Frequency Distribution Analysis in 23S rRNA Sequences of Streptococcus and Helicobacter Sp. Using Mutual Information Theory

Chakresh Kumar Jain* *Member, IAENG,* Anurag Sharma, S.K. Sharma, Yamuna P. Shukla

***Abstract*— Resistance to antibiotics for both *helicobacter* & *streptococcus* has been found in great prevalence and mainly the macrolides. The target site for macrolides is the large (50S) subunit of the bacterial ribosome. a new regime of antibiotics such as triple eradication therapy employing a large no. of experimentations and trials are needed every time. So to reduce the cost associated with these trials, we are trying to find an alternative procedure by applying statistical tool like Mutual Information Theory which is a measure of reduction of uncertainty and can be used to associate sequence positions and groups of sequence positions with drug resistance. In order to evaluate the significance of mutual information shared by the adjacent bases, we have subjected the sequences of both bacterium, in the form of databases to find the mutual information scores. The results obtained after applying the calculations were plotted on the graph. The results were analyzed after adding the trend line (of moving averages) which gave us a better idea to find out the most probable region of variability. The results were in conformation with the known point mutations as the scores found here were high at these positions. The probable occurrence of most variable regions was found for both the bacterium and was documented.**

Index Terms— **Mutual information, variability, streptococcus, helicobacter, 23S rRNA.**

I. Introduction

Antibiotic resistance is becoming a great problem as most of the bacterium are modifying themselves to evade out the effect of antibiotics. The reason for the resistance is mainly due to the point mutations that are occurring in the 23S rRNA region of bacterium where most of the antibiotic bind and stop the translation process. Along with that efflux pumps methylation of the sequences are some of the other reasons conferring antibiotic resistance [1]. Even after employing different drugs regimes after a lot of experimentations and trials, these bacterial species are still able to skip their effect by modifying their sequence in course of time [2]. Hence to identify the probability of occurrence of variable region this could be the epicenter for creating mutation directly in that region which help in to figure out the region responsible for the resistance. Mutual Information Theory could be employed to find out the variable region [3].

Chakresh Kumar Jain* is lecturer in Department of Biotechnology. Jaypee Institute of Information Technology University, Noida, India .
(e-mail: ckj522@lyahoo.com).
A.Sharma. was the M.Tech Student in Department of Biotechnology.Jaypee Institute of Information Technology University, Noida, India.
S.K. Sharma is Assistant Professor in Department of Computer Science, Inderprastha Engineering College, Shahibabad, Ghaziabad, India.
Yamuna P. Shukla is Lecturer in Department of Computer Science , Jaypee Institute of Information Technology University, Noida, India .
(e-mail: shukla.yamuna@gmail.com).

II. METHODOLOGY

Around 53 entries of 23S rRNA were taken to form the dataset[8]. These entries belong basically to the strains of different species of *helicobacter* and *streptococcus* with an average of 2780 bases. This dataset was maintained in the form of database using MS Access and was subjected to the calculation used for finding Mutual Information Score. The results were obtained by developing a software using visual basic which found out the results for each and every possible combination in the dataset. The results were analyzed by plotting graphs for the values obtained for mutual information score. To find the probable region of highest variability, we plotted graphs by increasing the value of the MI score by a constant number and thus zeroing in to our required result.

(Table 1 List of sequences used in this study)
Helicobacter sp.

Entry_no	Accession_no
1	DQ307738
2	DQ307737
3	AY596244
4	AY596250
5	DQ418750
6	AY596256
7	DQ418749
8	AY596229
9	AY596245
10	AY596249
11	AY596220
12	AY596248
13	AY596237
14	AY596234
15	AY596221
16	AY596231
17	AY596223
18	AY596222
19	AY596233
20	AY596255
21	AY596253

22	AY596251
23	AY596235
24	AY596258
25	AY596239
26	AY596240
27	AY596243
28	AY596254
29	AY596259
30	AY596238

Streptococcus sp.

Entry_no	**Accession_no**
1	AJ544682
2	AJ544681
3	AB168125
4	AB096747
5	AB096741
6	AB096750
7	AB096748
8	AB096742
9	AB096753
10	AB096744
11	AB096751
12	AB096752
13	AB096740
14	AB096743
15	AB096749
16	AB096745
17	AB096746
18	AB168119
19	AB096754
20	AB168123
21	X68038
22	X68429
23	X68036

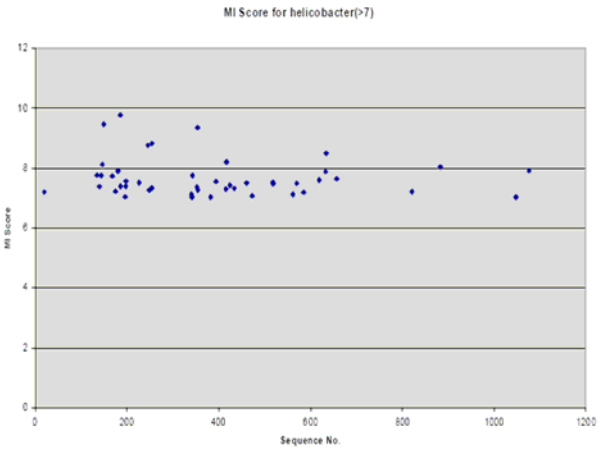

Fig 1. By reducing the range for the MI values, we ere able to find almost the exact probability of finding the most variable region of 23 S rRNA of helicobacter sp. as between (nucleotide number) 150 to 857.

