Spectral Domain Characterization of Genome Sequences

Pedakolmi Venkateswarlu
Department of Computer Science
University of Mysore, Manasa Gangotri
Mysore, Karnataka, India
venkat123.pedakolmi@gmail.com

E. G. Rajan
Director, Avatar MedVision US. LLC, USA
Director, Pentagram Labs Inc, California, USA
Scientific Advisor, Helios & Matheson Analytics, USA
rajaneg@yahoo.co.in

Abstract—Enormous amount of genomic and proteomic data is available to us in the public domain and there is need for faster and more sensitive algorithms for genomic study in view of this rapidly increasing amounts of genomic sequences. Basically genomic data falls under the category of ‘Big Data’ and thus processing these large data has become a difficult task. In this regard, people working in the area of data mining and pattern recognition have realized the importance of ‘Machine Learning’ techniques in analyzing such big data. In this context, this paper proposes a novel technique of ‘Spectral characterization of genomic sequences’ for DNA analysis purpose.

Keywords—Genome Analytics, Big Data, Spectral Characterization, Pattern Recognition

I. INTRODUCTION

Genome data is now treated as ‘Big Data’. Parallel Programming would then be the most important tool to calculate and compute like a supercomputer and yet, be economical. The power spectrum of a sequence is a transformation of that sequence of variables in the frequency space. It has a significant advantage that periodic patterns in the sequence - hidden or latent - become evident after transformation. That is, hidden periodic information are shown as peaks in the spectrum. Periodicity of DNA sequences have already been examined using various methods like autocorrelation and auto convolution. Various discrete transforms have been tried on genomic sequences like Discrete Fourier transform, Wavelet Transform and their power spectra examined. One such transform is Rajan Transform which unlike other transform exhibits homomorphism property in addition to being an isomorphic map. Homomorphism means classification or pattern recognition. A case study was made on a genome sequence of Brucella Suis 1330. This sequence was obtained from NCBI. Discrete Fourier Transform based power spectrum of the adjoint sequences were obtained and the spectral data analyzed for extracting hidden information pertaining to various codons. Rajan Transform based power spectrum of the adjoint sequences were also obtained and examined for codon information. Results were compared and analytical reports presented in this paper.

II. POWER SPECTRA OF GENOME STRANDS

A genome strand has to be converted into a number sequence so that signal processing methods could be applied to it for analysis of genomic data and extraction of features of chromosome that would be difficult to obtain using standard statistical methods. The question that arises at this juncture is whether spectral analyses useful for DNA sequence analysis. The answer is affirmative. Spectral analysis plays an important role in detecting latent periodicities and in distinguishing them from other tandem repeats. Usually latent periodicities indicate biological meanings. One can filter a genome sequence based on single nucleotide, base type, dinucleotides, to name a few, and can enhance the interpretation of signal. Spectral analysis could also be used to characterize large-scale fluctuations of base compositions and thus one can analyze genome-wide patterns. Let X[n]: 0≤n≤N-1 denote DNA of length N. X[i] represents a sequence residue. A DNA sequence may be thought of as the one comprising m different residues, which come from an alphabet set A_m = {r; r = 1, 2,..., m}. Typical alphabet examples are (i) for m = 4, the nucleotide alphabet is A_4 = {A, C, G, T} and (ii) for m = 20, A_20 is the amino acid set. This generalized designation allows one to have additional sequence residues to be studied, like codons or other functional groups. Let us consider a nucleotide sequence X(n) = GCCAAAAATCAGCTAAT CGC. Now adjoint of A is defined as A(n) = { x = 1 for X(i) = A ; 0 for X(i) ≠ A}. Table 1 shows this coding pattern.

| Adjoint of GCCAAAAATCAGCTAAT CGC | 0 0 0 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 |

Hence A(n) = 00011110000000000. Similarly one can define G(n), C(n), T(n). Triplets of nucleotides are called CODONS. Given a genome character string of ATGC, that is a strand, one can obtain codon adjoint by searching for a specific codon and replacing a match by 1 and others by 0s. Four of the 64 codon adjoints obtained from X(n) = GCCAAAAATCAGCTAAT CGC are shown in table 2.

| Codon Adjoint | 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 |

Table 1: Adjoint of GCCAAAAATCAGCTAAT CGC

Table 2: Some adjoints of X(n)
This amounts to saying that one can construct 64 triplet codon adjoints for any given genome strand. From these 64 codon adjoints, one can construct a 'character sequence' using the procedure given below.

