Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 01, January-2020

State Machine based Framework for Genomic Analysis

[1.] Lakshmi Bharathi S, [2.] Dr. Vidya Niranjan, [3.] Sudhamshu Mohan S
Mysore Road, R V Vidyanikethan, Bengaluru 560059,
Karnataka, India

INTRODUCTION

The extensibility of Finite State Machines (FSM) to different disciplines viz., networking, compiler design, marketing, etc., is highly appreciated for its optimal and precise solutions to respective problems[12]. The attempt of attributing states of a FSM to biological data, specifically to genes is rarely endeavored because it is very difficult to visualize and represent states and events corresponding to genomic data. This constraint paved the way in looking for a transform using which representable genomic states and events could be realized. The subsequent sections of this paper introduce to such a linear transform thereby the mapping of genes leads to mathematically representable states[17].

METHODOLOGY

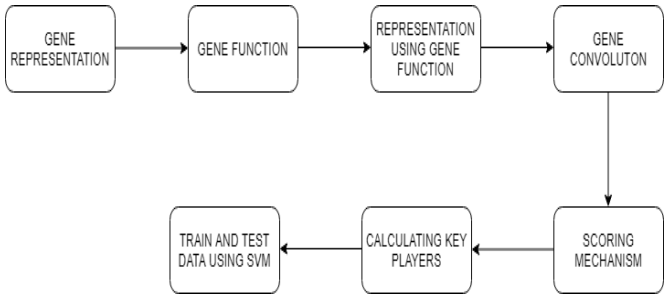
A. Gene Representation

Let $\varphi(G) = \{G_1, G_2, \ldots G_N\}$ be the genes present in a genome, where G_1, G_2,…G_N are genes. As there we be a really large set of genes present in genome. This representation can lead to a very large set.

Mathematically $\varphi(G)$ contains finite set of genes, however large it might be. We can use mathematically convenient representations which are computationally appreciable

$$\varphi(G) = \{G_1, G_2, \ldots G_N\}$$

$$\varphi = \sum_{k=1}^{N} Gk * P\delta k$$

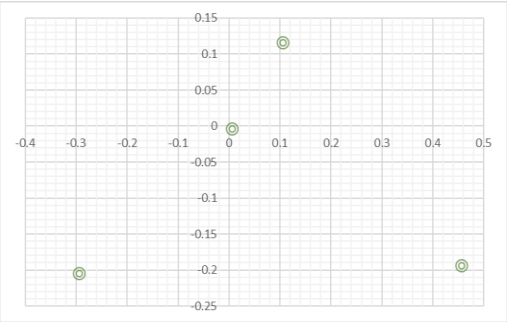


General flow of the framework

Where, $P\delta k$ is impulsive only at k^{th} index where gene is present. In fact, $P\delta k$ will have a value unity where gene-k exists and is equal to 0 elsewhere. This throws focus on mathematical representation of nucleotides A T G C.
Bergen and Antoniou proposed a method based on complex representation, parametric window function and STDFT to maximize SNR (Signal to Noise Ratio) to identify coding regions of the genes

A= 0.10+0.12j;
T= -0.30-0.20j;
G=0.45-0.19j;
C=0

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 01, January-2020



Graphical representation of nucleotides

B. Gene Function

$\eta(G) = \{ATAGCCT \ldots TGAC\}$ which is a $function\ of\ \{ATAGCCT \ldots TGAC\}$ Such that $\eta(G)$ leads to a numerical value V which is complex in nature.

$\eta(G)$ should be a convergent function.

$$\eta(G) = \mu_N e^{-\sqrt{(\mu_N+\sigma_N)(\mu_N-\sigma_N)}\sqrt{(\mu_G+\sigma_G)(\mu_G-\sigma_G)}}$$

Where $f(n) = \{A, T, G, C \ldots\}$ Where, 'n' is the position of nucleotide within a gene

$\mu_G = \frac{\sum_n f(n)}{l}$ Where, μ_G is the arithmetic mean of gene, *l* is the number of nucleotides within a gene

$\mu_N = \frac{A+T+G+C}{4}$ μ_N is the mean of the nucleotide

$\sigma_N = \sqrt{\sum_m \frac{(\mu_N - g(m))^2}{4}}$ σ_N the variance of the nucleotide

$$\sigma_G = \sqrt{\sum_n \frac{(\mu_G - f(n))^2}{l}}$$

So, we will have each gene getting mapped to a number, uniqueness of this value increases with increase in number of parameters, that is if we go for μ and σ, the value we get will be more unique.