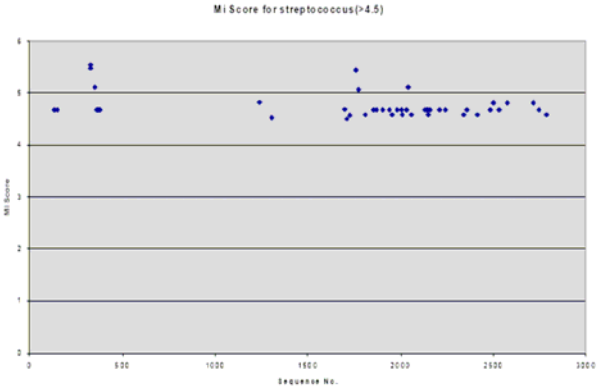

Fig 2. After analyzing the graph obtained for the values (>4.5) we can infer that the probability of finding the most variable region in *streptococcus* 23S rRNA region would be between nucleotide position1698 to 2787.

III. RESULT

From the experimental analysis carried out on the sequences of helicobacter & streptococcus sp., it could be easily figure out from the graphs obtained that the region of variability was between 21-1264 (nucleotide position) for helicobacter and 133-2801(nucleotide position) for streptococcus as the Mutual Information scores were found to be high in this region which is a reflection of variability. But after noise reduction and analyzing the concentration of mutual information scores from the graphs we have concluded that the most variable region for helicobacter lies between the nucleotide position 150 to 657 fig-1 & for streptococcus the region found was between 1698 to 2787 as depicted in fig-2.

IV. CONCLUSION

In this work we attempted to show the most variable region for both the bacterium. We proposed using Mutual Information approach consisting of making databases and finding the scores. The results obtained from the work set well with positions of mutation found as the MI scores were found to be high on these positions. Similarly the region of variability depicted for *streptococcus sp*. by the results shows that the mutation is generally occurring in that region (V domain) only. Few classical positions [4] are known for helicobacter but as we observed that in our study positional variability is higher in the segment where the mutation is generally not observed for most of the antibiotics used for treatment except for the clarithromycin which interacts with both II and V domain while interacting with 23S rRNA of the bacteria [5], also till date the mechanism working behind the frequent change in the antibiotic resistance region is still not clear [6]. This manifestation supports the region of variability defined by our work. In some of the different geographical regions it has been found out that some strains are showing resistance to the antibiotic with out mutations [7]. Furthermore, the lower values of MI score suggest the higher conservation of sequences for streptococcus sp. as compared to helicobacter sp. Since conserved sequences are responsible for storing the biological information. To reveal a novel, potentially important biological phenomenon, we had employed an information-theoretic tool, especially the mutual information, to statistically determine the dependent segments of DNA sequences. The proposed approach will provide a key to fundamental advances in understanding and quantifying biological information. The work addresses the problem of Identifying the leads based on the hyper variable region suitable for all the strains found in different geographical conditions and to find the probable region which is responsible for the frequent resistance regions shown by the bacterium. This work can be further extended to other species

ACKNOWLEDGMENT

We are thankful to our Jaypee Institute of Information Technology University, Noida for providing necessary support to complete this project, Further we are obliged to Dr. G.B.K.S Prasad, Coordinator, Department of Biotechnology, Jiwaji University, Gwalior for academic help.

REFERENCES

[1].Kataja J, *et.al. Antimicrobial Agents and Chemotherapy. 43:* 48 (1999) [PMID: 9869564]

[2] Debets-Ossenkopp, *et al. Journal of Antimicrobial Chemotherapy*, 43:511(1999).[PMID:10350380]

[3]. Identifying Statistical Dependence in Genomic Sequences via Mutual Information Estimates *www.cs.purdue.edu/homes/spa/papers/jbsb07.pdf*

[4].James Versalovic, *et al. Journal of Antimicrobial Chemotherapy*,40:283 (1997) [PMID: 9301997]

[5]. Stephen Douthwaite, *et al. Molecular Microbiology*, 36:183 .(2000) [PMID: 10760175]

[6].Qing Hao, *et al. World Journal of Gastroenterology*, 10:1075 (2004) [PMID: 15052698]

[7].Carla Fontana, *et al.Antimicrobial Agents and Chemotherapy*, 46:3765 (2002) PMID: 12435674

[8] National Center for Biotechnology Information *http://www.ncbi.nlm.nih.gov*