**Procedure for obtaining character sequence**

Find all the 64 codon adjoints.

Find the sum of each adjoint.

Obtain all 64 sum values.

This sequence is called 'character sequence' of the strand. The character sequence for X(n) = G C C A A A A A T C A G C T A A T C G C is shown in table 3.

**Table 3: Sum of codon adjoints given in table 2**

| AAA-adjoint | AAT-adjoint | GCC-adjoint | CGC-adjoint |
|-------------|-------------|-------------|-------------|
| Sum Value   | Sum Value   | Sum Value   | Sum Value   |

After obtaining the given strand's character sequence of length 64, one can apply all signal processing tools like 'autocorrelation', 'auto convolution', 'Fourier Transform', 'Wavelet Transform', 'Rajan Transform' to name a few to analyze the sequence. This paper discusses the spectral analysis of genome character sequences using Discrete Fourier Transform and Rajan Transform. The basic definitions of these two transforms are given below.

**Discrete Fourier Transform**

Consider a sequence x(n) of length N. Then, its DFT (Discrete Fourier Transform pair) is given by the expressions

\[ X(k) = \sum_{n=0}^{N-1} x(n) e^{-j2\pi nk/N}, \quad k = 0, ..., N-1 \]  
\[ x(n) = \frac{1}{N} \sum_{k=0}^{N-1} X(k) e^{j2\pi nk/N}, \quad n = 0, ..., N-1 \]  

(1)  
(2)

All details relevant to this transform could be obtained from standard literature.

**Rajan Transform**

Assume a sequence x(n) of length N, which is a power of 2. First it is divided into the first half and the second half each consisting of N/2 points so that the following hold good.

\[ g(i) = x(i) + x(i+N/2) \quad 0 \leq i < N/2 \]  
\[ h(i) = x(i) - x(i+N/2) \quad 0 \leq i < N/2 \]  

Now each (N/2) point segment is divided into two halves each consisting of (N/4) points so that the following hold good

\[ g(i) = x(i) + x(i+N/4) \quad 0 \leq i < N/4 \]  
\[ h(i) = x(i) - x(i+N/4) \quad 0 \leq i < N/4 \]  
\[ g(i) = x(i) + x(i+N/4) \quad N/4 \leq i < N/2 \]  
\[ h(i) = x(i) - x(i+N/4) \quad N/4 \leq i < N/2 \]  
\[ g(i) = x(i) + x(i+N/4) \quad N/2 \leq i < N \]  
\[ h(i) = x(i) - x(i+N/4) \quad N/2 \leq i < N \]  

This process is continued till the final division. Consider a sequence x(n) = 3, 8, 5, 6, 0, 2, 9, 6. The total number of stages of computation would then be log_2N = 3. The sum and difference operators are respectively denoted as '+' and '-'.

The signal flow graph for computing Rajan Transform (RT) of x(n) of length 8 would be of the form shown in figure 1. Once the character sequence is calculated for a genome sequence, one can evaluate its spectral sequences due to 'Fourier Transform' and 'Rajan Transform', for instance, and the power spectral sequences of the character sequence by applying 'Fourier Transform' and 'Rajan Transform' to its 'auto correlated', or 'auto convolved' sequences.

---

**III. POWER SPECTRAL ANALYSIS OF BRUCELLA SUIS 1330**

X(n), a part of the Brucella Suis 1330 genome of length 5806 is considered here for a case study. This genome data was retrieved from NCBI web site.
Adjoins of A, T, G and C of the genome and 64 codon adjoints were obtained from the genome sequence. Sums of all 64 codon adjoints were calculated and the character sequence of the genome formed. For brevity, we call this character sequence as ‘codon sequence’. Codon sequence of the genome sequence of length 5806 is given below.

