Graphical Representation and Comparison of Whole Genome Sequences of different strains of Mycobacterium Tuberculosis

SHIWANI SAINI
Department of Electrical Engineering
National Institute of Technology
Kurukshetra, Haryana
INDIA
shiwani_saini76@yahoo.com

LILLIE DEWAN
Department of Electrical Engineering
National Institute of Technology
Kurukshetra, Haryana
INDIA
l_dewanin@yahoo.com

Abstract: - Tuberculosis (TB) is global health problem and is the leading cause of mortality. Global TB control is a difficult task due to the prevalence of multidrug resistant (MDR) and extensively drug resistant (XDR) strains of TB. New tools for faster and accurate diagnosis of drug-resistant TB are urgently needed as early detection of drug resistance allows starting of an appropriate treatment. Need for faster assessment can be addressed by genomic signal processing based methods. In this paper, a graphical method has been used to compare the DNA(deoxyribonucleic acid) sequences of different strains of tuberculosis that predict the nature of the mycobacterium tuberculosis (MTB) strains thus saving time for initiating adequate therapy.

Key-Words: Tuberculosis, Extensive Drug Resistant (XDR), Multidrug Resistant (MDR), Drug susceptible (DS), Genomic Signal Processing, Z-curves, graphical representations.

1 Introduction

Human tuberculosis is caused by an intracellular pathogen, mycobacterium tuberculosis and it replicates rapidly in the lungs where oxygen concentration is high as MTB is aerobic. Resistance of MTB to anti-TB drugs is caused by chromosomal mutation. Surveys conducted by the World Health Organization and the International Union against Tuberculosis and Lung Disease give the most recent estimates on the prevalence of anti-TB drug resistance. According to WHO statistics [1], 8.6 million people fell ill with TB in 2012 in EU, including 1.1 million cases among people infected with HIV. The Region accounts for 39% of the global burden of TB in terms of incidence, and India alone accounts for 26% of the world's TB cases. It is estimated that about 3.4 million new cases of TB continue to occur each year and that about 4,50000 people died of TB in 2012, most of these in five countries, namely Bangladesh, India, Indonesia, Myanmar and Thailand [2].

One of the major challenges associated with drug resistant tuberculosis is the lack of diagnostic capacity. WHO puts the number of MDR-TB cases detected globally around 18% and an even smaller fraction of detected cases of XDR-TB. This lack of diagnostic capability is due to critical gaps in laboratory capacity for culture and drug susceptibility testing (DST). Therefore there is a need to expedite the efforts needed for the global surveillance and control of drug-resistant TB which can be achieved by the expanded capacity to diagnose MDR and XDR TB.

Adequate treatment of patients with TB starts with a preliminary diagnosis, obtained by identifying Mycobacterium Tuberculosis from clinical specimens and conducting DST of the organism to confirm or exclude resistance. Conventional laboratory methods for preliminary diagnosis require 3-4 weeks of culture of sputum samples. Apart from conventional methods, sequence-based diagnostic methods have been developed that detect specific mutations associated with drug resistance. These tools have the advantage of being rapid, high throughput, and easily compared between laboratories. However, the development of such diagnostic tools relies on detailed information about the mutations that lead to drug resistance and their relative frequency.

In comparison to the laboratory based methods and sequence based diagnostic methods, genomic signal processing based methods offer the advantage of faster analysis and assessment due to the availability of whole genome sequence information.

Genomic signal processing methods help in visualization of whole genomic data by completely understanding the underlying biological functions and are capable of extracting relevant embedded information, in comparison to standard methods of laboratory testing [3]. One such graphical representation method of genomic signals had been used in analysing MTB resistance to rifampicin based on the assessment of rpoB gene [4]. Graphical representations of genomic sequences can be used as a tool to determine local and global similarities, identify repetitive motifs, and differentiate between coding and non-coding regions in genomic sequences [5].