C. Representation using Gene Function

$$\varphi(G) = \{G1, G2, \ldots GN\}$$
$$= \{\eta(G1), \eta(G2), \ldots \eta(GN)\}$$
$$\varphi(G) = \{val1, val2, \ldots valN\}$$

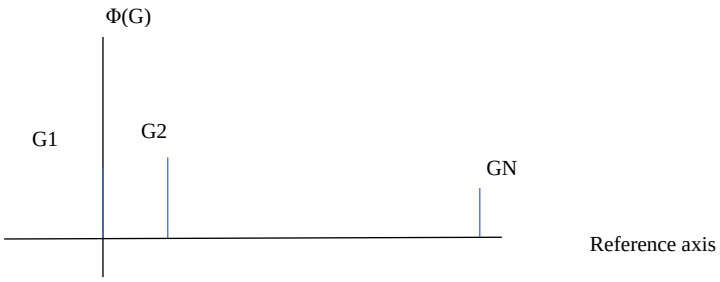
$\varphi(G)$ can be represented as a right-handed sequence starting from origin till valN.

Similarly, we can also represent the query sequence.
Now we have mathematically representable sequences and numbers which are unique.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 01, January-2020

D. Gene Convolution

Genome convolution is the convolution of these sequences



Here, G1, G2, … GN are complex in nature.

Similarly, Reference gene R_G corresponds to some value.

With all these assumptions, we need to find the Ψ1 and Ψ2
Where,
$\psi 1(G) = \phi(G) * Q(G)$ and
$\psi 2(G) = \phi(G) * RG(G)$
Ψ1(G) and Ψ2(G) are the resultants obtained by convolving genomic sequence with the query sequence and genomic sequence with the reference sequence.[13]
If Ψ1(G) and Ψ2(G) are of bigger length, we need to use transformation.

τ(Ψ1(G), Ψ2(G))

We need to first reduce the feature space of Ψ1(G) and Ψ2(G) using K means algorithm, if the feature space is very large to the best samples.

E. Scoring Mechanism

Now, scoring mechanism needs to be employed $\begin{bmatrix} \gamma 00 & \cdots & \gamma 0q \\ \vdots & \ddots & \vdots \\ \gamma p0 & \cdots & \gamma pq \end{bmatrix}$ This will be a rectangular matrix of order p*q.

γpq = (Ψ1(G) - Ψ2(G))/Δ
where, Δ is the determinant.
We need to make order (r*r) where R is the LCM(p,q) = We need to make all the remaining entries 0 by appending it everywhere

$$\begin{bmatrix} \gamma 11 & \cdots & \gamma pr \\ \vdots & \ddots & \vdots \\ \gamma r1 & \cdots & \gamma rr \end{bmatrix}$$

Now, the determinant needs to be calculated

F. States (Key Players)

Now try to make the matrix upper/lower triangular or try to diagonalize it (Echelon forms).
The remaining entries now are the key players of our analysis. If the matrix is diagonalizable then analysis becomes easy. Judgement (J) from the key players will be based on the learning methods.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 01, January-2020

Inputs for learning are

[Principal diagonal elements or Triangular matrix] + [Determinant obtained from matrix] + [Rank of the matrix]

Using some machine learning techniques like multi-layered perception or SVM we can train some of the inputs. Involving reference genes for a given genome with some of the query sequences (around 30-40 samples)
Then by using this binary classifier we can make a decision for a particular analysis.