**Codon sequence:**

19612, 16291, 19662, 14560, 12445, 16340, 23517, 23652, 13051, 6463, 16850, 22298, 11730, 6465, 16797, 16907, 6066, 12602, 4722, 5620, 6150, 19437, 20315, 24409, 22283, 28775, 19017, 13537, 26124, 17876, 24188, 23675, 12504, 15728, 5495, 14602, 15683, 15623, 17799, 16447, 15403, 37796, 28608, 22727, 40187, 37797, 20260, 23385, 21776, 15985, 4921, 19932, 22299, 12928, 26226, 16754, 32909, 40451, 22004, 17346, 33230, 15951

**Spectral Analysis Report of adjoints of X(n).**

Fig. 2 shows the graphical representation of the adjoints

**Fig. 2:** Graphs of adjoints of X(n)

Auto correlation of a sequence y(n) of length N is calculated using the formula

\[ R_{yy}(l) = \sum_{n=0}^{N-1} y(n) y(n+l) \]  \hspace{1cm} (3)

The auto correlation of y(n) would be of length 2N-1. The auto correlation of adjoints of X(n) are shown in Fig. 3.

**Fig. 3:** Auto correlation of adjoints of X(n)

Auto convolution of a sequence y(n) of length N is calculated using the formula \( R_{yy}(t) = \sum_{n=0}^{N-1} y(n) y(n-t) \). The auto convolution of y(n) would be of length 2N-1. The auto convolution of adjoints of X(n) are shown in Fig. 4.

**Fig. 4:** Auto convolution of adjoints of X(n)

Cross correlation of sequences y(n) and z(n) each of length N is calculated using the formula \( R_{yz}(l) = \sum_{n=0}^{N-1} y(n) z(n+l) \). The cross correlation of y(n) would be of length 2N-1. The cross correlation of adjoints of X(n) are shown in Fig. 3.

**Fig. 3(a):** Cross correlation of adjoints of X(n) – Set #1

**Fig. 3(b):** Cross correlation of adjoints of X(n) – Set #2

**Fig. 3(c):** Cross correlation of adjoints of X(n) – Set #3

From Fig. 3(a), (b) and (c) one can observe that the operation of cross correlation is a not commutative. Similarly, one can construct cross convolution of difference sequences. Fourier Spectra of adjoints of X(n) could be obtained by applying the DFT to them. Fig. 4 shows all such spectra.

**Fig. 4:** Fourier Spectra of A(n), G(n), T(n), C(n)

Fig. 5 shows the Fourier power spectra of the auto correlations of the adjoints of X(n).

**Fig. 5:** Fourier Power Spectra of RA(\(\omega\)), RG(\(\omega\)), RT(\(\omega\)), RC(\(\omega\))

Similar to what have been discussed so far, one can apply Rajan Transform to adjoints of X(n) so that the resulting spectra could be called as Rajan Spectra. Fig. 6 shows all such Rajan Spectra of the adjoints of X(n).

**Fig. 6:** Rajan Spectra of A(n), G(n), T(n), C(n)

**Fig. 7:** Rajan Power Spectra of RA(\(\tau\)), RG(\(\tau\)), RT(\(\tau\)), RC(\(\tau\))

2016 2nd International Conference on Next Generation Computing Technologies (NGCT-2016)
Dehradun, India 14-16 October 2016
Fig. 7 shows all such Rajan Spectra of the auto correlation of adjoints of X(n). Similarly one can evaluate auto convolutions, cross correlations and Rajan power spectra of the adjoints of X(n). Codon sequence of the Brucella Suis 1330 genome sequence of length 5806 is given below.

19612, 16291, 19662, 14560, 12445, 16340, 23517, 23652, 13051, 6463, 16850, 22298, 11730, 6465, 16797, 16907, 6066, 12602, 4722, 5620, 6150, 19437, 20315, 24409, 19017, 11746, 22283, 28775, 19070, 13537, 26124, 17876, 24188, 23675, 12504, 15728, 5495, 14602, 15683, 15623, 17799, 16447, 15403, 37796, 28608, 22727, 40187, 37797, 20260, 23385, 21776, 15985, 4921, 19932, 22299, 12928, 26226, 16447, 15403, 37796, 28608, 22727, 40187, 37797, 20260, 23385, 21776, 15985, 4921, 19932, 22299, 12928, 26226, 16754, 32909, 40451, 22004, 15951

As done in the case of adjoints of X(n), one can implement all the signal processing operations on the codon sequence of X(n). Fig. 8 shows the Fourier spectrum and Rajan spectrum of the codon sequence.