DNA is the main nucleic genetic material of the cells with a double helix structure and two antiparallel intertwined complimentary strands. There are four kinds of nitrogenous bases found in DNA that constitute the genomic sequences: thymine (T) and cytosine (C) - called pyrimidines, adenine (A) and guanine (G) - called purines. Base A always pairs with base T while base C always pairs with base G. Hence, the two strands of a DNA helix are complementary and contain exactly the same number of A,T bases and the same number of C,G bases. There are three main biochemical properties of nitrogenous bases [6] according to which they can be categorized:

1. Molecular structure— bases A and G are purines (R), while C and T are pyrimidines (Y)
2. Strength of links— bases A and T are linked by two hydrogen bonds (W- weak bond), while C and G are linked by three hydrogen bonds (S- strong bond).
3. Radical content— bases A and C contain the amino (NH3) group in the large groove (M class), while T and G contain the keto (C=O) group (K class).

In order to apply graphical representation techniques, DNA sequences need to be mapped into their corresponding numerical values for visualization and analysis with digital signal processing methods. Numerical representation gives the DNA sequences a characteristic signature in the composition and distribution of the nucleotides throughout the whole genome. Mutations however cause deviation from uniqueness which can be easily visualised by signal processing methods.

Several mathematical representations for genomic sequences have been reported in literature such as Voss representation [7], tetrahedral representation [8], DNA walk [9], Z-curves [10], Fourier transforms [11] and wavelet transforms [12]. Numerical representations of a DNA sequence and graphical analysis facilitates sequence identification and comparison of similarities and dissimilarities of sequences [13]. Frequency domain analysis of Voss representations has been used to determine coding regions in genomic sequences [14]. DNA walk has been used as a tool to visualize changes in nucleotide composition, locating coding and non coding regions, identifying periodicities and large scale local and global features present in many genomes [15], [16]. Fourier transforms have been used to determine periodicities in proteins, identification of protein coding DNA regions and open reading frames [17]. Wavelet transforms have been used to determine long-range correlations, locating periodicites in DNA sequences [18]. Z-curves have been used in identifying replication origins of archeal genomes [19].

Table1 Different Categories of MTB Sequences

Sequence Number	MTB Genome/ Genbank Accession Number	Type of MTB Sequence
1	H37Rv / NC_000962.3	DS
2	H37Ra / NC_009525	DS
3	F11/ NC_009565	DS
4	CDC1551/ NC_002755	DS
5	CCDC5079 /CP002884	DS
6	CCDC5079/NC_021251	DS
7	KZN605/NC_018078	XDR
8	KZN1435/NC_012943	MDR
9	KZN1435/ CP001658	MDR
10	CCDC5180/ CP001642	DR

2 Method

This paper discusses the plotting (3-dimensional, 2-dimensional and 1-dimensional) of the whole genome sequences of different strains of MTB to determine the deviations in the patterns of the resistant and susceptible sequences. The 3-dimensional plots are called Z-curves. The DNA sequences of MTB are first converted to mathematical representations and then plotted. The graphical representations show patterns which can be compared visually while highlighting significant features for further analysis.

Z curve constitutes a unique representation of a DNA sequence in three-dimensional space that contains all the information which reflects their symmetry, periodicity and global features of the

distribution of bases along the length of the entire DNA sequence. Z-curve helps to analyse a DNA sequence by visualising both global and local compositional features of genomes. The values of x-axis, y-axis and z-axis represent the purine minus pyrimidine (R–Y) distribution, amino minus keto (M–K) distribution and weak minus strong hydrogen bonded nucleotide (W–S) distribution respectively along the sequence. These values on the curve are represented by a series of nodes P_0, P_1, P_2, ..., P_N each with coordinates x_n, y_n and z_n , n = 0, 1, 2, …, N ; where N is the length of the DNA sequence [10] and x_n, y_n, z_n are defined as

$x_n=(A_n+G_n)-(C_n+T_n)= R_n - Y_n$

$y_n=(A_n+C_n)-(G_n+T_n)= M_n - K_n$

$z_n=(A_n+T_n)-(G_n+C_n)= W_n - S_n$

$x_n , y_n , z_n \in [-N, N]$;

A_n, G_n, C_n, T_n are the cumulative values of occurrences of the bases A, G, C and T respectively. Two more significant values namely AT and GC disparity can also be calculated from the values of x_n, y_n and z_n. AT disparity, defined by $(x_n + y_n)/2$ determines excess of A over T whereas GC disparity, defined by $(x_n - y_n)/2$ determines excess of G over C along the sequence length.