G. SVM based Binary Classifier

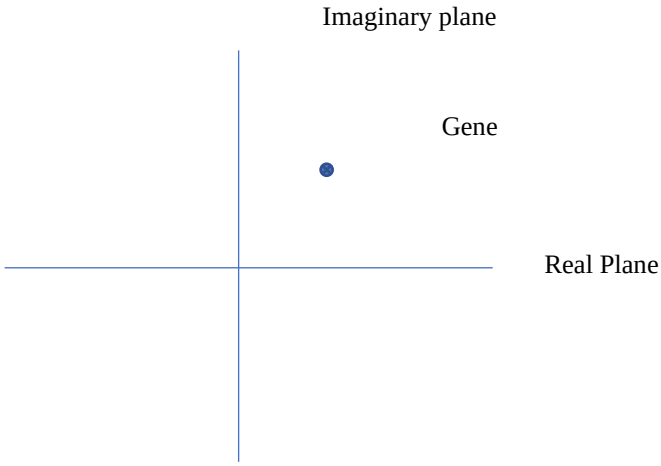
Training inputs	Output Range
Ref Gene +Genome +Query sequence	Decide on output range of values

II. CONSTRAINTS FOR THE APPLICATION OF THIS METHOD

1. Gene Representation

To mathematically represent a gene, following parameters need to be decided.

a. All the nucleotides must be unique i.e., the number corresponding to each nucleotide must be different from others.
b. These numbers must be linearly independent and should not belong to same linear space.
c. Orthogonality is the most preferred feature to introduce the uniqueness in the analysis.
d. Each nucleotide will be a point on complex plane. (S-plane in Laplace plane)



2. Gene Function
   a. It should be convergent
   b. It should be definitive, continuous and differentiable
   c. Periodic and non-periodic properties need to be studied further
3. Scoring Mechanism
   a. As this involves finding key players (numbers) that is finding triangular elements or diagonal elements, the values are highly uncorrelated.
   b. Dividing the values by determinant will normalize the values.

III. APPLICATIONS

This idea could be transformed as a framework for Genomic Analysis, primarily intended to characterize nucleotide, gene, genome and their expressions in a mathematically coherent way.[14] The inherent modularity of this framework ensures to enhance and improve performance of each stage independently. As the crux of the framework is principally derived harnessing the concept of linear systems viz. Convolution, Linear Transformation- Cartesian Association, K- means Clustering, Discrete Differentiation, Characteristic Equations followed by a training method based on Support Vector Machines, the approach employed is robust, simple, conclusive and reliable.[16]

The tangibility of this approach is derived from the philosophy of Linearization of Sample Space.[13] This novel approach shapes & fits the non-characterized biological data into a framework where one can apply any of the Linear Discriminative Techniques which are deterministic in nature, thus leading to conclusions which are reliable and specific in nature. [18]

IV. FUTURE SCOPE

As this framework is modular and generic in nature, this could be enhanced and extended to techniques based on Fuzzy Logic, for better decisiveness and quantifiability of conclusions. For huge training data sets a layered approach involving Artificial

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 01, January-2020

intelligence-based approach could be studied. For better characterization of Gene Functions, we can even go for multi parameter based statistical techniques for Modelling, which are based on Curvature based analysis.