Fig. 8: Fourier and Rajan spectra of the codon sequence of X(n) with auto correlation and auto convolution

Fig. 9 shows Fourier and Rajan power spectra of the auto correlation and auto convolution of codon sequence

Fig. 9: Fourier and Rajan power spectra of the auto correlation and auto convolution of codon sequence

IV. CONCLUSION
From Fig. 8 and Fig. 9 it is evident that Rajan spectra of adjoints and Rajan power spectra yield better results and better visualization standards.

ACKNOWLEDGMENT
The authors place their gratitude on record for the immense technical, moral and administrative support extended by Mrs. G. Chitra, Managing Director of Pentagram Research Centre Pvt Ltd., India, Dr. Michael Patrick Coyle, CEO of Avatar MedVision U.S., LLC., NC, USA and Mr. Pat Krishnan, CEO of Pentagram Labs Inc., California, USA while carrying out research in this advanced state-of-the-art technology.

REFERENCES
[1] Rajan E. G., “Symbolic Computing – Signal and Image Processing”, Anshan, CBS, UK, 2004.
[2] Rajan, E.G., Srinivas, S.S.S., Srinivas, P.L., Mallikarjuna Rao, G., Srinivasa Rao, N., “On the notion of Generalised Rajan Transform”, World multiconference on Systemics, Cybernetics and Informatics, Caracas, Venezuela, July 7-11, 1997.
[3] Rajan, E.G., Prakash Rao, K., Vikram, K., Harish Babu, S., Rammurthy, D., Rajasekhar, “A Generalised Rajan Transform Based Expert System for Character Undersstanding”, World multiconference on Systemics, Cybernetics and Informatics, Caracas, Venezuela, July 7-11, 1997.
[4] J. P margues de sa, “Pattern Recognition Concepts, Methods and Applications, Springer, May 2001, Prtugal.
[5] Rajan, E.G., Ramakrishna Rao, p., Mallikarjuna Rao, G., and Venugopal, S., “Use of Generalised Rajan Transform in Recognising Patterns due to defects in the weld”, JOM—8 International Conference on the Joining of Materials, American Welding Society, JOM Institute, Helsingor, Denmark, May 1997.
[6] Richard O. Duda, Peter E. Hart, David G. Stork, “Pattern Classification”, Second edition.
[7] Rajan, E.G., Venugopal, S., “The role of RT in the machine vision based factory automation”, National conference on Intelligent manufacturing systems, February 6 – 8, 1997, Coimbatore Institute of Technology.
[8] Srinivas Rao, N., “Data Encryption and Decryption Using Generalised Rajan Transform”, M. Tech. Thesis, department of Electronics and Communication Engineering, Regional Engineering College, Warangal, 1997.
[9] Robert J. Schalkoff, Pattern Recognition: Statistical, structural and Neural approaches. ISBN : 0-0471-52974-5, June 1991.
[10] Dr. E. G. Rajan, Symbolic Processing of Signals and Images. First edition, 2003.
[11] R. C Gonzales and R. E Woods, Digital Image Processing Addison Wesley, 2000.
[12] Richard. O Duda and Peter E. Hart : Pattern Recognition (Pattern classification and Scene analysis) Second Edition 2000.
[13] Julius J. Tou rafael C. Gonzales, Pattern Recognition Principles, Addison Wesley 1974.
[14] Don Pearson, Image Processing, Tata McGraw Hill, U.K. 1991.
[15] Chou-Ting Hsu and Ja-ling Wu, Hidden Digital Watermarks in Images, Senior Member, IEEE. IEEE Transactions on Image Processing.
[16] Markov A A, Theory of Algorithms, Israel Program for scientific Translation, DC, 1951.
[17] Rajan E.G., Cellular Logical Array Processing Techniques for High Throughput Image Processing, Special issue on computer vision, June 1992, Indian Academy of Science, Bangalore.
[18] Kishore K, Subramanyam J. V., and Rajan E. G., “Thinning Lattice Gas Automation Model for Solidification Processes, National Conference by ASME, USA, Hilton Hotel, California.
[19] Rajan E. G., “Fast algorithm for detecting volumetric and superficial features in 3-D images”, International Conference on Biomedical Engineering, Osmania University, Hyderabad, 1994.
[20] Sergios Teodoridis, Konstantinos Koutroumbas, “Pattern Recognition” 2nd edition, Elsevier Academic Press, 2001.
[21] Andrew Webb, “Statistical Pattern Recognition”, 2nd Edition, Wiley 2002.