Ten different MTB sequences listed in Table 1 [20] were downloaded from NCBI [21] (National Centre for Biotechnology Information) database. Whereas sequences 1-6 were drug susceptible (DS), sequence 7 was XDR, sequences 8 and 9 were MDR, sequence 10 was DR. Different graphical representations such as 3-dimensional plots (Z curves), 2-dimensional plots showing R minus Y versus M minus K disparity and 1-dimensional plots showing R minus Y disparity, M minus K disparity, W minus S bond disparity, GC content and AT content with respect to the sequence length were plotted and compared.

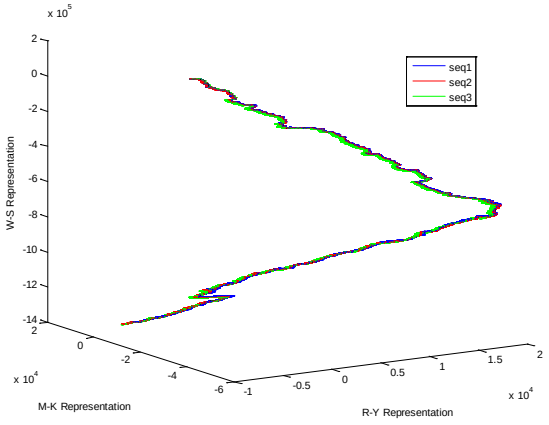

Fig. 1 Z Curves for Sequences 1,2,3

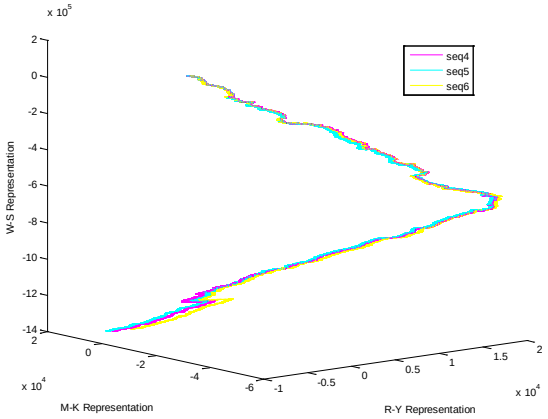

Fig. 2 Z Curves for Sequences 4,5,6

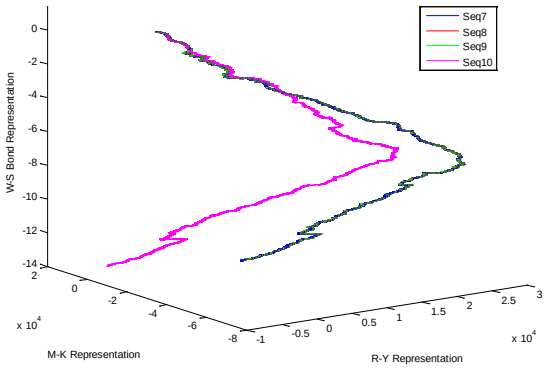

Fig. 3 Z Curves for Sequences 7,8,9,10

3 Results

The 3-dimensional Z-curve representations of the sequences were plotted as shown in figs. 1-3. From these figures it is observed that

i) DS sequences 1-6 (seq1-6) show overlapping plots (figures 1 and 2). The overall shape of the curves being similar, suggests global similarity in these sequences. The MDR sequences 8 and 9 and XDR sequence 7 show identical and overlapping plots but are significantly deviating from the Z-curve of the DR sequence 10 as shown in figure 3. Despite the fact that the sequences 7, 8 and 9 have different resistance characterisations, they exhibit similar Z- curves suggesting that they have significant underlying similarity along the whole genome. However the plot of the DR sequence (sequence 10) is different from the plot of the XDR and MDR sequences.

