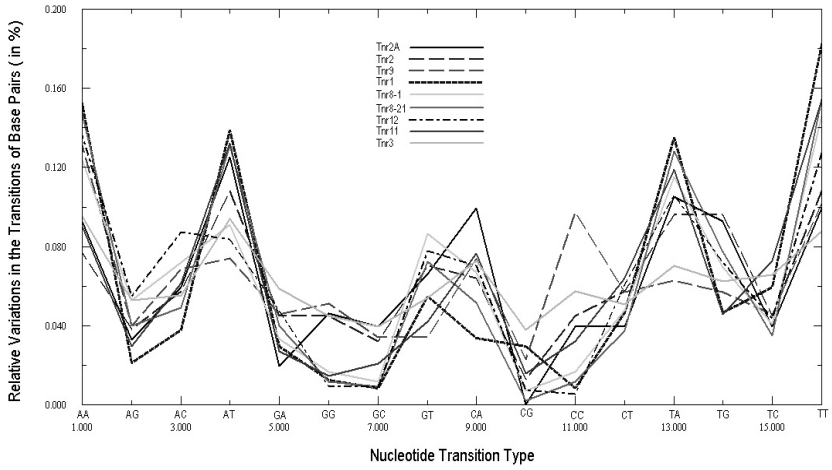

Figure 1: Trend analysis of the sequence data set tested for nucleotide transition types, showing similarity in the richness of AT, CA and TA transitions and poorness of AG, GA, GC, CG and TC transitions.

The results of the analysis of the sequences in the numerical domain are presented in Table 2. The real eigen values of matrix **P** hovers around a very narrow range. Irrespective of sequence lengths, even the real dominant eigen values occur in a narrow range of 0.2574, for TNR3 to 0.3293, for TNR1. The narrow band of the real eigen values may be attributed to the consensus nature of the repeat sequences.

The results of numerical characterizations of the transition proportion matrices of the tested sequences are provided in Table 3. The superimposed plot of the linear fits of the consensus Tnr category repetitive sequences is given in Figure 2. All the straight line fits, except that of TNR2A intersect each other at a point, while that of TNR2A makes four different intersections with the linear fits of TNR1, TNR9, TNR11and TNR12. The slopes and y-intercepts of the linear fits also lie in a narrow range (Table 3). The straight line representations of the sequences TNR3, TNR9 and TNR12 have negative slopes while that of the rest of the sequences have positive slopes.

Table 2: Results of algebraic characterizations of the transition proportion matrices of the tested sequences.

Repeat Sequence Id.	Total number of bases	Real Eigen Values (REV)	Dominant Eigen Value (DEV)
TNR2A	152	0.2723**,** -0.04925	0.2723
TNR2	157	0.2681, -0.02384, 0.02681, 0.00457	0.2681
TNR9	176	0.2626, 0.0357, 0.04389, 0.0406	0.2626
TNR1	237	0.3293, 0.03751	0.3293
TNR8-1	418	0.2913, 0.04956	0.2913
TNR8-21	428	0.3159,0.0387, -0.0232, -0.0058	0.3159
TNR12	528	0.2876, 0.0472, -0.01543, 0.04048	0.2876
TNR11	811	0.2992, -0.0140, 0.0031, 0.0031	0.2992
TNR3	1536	0.2574, 0.010	0.2574

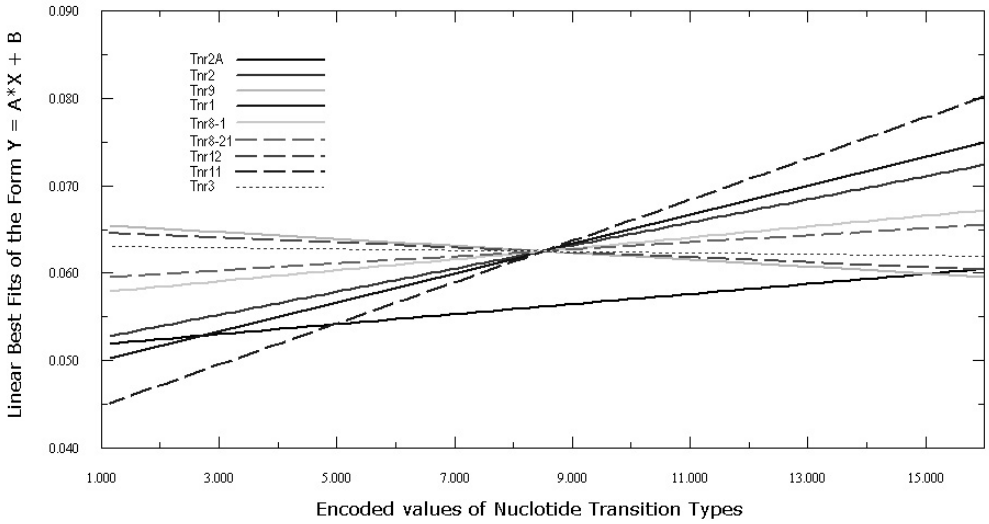

Figure 2: An illustration of the linear fits of the consensus TNR category repetitive sequences found in the genome of Oryza sativa.