V. REFERENCES

[1] FrancisDutil, JosephPaulCohen, MartinWeiss, GeorgyDerevyanko, YoshuaBengio *Towards Gene Expression Convolutions using Gene Interaction Graphs* arXiv:1806.06975v1 [q-bio.GN] 18 Jun 2018
[2] Michaël Defferrard, Xavier Bresson, Pierre Vandergheynst *Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering* EPFL, Lausanne, Switzerland
[3] Yifei Chen1,4,†, Yi Li1,†, Rajiv Narayan2, Aravind Subramanian2 and Xiaohui Xie, *Gene expression inference with deep learning* Bioinformatics Advance Access published February 11, 2016
[4] Tanya Barrett*, Dennis B. Troup, Stephen E. Wilhite, Pierre Ledoux, et.al., *NCBI GEO: archive for high-throughput functional genomic data* Nucleic Acids Research, 2009, Vol. 37, Database issue D885–D890 doi:10.1093/nar/gkn764
[5] David Warde-Farley, Sylva L. Donaldson, Ovi Comes, Khalid Zuberi, et.al., *The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function* Nucleic Acids Research, 2010, Vol. 38, Web Server issue doi:10.1093/nar/gkq537
[6] Minoru Kanehisa, Miho Furumichi, Mao Tanabe, Yoko Sato et.al., *KEGG: new perspectives on genomes, pathways, diseases and drugs* Nucleic Acids Research, 2017, Vol. 45, Database issue D353–D361 doi: 10.1093/nar/gkw1092
[7] James M. Heather, Benjamin Chain, The Sequence of Sequencers: *The History of Sequencing DNA*, Genomics (2015), doi: 10.1016/j.ygeno.2015.11.003
[8] Stanley H. Chan, *CONSTRUCTING A SPARSE CONVOLUTION MATRIX FOR SHIFT VARYING IMAGE RESTORATION PROBLEMS Proceedings of 2010* IEEE 17th International Conference on Image Processing
[9] Yanwei Pang, Senior Member, Manli Sun, Xiaoheng Jiang, and Xuelong Li, *Convolution in Convolution for Network in Network* IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS
[10] S. H. Chan, "Constructing a sparse convolution matrix for shift varying image restoration problems," *2010 IEEE International Conference on Image Processing*, Hong Kong, 2010, pp. 3601-3604.
[11] Y. Pang, M. Sun, X. Jiang and X. Li, "Convolution in Convolution for Network in Network," in *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 5, pp. 1587-1597, May 2018.
[12] R. Xi, M. Hou, M. Fu, H. Qu and D. Liu, "Deep Dilated Convolution on Multimodality Time Series for Human Activity Recognition," *2018 International Joint Conference on Neural Networks (IJCNN)*, Rio de Janeiro, 2018, pp. 1-8
[13] L. Gao, P. Chen and S. Yu, "Demonstration of Convolution Kernel Operation on Resistive Cross-Point Array," in *IEEE Electron Device Letters*, vol. 37, no. 7, pp. 870-873, July 2016.
[14] J. Shangguan, Y. Li, Y. Wang and H. Li, "Fast algorithm of modified cubic convolution interpolation," *2011 4th International Congress on Image and Signal Processing*, Shanghai, 2011, pp. 1072-1075.
[15] X. Gao and H. Xiong, "A hybrid wavelet convolution network with sparse-coding for image super-resolution," *2016 IEEE International Conference on Image Processing (ICIP)*, Phoenix, AZ, 2016, pp. 1439-1443.
[16] S. Shrivastava and P. Rawat, "High speed and delay efficient convolution by using Kogge Stone device," *2017 International Conference on Computer Communication and Informatics (ICCCI)*, Coimbatore, 2017, pp. 1-5.
[17] B. Bipin and J. J. Nair, "Image convolution optimization using sparse matrix vector multiplication technique," *2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, Jaipur, 2016, pp. 1453-1457.
[18] S. Jain and S. Saini, "High speed convolution and deconvolution algorithm (Based on Ancient Indian Vedic Mathematics)," *2014 11th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, Nakhon Ratchasima, 2014, pp. 1-5.
[19] C. Radhakrishnan and W. K. Jenkins, "Modified Discrete Fourier Transforms for fast convolution and adaptive filtering," *Proceedings of 2010 IEEE International Symposium on Circuits and Systems*, Paris, 2010, pp. 1611-1614.
[20] R. Krutsch and S. Naidu, "Monte Carlo method based precision analysis of deep convolution nets," *2016 Conference on Design and Architectures for Signal and Image Processing (DASIP)*, Rennes,
[21] 2016, pp. 162-167.
[22] S. Kambhampati, "Power efficient modulo convolution," *2016 International Conference on Inventive Computation Technologies (ICICT)*, Coimbatore, 2016, pp. 1-6.
[23] P. Katkar, T. N. Sridhar, G. M. Sharath, S. Sivanantham and K. Sivasankaran, "VLSI implementation of fast convolution," *2015 Online International Conference on Green Engineering and Technologies (IC-GET)*, Coimbatore, 2015, pp. 1-5.
[24] A. Khumaidi, E. M. Yuniarno and M. H. Purnomo, "Welding defect classification based on convolution neural network (CNN) and Gaussian kernel," *2017 International Seminar on Intelligent Technology and Its Applications (ISITIA)*, Surabaya, 2017, pp. 261-265.