ii) The peak values of the 3-dimensional curves of DS strains (sequences 1-6) and DR strain (sequence 10) are similar but the peak values of MDR and XDR sequences are significantly

higher than those of DS and DR sequences. Thus while MDR and XDR sequences form one cluster, the DR and DS sequences form a separate cluster.

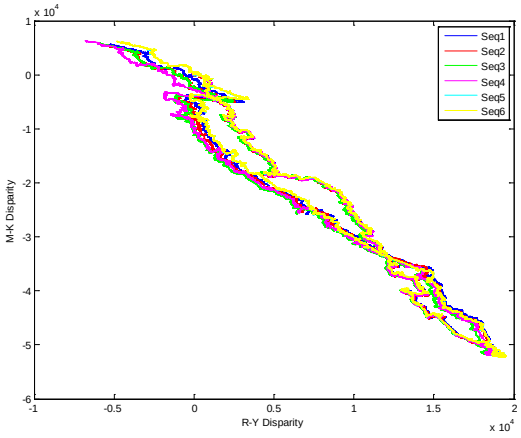

Fig. 4 R–Y vs M–K Representation: Sequences 1-6

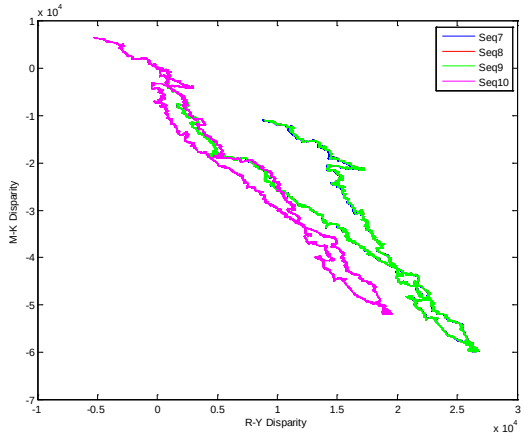

Fig. 5 R–Y vs M–K Representation: Sequences 7-10

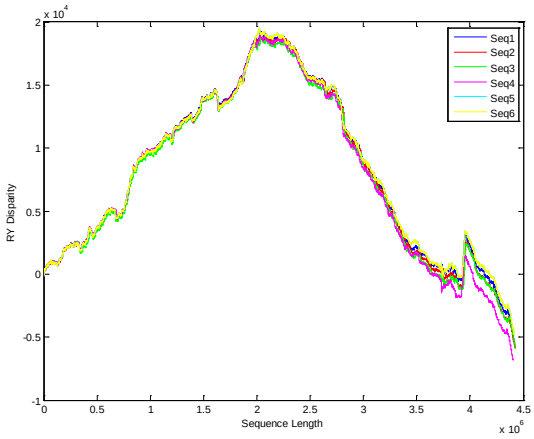

Fig. 6 R–Y Content: Sequences 1-6

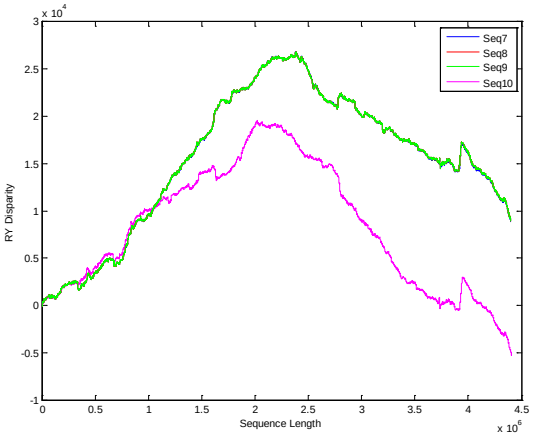

Fig. 7 R–Y Content: Sequences 7-10

From the 2-dimensional plots of M minus K content with respect to R minus Y content for all the sequences as shown in figs. 4, 5 it is observed that