Table 3: Results of numerical characterization of the transition proportion matrices of the tested sequences.

S.No.	Repeat Sequence Id.	Line of Best Fit	Slopes
1	TNR2A	$5.7460x10^{-3}X + 5.7619x10^{-2}$	$5.7460x10^{-3}$
2	TNR2	$1.3198x10^{-3}X + 5.1282x10^{-2}$	$1.3198x10^{-3}$
3	TNR9	$-3.9496x10^{-4}X + 6.5857x10^{-2}$	$-3.9496x10^{-4}$
4	TNR1	$1.6699x10^{-3}X + 4.8305x10^{-2}$	$1.6699x10^{-3}$
5	TNR8-1	$6.2421x10^{-4}X + 5.7194x10^{-2}$	$6.2421x10^{-4}$
6	TNR8-21	$4.0295x10^{-4}X + 5.9075x10^{-2}$	$4.0295x10^{-4}$
7	TNR12	$-2.8184x10^{-4}X + 6.4896x10^{-2}$	$-2.8184x10^{-4}$
8	TNR11	$2.3566x10^{-3}X + 4.2469x10^{-2}$	$2.3566x10^{-3}$
9	TNR3	$-7.3769x10^{-5}X + 6.3127x10^{-2}$	$-7.3769x10^{-5}$

The correlation analysis between various pairs (36 pairs) yielded correlation coefficients in the range 0.539 (minimum value for the pair of TNR2 and TNR9) to 0.971 (maximum value for the pair of TNR8-1 and TNR12). A **r** value of more than 0.8 was obtained in 21 cases, between 0.6 and 0.8 in 12 cases and less than 0.6 in only 3 cases. The computation of Sakoda coefficients for these pair of sequences, also confirm the high degree of association between these sequences, thus confirming the consensus nature of them. A further study of these sequences, with the aid of obtained mathematical indicators, in a biological perspective may shed more light on the genomic significance.

Conclusion

A novel technique containing the combined power of matrix, statistical and numerical procedures is proposed and illustrated for repetitive elements present in the genome sequence of Oryza sativa. The mathematical indicators used in this technique help to infer the sequence patterns in a direct manner. As the sequence data being normalized before analysis, this method can be useful for rapid computation and comparative analysis of lengthy sets of sequences and the result indicator would differ by a scale factor only. The above demonstrated technique thus works as a handy tool in recognizing the sequence variations. The method can easily be applied to study both intronic and exonic regions and serve as a codon region analyzer. The transformation of the biological sequences from character strings to the numerical domain not only helps for the multitude of analysis strategies but also enhances the human understanding. The pictorial representation and the related statistical parameters will help biologists in arriving at some concrete conclusions easily and to decide further course of actions.

References

[1] Elias Daura-Oller, Maria Cabre, Miguel A. Montero, Jose L. Paternain, Antoni Romeu (2009). Specific gene hypomethylation and cancer: New insights into coding region feature trends. *Bioinformation* 3(8): 340-343.

[2] Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O., Walichiewicz, J. (2005) *Repbase Update*, a database of eukaryotic repetitive elements. Cytogentic and Genome Research, 110, 462-467. Repbase version 15.01, (2010). FASTA and EMBL formats. GIRI: Genetic Information Research Institute. www.Girinst.org

[3] Davis, J. C. Statistics and Data analysis in Geology, John & Sons, INC, New York, London , Sydney, (1973).

[4] Zu-Guo yu and Guo-yi Chen (2000). Rescaled range and transition matrix analysis of DNA Sequences. *Comm. Theor. Phys.* 33(4) 673-678.

[5] Han C.G., Frank M.J., Ohtsubo H. and Ohtsubo E. (2000). New transposable elements identified as insertions in rice transposon Tnr1. *Genes Genet Syst.* Apr; 75(2):69-77.

[6] Yu, J., Hu, S., Wang, J., Wong, G.K., Li, S., Liu, B., Deng, Y., Dai, L., Zhou, Y., Zhang, X., et al. (2002). A draft sequence of the rice genome (*Oryza sativa L. ssp. indica*). *Science* 296: 79–92.