i) The plot for DS sequences 1-6 and DR sequence 10 shows peak value of R–Y content is approximately 1.9×10^4 and the corresponding M-K value is -5.2×10^4 while the peak value of R-Y content of XDR and MDR strains is approximately 2.7×10^4 and the corresponding M–K value is -6×10^{-4}. Thus XDR and MDR sequences apparently have higher purine content and higher keto content than the DR and DS sequences.

ii) Visual comparison of the plots for DR and DS sequences also suggest that DR and DS sequences exhibit similarity.

From the 1-dimensional representations of R minus Y content (figs. 6, 7), M minus K content (figs. 8, 9), W minus S bond content (figs. 10, 11), GC disparity (figs. 12, 13) and AT disparity (figs. 14, 15) it is observed that

i) The R-Y curve divides the whole sequence into two regions: Purine rich and Pyrimidine rich. Purine rich region exists from beginning of the sequence upto approximately 2.25M bases (depicted by rising curve) and pyrimidine rich region beyond 2.25M bases (depicted by falling curve). Peak value of R–Y content for DS and DR sequences was significantly lower than the peak value for XDR and MDR sequences.

ii) The M–K plots show two different regions of the sequences: Keto rich and Amino rich. Keto rich region for DR and DS sequences exists from beginning of the sequence to approximately 2M bases but for XDR and MDR sequences keto rich

region occurs upto 2.25M bases. Amino rich region for DR and DS sequences exists beyond 2M bases whereas for XDR and MDR sequences, it exists beyond 2.25M bases. Keto content in XDR and MDR sequences is more than the keto content of DR and DS strains.

iii) Comparison of the plots of W–S bond for all the sequences do not show any deviation in the curves suggesting that the GC content in all the sequences is identical.

iv) Comparison of the cumulative GC disparity profile curves for all the sequences shows that the XDR and MDR sequences have a higher cumulative GC profile (peak value of 4.5x104) than those of the DS and DR sequences (peak value of 3.5x104).

v) Plots of Cumulative AT profiles suggest the peak value of the curve for XDR and MDR sequences is higher than that for DS and DR sequences.

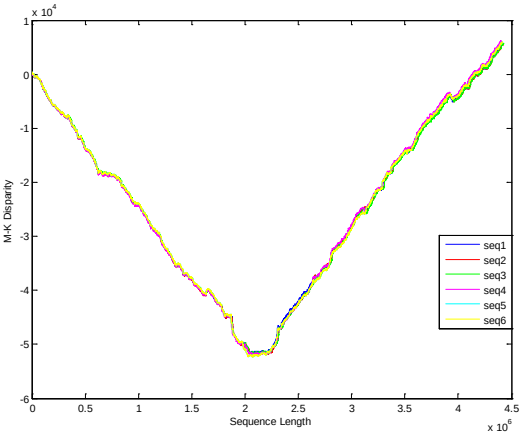

Fig. 8 M–K content: Sequences 1-6

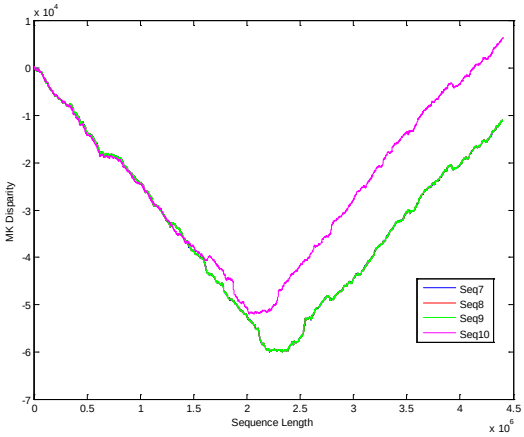

Fig. 9 M–K content: Sequences 7-10

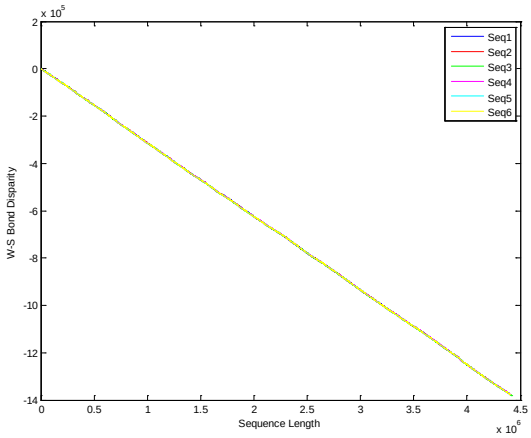

Fig. 10 W–S bond content: Sequences 1-6

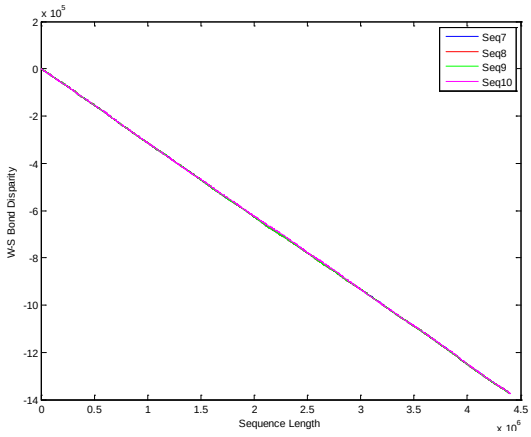

Fig. 11 W–S bond content: Sequences 7-10

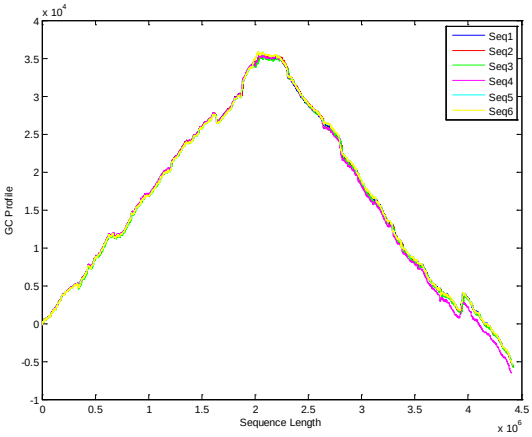

Fig. 12 Cumulative GC profile: sequences 1-6

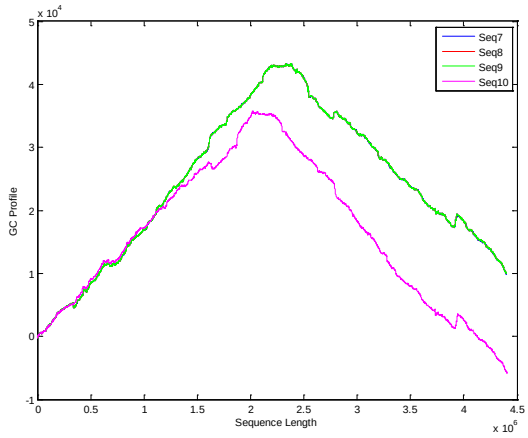

Fig. 13 Cumulative GC profile: sequences 7-10

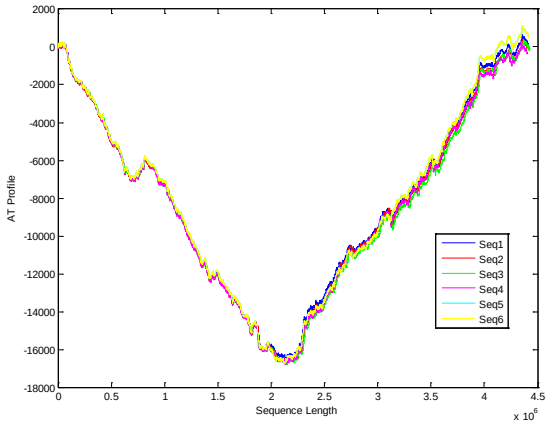

Fig. 14 Cumulative AT profile: Sequences 1-6

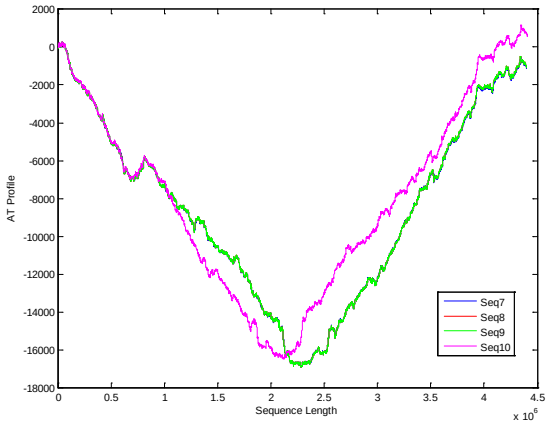

Fig. 15 Cumulative AT profile: Sequences 7-10

Thus it is observed from 3-D, 2-D and 1-D plots of whole genome sequences that XDR and MDR sequences can be differentiated from the DR and DS sequences by visual comparison.While MDR and XDR strains form one cluster, DR and DS together form a separate cluster. MDR and XDR TB sequences show higher peaks in 3-dimensional representations in comparison to DS and DR strains. 2-dimensional and 1-dimensional representations also show that XDR and MDR strains in comparison to DR and DS strains have higher values of purine content, keto content, cumulative GC content and cumulative AT content. Thus these representations can be used to differentiate between MDR, XDR and DR, DS strains

4 Conclusion

Global TB control is a difficult task due to the emergence of MTB drug resistance (MDR and XDR) in response to inadequate anti-TB therapy. This hampers the further choice of adequate treatment. But the conversion of genomic sequences into mathematical representations followed by signal processing methods can be used for faster processing and analyzing of genomic data. These graphical representation methods allow predicting possible strains of MTB by direct comparison so that different treatments for different strains of MTB can be initiated. Thus, the described method can be used in addition to the variety of pre-existing techniques for faster analysis and resistance assessment.

References:

[1] TUBERCULOSIS: WHO Global Tuberculosis Report 2013. http://www.who.int/tb/publications/global_report/en/.

[2] Tuberculosis control in the South-East Asia Region WHO Annual Report 2014. http://www.searo.who.int/entity/tb/documents/sea_tb_334/en/.

[3] Cristea, P. D., Tuduce, R., Banica, D., and Rodewald, K., Genomic Signals for the Study of Multiresistance Mutations in M Tuberculosis, *ISSCS 2007: International Symposium on Signals, Circuits and Systems*,Vol. 1, 2007, pp.1- 4.

[4] Cristea, P.D., Tuduce, R., Molecular Investigation in Support of the Clinical Decision: Early Diagnosis and Detection of Pathogen Drug Resistance, *in Proceedings of the 5th International Workshop on Wearable and Implantable Body Sensor Networks, in conjunction with The 5th International Summer School and Symposium on Medical Devices and Biosensors (ISSS-MDBS 2008), The Chinese University of Hong Kong, HKSAR, China*, 2008, pp. 239-242.

[5] Arniker, S., and Kwan, H., Graphical Representation of DNA sequences' *in EIT 2009: Proceedings of IEEE International Conference on Electro/Information Technology, Windsor, Ontario, Canada*, 2009, pp. 311-314.

[6] Paul Dan Cristea, "Phase Analysis of DNA Genomic Signals", *ISCAS '03 in Proceedings of the 2003 International Symposium on Circuits and Systems*, vol.5, 2003, V-25 - V-28.

[7] Voss, R. F., Evolution of Long-range Fractal Correlations and 1/f noise in DNA base sequences, *Physical Review Letters*, Vol. 68, 1992, pp. 3805-3808.

[8] Cristea, P.D., Phase Analysis of DNA Genomic Signals, *in ISCAS 2003: Proceedings of International Symposium on Circuits and Systems*, vol.5, 2003, pp. V25 - V28.

[9] Berger, J. A., Mitra, S. K. and Astola, J., Power spectrum analysis for DNA sequences, *in Proceedings of Seventh International Symposium on Signal Processing and its Applications*, Vol. 2, 2003, pp. 29-32.

[10] Zhang,C., Zhang,R., and Ou,H., The Z curve database: a graphic representation of genome sequences, *Bioinformatics*, Vol. 19 (5), 2003, pp. 593–599.

[11] Lorenzo-Ginori, J., Rodríguez-Fuentes, A., Grau Ábalo, R., and Rodríguez, R., Digital Signal Processing in the Analysis of Genomic Sequences , *Current Bioinformatics*, Vol. 4, 2009, pp. 28-40.

[12] Li `o,P. , Wavelets in bioinformatics and computational biology: state of art and perspectives, *Bioinformatics Review*, Vol. 19 no. 1, 2003, pp. 2–9.

[13] Nandy, A., Harle, M. and Basak, S.C., Mathematical descriptors of DNA sequences: development and Applications, *ARKIVOC (ix)*, 2006, pp. 211-238.

[14] Anastassiou, D., Frequency-domain analysis of biomolecular sequences, *Bioinformatics*, Vol. 16, No. 12, 2000, pp. 1073-1081.

[15] Berger, J. A., Mitra, S. K., Carli, M. and Neri, A., New Approaches to genome sequence analysis based on digital signal processing, *in GENSIPS 2002: Proceedings of IEEE Workshop on Genomic Signal Processing and Statistics, Raleigh, North Carolina, USA*, 2002, pp. 1-4.

[16] Haimovich, A. D., Byrne, B., Ramaswamy, R., Welsch, W. J., Wavelet analysis of DNA walks, *Journal of Computational Biology*, Vol. 13, 2006 , pp.1289- 1298.

[17] Zhou,Y., Zhou,L., Yu,Z. and Anh,V., Distinguish Coding And Noncoding Sequences In A Complete Genome Using Fourier Transform' *in ICNC 2007: Proceedings of Third International Conference on Natural Computation, Haikou, China*, 2007, pp. 295-299.

[18] Arneodo, A., D'Aubenton-Carafa, Y., Audit, B., Bacry, E., Muzy, J. F., and Thermes, C., What can we learn with wavelets about DNA sequences?, *Physica A*, Vol. 249, 1998, pp. 439–448.

[19] Zhang, R. and Zhang, C. T., Identification of replication origins in archaeal genomes based on the Z-curve method, 2005, *Archaea*, Vol. 1, pp. 335-346.

[20] Ilina, E. N., Shitikov, E A., Ikryannikova, L. N., Alekseev, D. G., Kamashev, D. E., Malakhova, M. V., Parfenova, T. V., Afanas'ev, M. V., Ischenko, S., Bazaleev, N. A., Smirnova, T. G., Larionova, E. E., Chernousova, L. N., Beletsky, A. V., Mardanov, A. V., Ravin, N. V., Skryabin, K. G., Govor, V. M. (2013), Comparative Genomic Analysis of Mycobacterium tuberculosis Drug Resistant Strains from Russia, *PLoS One*, Vol. 8 (2):e56577, 2013, DOI: 10.1371/journal.pone.0056577.

[21] National Centre for Biotechnology Information, National Institute of Health, National Library of Medicine website http://www.ncbi.nlm.nih.gov/genoms/,ftp://ftp.ncbi.nlm.nih.gov/genoms/,GenBank,http://www.ncbi.nlm.nih.gov/Genbank/index.html.